Méthodes ensemblistes pour l'estimation d'état et de paramètres

Tarek Raïssi

THESE

Présentée à l'Université Paris XII – Val de Marne
en vue de l'obtention du titre de

DOCTEUR de L'UNIVERSITE PARIS XII VAL de MARNE

Spécialité : Automatique

Méthodes ensemblistes pour l'estimation d'état et de paramètres

Par

Tarek RAISSI

Soutenance prévue le 29 novembre 2004 devant la commission d'examen :

Rapporteurs

M. Alain	RICHARD	Professeur
M. Eric	WALTER	Directeur de recherche

Examinateurs

M. Pierre	BORNE	Professeur
M. Gilbert	TEYSSEDRE	Chargé de recherche

Directeurs de Thèse

M. Yves	CANDAU	Professeur
M. Nacim	RAMDANI	Maître de Conférences

Centre Etude et de Recherche en Thermique, Environnement et Système
Université Paris XII Val de Marne
61 Avenue de Général De Gaulle 9401 Créteil Cedex

Table des Matières

Introduction

L'estimation d'état et de paramètres joue un rôle déterminant dans plusieurs domaines de l'ingénierie comme la commande de processus ou le diagnostic. L'élaboration d'une décision (une commande par exemple) est généralement basée sur ces grandeurs estimées à l'aide de données mesurées. Les méthodes classiques d'estimation de paramètres, dites dans un contexte statistique, consistent à trouver la valeur optimale de ces grandeurs au sens d'un critère liant les sorties du modèle aux données expérimentales. Cette valeur optimale est déterminée en utilisant un algorithme d'optimisation ponctuelle. Ceci revient généralement à supposer qu'une description statistique du bruit de mesure est disponible.

Dans certaines applications, il est difficile de décrire les perturbations par des lois de probabilité. Il est donc plus judicieux de considérer que l'erreur entre la sortie du modèle et celle du système est bornée et de bornes connues. Ces bornes tiennent compte du bruit de mesure et des erreurs de modélisation. Dans ce cas, on ne cherche plus une valeur du vecteur de paramètres permettant de minimiser le critère, mais un ensemble de valeurs *acceptables*. Cet ensemble contient d'une manière garantie toutes les valeurs du vecteur de paramètres, telle que l'erreur entre la sortie prédite et celle du système réel n'excède pas les bornes fixées *a priori*. Dans ce contexte, l'estimation de paramètres est alors un problème d'inversion ensembliste.

Lorsque le modèle est linéaire, l'ensemble solution est un polytope que l'on peut caractériser lorsque la dimension du vecteur des paramètres à identifier est assez réduite. Néanmoins, la structure de l'ensemble des valeurs acceptables devient rapidement très complexe lorsque le nombre de mesures augmente, et la description exacte reste un problème très difficile à résoudre. En conséquence, une approximation extérieure à l'aide de formes géométriques simples, par exemple des ellipsoïdes ou des orthotopes, est souvent privilégiée.

Par ailleurs, lorsque le modèle est non-linéaire, la caractérisation exacte de l'ensemble des valeurs acceptables n'est souvent pas possible. D'autres part, l'utilisation des formes géométriques, utilisées dans le cas des modèles linéaires, pour trouver une approximation extérieure de l'ensemble des valeurs admissibles, noté par $\mathbb{S}$, n'est pas pratique. Un algorithme d'inversion ensembliste par *analyse par intervalles*, SIVIA [JW93a] [JW93b], a été alors développé afin de résoudre ce problème. Cet algorithme permet de trouver des approximations intérieure et extérieure garanties de l'ensemble $\mathbb{S}$. Il est basé sur l'évaluation de fonctions sur des intervalles ; l'arithmétique des intervalles est alors utilisée.

L'analyse par intervalles a été introduite par Moore [Moo66] comme un outil permettant de tenir compte des erreurs d'arrondi dues à la précision finie des calculateurs. Cet outil est maintenant utilisé dans plusieurs domaines d'ingénierie et des bibliothèques mathématiques dédiées à l'analyse par intervalles sont disponibles.

Dans la première partie de cette thèse (chapitre 2 et 3), nous allons considérer le problème de l'estimation de paramètres d'une classe particulière de systèmes pouvant être représentés par une fonction de transfert. Dans ce dernier cas, le modèle utilisé est donné par une fonction

explicite à variable complexe. La deuxième partie de la thèse est dédié à l'estimation d'état et de paramètres pour une classe plus générale de systèmes, i.e. systèmes décrits par des équations différentielles ordinaires.

Dans le premier cas, les incertitudes sont définies par des intervalles complexes ; deux approches sont alors possibles pour propager ces incertitudes. La première consiste à décomposer la sortie du modèle en deux fonctions explicites à variables réelles, ceci permet donc d'utiliser les bibliothèques d'arithmétique des intervalles réels disponibles. Néanmoins, dans plusieurs applications, cette décomposition explicite n'est pas possible. La seconde approche consiste à utiliser l'arithmétiques des intervalles complexes. Par ailleurs, plusieurs formes géométriques simples permettent de représenter une incertitude complexe ; dans la littérature, les représentations rectangulaire et circulaire sont les plus utilisées. Néanmoins, l'utilisation de ces formes pour l'évaluation de fonctions non-linéaires conduit à des résultats souvent très pessimistes. Ce pessimisme est dû au fait que l'évaluation des fonctions élémentaires (i.e. log, sinh, cosh, etc.) ainsi que les opérations arithmétiques $\{*, /\}$ ne sont pas exactes. Un premier objectif de cette thèse consiste à développer une nouvelle arithmétique, plus pratique dans le cas non-linéaire, basée sur la représentation polaire des nombres complexes.

L'estimation d'état et de paramètres dans un contexte à erreurs bornées pour des systèmes décrits par des équations différentielles représente le second objectif de cette thèse. En effet, ce problème n'a pas été assez étudié dans le contexte ensembliste.

Ce document est structuré comme suit : dans le chapitre 1, nous allons détailler les principaux outils de l'analyse par intervalles qui seront utilisés tout au long de ce document.

L'arithmétique des intervalles complexes, représentés par la forme polaire, représente la première contribution de mon travail ; elle est présentée dans le chapitre 2. Dans ce cas, les incertitudes complexes sont représentées par des *secteurs* ; l'avantage de cette représentation vient du fait que les opérations de multiplication et de division de deux secteurs, telles qu'elles sont définies, sont exactes. Cette arithmétique est alors préférable pour l'évaluation de fonctions non-linéaires. Néanmoins, l'addition et la soustraction ne sont plus exactes. Nous proposons alors des algorithmes permettant de trouver le plus petit secteur contenant la somme (ou la différence) de deux secteurs. Nous verrons alors que l'addition définie dans ce chapitre garantit la condition de *minimalité,* i.e. on obtient le plus petit secteur contenant la somme (ou la différence). Dans cette thèse, la somme de deux secteurs est traitée comme étant un problème d'optimisation sous contraintes. Ce dernier est résolu analytiquement ; ceci est possible étant donné le nombre réduit de variables.

Dans le chapitre 3, les techniques d'estimation de paramètres dans un contexte à erreurs bornées sont appliquées à l'estimation de paramètres physiques pour deux applications différentes. Dans les deux cas, les modèles utilisés sont fortement non-linéaires et à variables complexes. La bibliothèque des intervalles complexes polaires développée dans le chapitre 2 est utilisée pour l'évaluation de ces modèles.

Dans la première application, consacrée à l'étude de modèles diélectriques, nous allons illustrer un avantage des méthodes d'estimation dans un contexte à erreurs bornées concernant

la *validation* de modèles. Nous verrons qu'il est possible de sélectionner, parmi un nombre restreint de modèles, le modèle permettant d'expliquer les données expérimentales.

Par ailleurs, dans cette partie, la sortie du modèle peut être décomposée explicitement en parties réelle et imaginaire ; l'arithmétique des intervalles réels peut alors être utilisée. Nous verrons alors que dans certains cas, il est plus judicieux d'utiliser la bibliothèque d'arithmétique des secteurs développée dans le chapitre 2.

Dans la deuxième application étudiée, consacrée à l'estimation de paramètres thermophysiques de matériaux, le modèle est donné par une fonction explicite à variables complexes. Néanmoins, il n'est pas possible de déterminer par le calcul symbolique une décomposition explicite en deux parties réelle et imaginaire. Dans ce cas, l'utilisation des intervalles complexes est indispensable. D'un autre côté, étant donné que le modèle est fortement non-linéaire, l'arithmétique des secteurs présentée dans le chapitre 2 est utilisée.

Dans le cadre de cette thèse nous allons considérer une seconde classe de modèles : modèles décrits par des équations différentielles ordinaires (EDOs). Deux approches sont alors possibles pour utiliser les techniques d'inversion ensembliste dans le cadre de l'estimation d'état et de paramètres pour ce type de modèles. La première méthode consiste à calculer analytiquement la solution de l'équation différentielle. En général, la résolution symbolique de ce type d'équations est très difficile, voire impossible à réaliser étant donné le caractère non linéaire de ces équations. Dans le cadre de cette thèse, nous allons opter pour des approximations numériques garanties de la solution des équations différentielles. Le chapitre 4 est alors consacré aux schémas d'intégration numérique garantie des équations différentielles ordinaires.

L'utilisation des méthodes garanties permet de calculer deux bornes inférieure et supérieure d'un pavé dont on garantit qu'il contient la solution de l'équation différentielle considérée. Ces méthodes permettent de s'assurer que le problème étudié contient ou non une solution. En effet, si le pavé retourné est vide, alors le système d'équations différentielles ne possède pas de solution ; dans le cas contraire son existence est assurée.

Les techniques d'intégration garantie des équations différentielles sont utilisées dans le chapitre 5 afin de développer des estimateurs d'états et de paramètres pour des systèmes non linéaires à temps continu. Cette partie constitue la seconde contribution de mon travail de thèse.

Nous proposons en premier lieu un estimateur basé sur l'approche *prédiction-correction* et semblable au filtre de Kalman. La phase de prédiction consiste à intégrer l'EDO à un instant t_{j+1} étant donné la solution à t_j ; cette étape produit alors un pavé $[\mathbf{x}_{j+1}]^+$. La correction consiste à contracter ce dernier pavé en utilisant la mesure disponible à t_{j+1}. On obtient alors un pavé $[\mathbf{x}_{j+1}]$ contenant les valeurs du vecteur d'état consistantes avec $[\mathbf{x}_j]$, la mesure à l'instant t_{j+1} et la borne d'erreur fixée *a priori*.

Cet estimateur a été étendu afin de développer un observateur à horizon glissant. Ce dernier permet d'estimer l'état à un instant t_j en utilisant les informations (mesures) disponibles sur un horizon de taille N. Les techniques de consistance permettent alors d'exploiter les données de tout l'horizon afin de réduire la taille du vecteur d'état. Cet observateur a été appliqué à l'étude d'un bioprocédé.

Enfin, les techniques d'inversion ensembliste ont été associées aux méthodes d'intégration garantie des équations différentielles afin d'étudier le cas où des paramètres inconnus doivent être estimés.

L'ensemble de ce travail a fait l'objets des articles [RRC04], [RIRC05] et [RRC05], des actes de conférences [RRC03a], [RRC03b], [RRC03c], [RIRC03], [RIRC04] et [RRIC04] et à la soumission de l'article [CRRI04].

Chapitre 1

Analyse par intervalles : Arithmétique et outils

1. Introduction

Dans certains domaines d'ingénierie, les données utilisées sont souvent de nature incertaine. Les sources de cette incertitude sont multiples, i.e. modèles mathématiques contenant des paramètres incertains, représentation des nombres réels sur des calculateurs numériques de précisions finies, données initiales incertaines. Dans certaines applications, il est nécessaire de connaître l'influence de ces incertitudes sur la solution calculée. Pour résoudre ce type de problèmes, des techniques basées sur l'analyse par intervalles ont été développées notamment par Moore [Moo66]. L'utilisation de cet outil permet de calculer un intervalle contenant d'une manière garantie la vraie solution.

Pour montrer l'intérêt de l'utilisation de l'analyse par intervalle, il suffit d'étudier l'exemple proposé par Rump [Rum88] illustrant les inconvénients de l'utilisation des précisions finies.

On considère la fonction suivante :

$$f(x,y) = \frac{1335}{4} y^6 + x^2 \left(11x^2y^2 - y^6 - 121y^4 - 2\right) + \frac{11}{2} y^8 + \frac{x}{2y}$$

L'évaluation de cette fonction pour x = 77617 et y = 33096 (qui sont exactement représentables) sur une machine S/370 donnent les valeurs suivantes [Han92] :

Précision	Résultat
Simple	1.172603
Double	1.1726039400531
Etendue	1.172603940053178

En utilisant ces trois précisions, les sept premiers chiffres sont les mêmes. Le résultat semble donc correct. Ceci n'est pas du tout le cas et le résultat est complètement faux pour les trois cas [Han92]. En effet, la valeur correcte de f est -0.8273960599468213. Pour obtenir l'évaluation exacte, il faut factoriser correctement la fonction f et l'évaluer en utilisant des entiers.

Par contre, si on fait l'évaluation de f en utilisant l'arithmétique d'intervalles on obtient un intervalle de grande taille qui contient la valeur correcte.

Chapitre 1

Durant ces dernières décennies, plusieurs algorithmes basés sur l'analyse par intervalles ont été développés dans plusieurs domaines, et ce dans le but d'étudier et de quantifier les effets des incertitudes (numériques et physiques) sur les données manipulées. Dans ce chapitre, seule l'influence d'incertitudes physiques sera pris en compte. Ainsi différentes méthodes de résolution de problèmes, donnés sous la forme de systèmes d'équations (ou contraintes), par analyse par intervalles seront présentées.

Ce chapitre est inspiré des différents travaux présentés dans le livre [JKDW01]. Il structuré comme suit : dans la section 2, nous allons rappeler les principales méthodes d'analyse par intervalles ; elles sont utilisées dans la section 3 dans le cadre de l'inversion ensembliste. La section 4 est dédiée aux principaux contracteurs étudiés dans la littérature. Enfin, nous allons donner un bref rappel de la différentiation automatique.

2. Arithmétique des intervalles

2.1. Intervalles

Définition 1 : Un intervalle, noté par $[x]$ est un ensemble connexe et borné de $\mathbb{R}$, il est défini par :

$$[x]=\left[x^-, x^+\right] = \left\{x\in \mathbb{R} \mid x^- \le x \le x^+\right\} \tag{1}$$

Les nombres réels x^- et x^+ sont respectivement les bornes inférieure et supérieure de $[x]$. Un intervalle est dit dégénéré lorsque $x^- = x^+$. Les intervalles dégénérés permettent la représentation des nombres réels représentables sur une machine. L'ensemble des intervalles de $\mathbb{R}$ est noté par $\mathbb{IR}$.

Remarque 1 : Il est probable que les nombres réels x^- et x^+ ne soient pas représentables d'une manière exacte sur une machine (par exemple $x^- = 0.1$). Dans ce cas, il est indispensable de prendre pour x^- le plus grand nombre réel inférieur à x^- représentable sur une machine. De même, on prend pour x^+ le plus petit nombre réel supérieur à x^+ et représentable sur un calculateur. Dans la suite de ce document, on supposera que les nombres x^- et x^+ sont parfaitement représentables. ♦

Les opérations mathématiques élémentaires sont étendues aux intervalles. Le résultat d'une opération entre deux intervalles est un intervalle qui contient tous les résultats des opérations entre les éléments des deux intervalles. Le résultat d'une opération entre deux intervalles de bornes finies est obtenu en travaillant uniquement sur leurs bornes.

Soient $[x], [y] \in \mathbb{IR}$, et $\circ \in \{+,-,*,/\}$, alors :

$$[x]\circ[y] \;=\; \left\{x\circ y \;\middle|\; x \in [x],\; y \in [y]\right\} \tag{2}$$

La définition (2) est valable pour toutes les opérations à l'exception de la division lorsque $0 \in [y]$. Dans ce dernier cas, le résultat n'est pas un intervalle.

L'expression (2) peut s'écrire comme suit :

$$[x]+[y]=\left[x^- + y^-, x^+ + y^+\right] \tag{3}$$

$$[x]-[y]=\left[x^- - y^+, x^+ - y^-\right] \tag{4}$$

$$[x]*[y]=\left[\min\left(x^-y^-, x^-y^+, x^+y^-, x^+y^+\right), \max\left(x^-y^-, x^-y^+, x^+y^-, x^+y^+\right)\right] \tag{5}$$

$$[x]/[y]=[x]*\left[1/y^+, 1/y^-\right],\ 0 \notin [y] \tag{6}$$

Définition 2

Soit $[x] \in \mathbb{IR}$, on définit alors:

- sa borne inférieure : $inf\left([x]\right)=x^-$
- sa borne supérieure : $sup\left([x]\right)=x^+$
- sa largeur : $w\left([x]\right)=x^+-x^- \geq 0$
- son milieu : $mid\left([x]\right)=\left(x^+ + x^-\right)/2$
- son rayon : $rad\left([x]\right)=\dfrac{x^+-x^-}{2} \geq 0$

La largeur d'un intervalle (ainsi que son rayon) peut être interprétée en terme d'incertitude sur la variable représentée par cet intervalle.

2.2. Vecteurs et matrices d'intervalles

Un vecteur d'intervalles (ou pavé) noté par $[\mathbf{x}]=\left([x_1],[x_2],\ldots,[x_n]\right)^T$ est un vecteur dont les éléments sont des intervalles ; on note par $\mathbb{IR}^n$ l'ensemble des vecteurs d'intervalles de $\mathbb{R}^n$. Les fonctions élémentaires définies pour les intervalles sont aussi définies pour les vecteurs d'intervalles.

Soit $[\mathbf{x}] \in \mathbb{IR}^n$, alors :

- sa borne inférieure est : $inf\left([\mathbf{x}]\right)=\left(x_1^-, x_2^-, \ldots, x_n^-\right)^T$
- sa borne supérieure est : $sup\left([\mathbf{x}]\right)=\left(x_1^+, x_2^+, \ldots, x_n^+\right)^T$

- sa largeur est : $w([\mathbf{x}]) = \max_{j=1}^{n} \left(x_j^+ - x_j^- \right) \geq 0$
- son milieu est : $mid([\mathbf{x}]) = \left(mid([x_1]), mid([x_2]), \ldots, mid([x_n]) \right)^T$

2.3. Pessimisme

Généralement, le résultat d'une suite d'opérations entre deux ou plusieurs intervalles n'est pas minimal ; l'intervalle obtenu est donc pessimiste. Ce problème est dû principalement à deux phénomènes : *dépendance* et *enveloppement*.

2.3.1. Phénomène de dépendance

On considère un intervalle non dégénéré $[x] = \left[x^-, x^+ \right]$ et une opération $\circ \in \{+, -, *, /\}$, alors en utilisant la définition (2), on obtient :

$$[x] \circ [x] \; = \; \{ x \circ y \;\; | \;\; x \in [x], \;\; y \in [x] \} \tag{7}$$

D'après (7), on constate que les variables *x* et *y* sont considérées comme différentes malgré le fait que l'on manipule le même intervalle ; ce problème est appelé phénomène de *dépendance*.

Exemple 1: Soit $[x] = [-1,1]$, alors $[x] - [x] = [-1,1] - [-1,1] = [-2,2] \neq \{0\}$

2.3.2. Phénomène d'enveloppement

L'effet d'enveloppement caractérise le pessimisme dû à la représentation d'un ensemble quelconque par un pavé (vecteur d'intervalles).

Exemple 2 [Moo66]: On considère des rotations successives d'un pavé $[\mathbf{x}] = [x_1] \times [x_2]$ à l'aide d'une rotation définie par la matrice

$$\mathbf{A} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}$$

Une rotation du pavé $[\mathbf{x}]$ donne lieu à un « rectangle » de même taille, tracé sur la figure 1. La représentation de ce rectangle par un pavé se fait en l'enveloppant par un autre rectangle dont les cotés sont parallèles aux axes du repère. La rotation suivante de ce pavé dont la taille est plus grande que celle du pavé initial donne lieu à un autre rectangle de même taille. Sa représentation par un pavé augmente sa taille. Par suite, la rotation successive d'un pavé génère une suite de pavés de tailles croissantes alors que la rotation est une opération conservatrice.

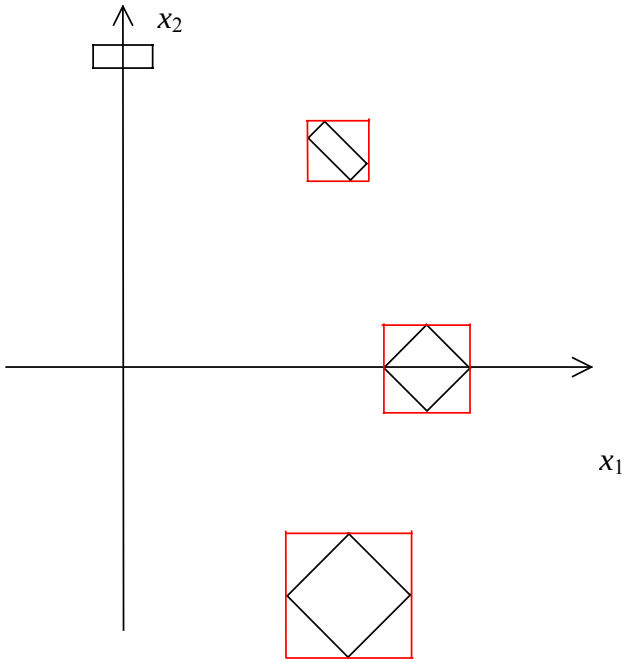

Figure 1 : Effet d'enveloppement pour $\theta = -\pi/4$

2.4. Fonctions d'inclusion

Soit $\mathbf{f}:\mathbb{R}^n \to \mathbb{R}^m$ une fonction vectorielle contenant un nombre fini d'opérations arithmétiques et de fonctions élémentaires {cos, sin, log, exp…}. Une fonction d'inclusion de **f** notée par [**f**] est une fonction de $\mathbb{IR}^n$ dans $\mathbb{IR}^m$ vérifiant

$$\mathbf{f}([\mathbf{x}]) = \{\mathbf{f}(\mathbf{x}) \mid \mathbf{x} \in [\mathbf{x}]\} \subseteq [\mathbf{f}]([\mathbf{x}]) \tag{8}$$

En général, la fonction d'inclusion n'est pas unique et dépend de la manière dont **f** est écrite. L'objectif général de l'analyse par intervalles est de pouvoir utiliser des fonctions d'inclusion peu pessimistes dans le sens où la taille de $([\mathbf{f}]([\mathbf{x}]) - \mathbf{f}([\mathbf{x}]))$ est assez petite.

Définition 3 - *Monotonie au sens de l'inclusion :* Une fonction d'inclusion [**f**] d'une fonction **f** est dite monotone si

$$[\mathbf{x}] \subset [\mathbf{y}] \Rightarrow [\mathbf{f}]([\mathbf{x}]) \subset [\mathbf{f}]([\mathbf{y}]) \tag{9}$$

Définition 4 – *Convergence :* Une fonction d'inclusion [**f**] d'une fonction **f** est dite convergente si

$$\lim_{k \to \infty} w([\mathbf{x}](k)) = 0 \quad \Rightarrow \quad \lim_{k \to \infty} w(\mathbf{f}([\mathbf{x}](k))) = 0 \tag{10}$$

2.4.1. Fonction d'inclusion des fonctions élémentaires

La construction des fonctions d'inclusion pour les fonctions élémentaires {exp, log, cos, sin,…} est basée sur des propriétés de monotonie.

Propriété 1 : Soit *f* une fonction continue croissante (respectivement décroissante) sur un intervalle $D \subset \mathbb{R}$, alors la fonction intervalle qui, à tout intervalle $[x] \subset D$ associe l'intervalle $\left[f\left(x^-\right), f\left(x^+\right) \right]$ (respectivement $\left[f\left(x^+\right), f\left(x^-\right) \right]$) est une fonction d'inclusion minimale de *f* (i.e. la fonction qui donne l'évaluation la moins pessimiste). ♦

Exemple 3 :

- fonction Log : le domaine de la fonction Logarithme est $D =]0, +\infty[$, alors :

$$\forall [x] = \left[x^-, x^+ \right] \subset D, \mathrm{Log}([x]) = \left[\mathrm{Log}\left(x^-\right), \mathrm{Log}\left(x^+\right) \right]$$

- fonction Exp : le domaine de la fonction Exponentielle est $D =]-\infty, +\infty[$, alors :

$$\forall [x] = \left[x^-, x^+ \right] \subset D, \mathrm{Exp}([x]) = \left[\mathrm{Exp}\left(x^-\right), \mathrm{Exp}\left(x^+\right) \right]$$

2.4.2. Fonction d'inclusion naturelle

Soit **f** une fonction de $\mathbb{R}^n$ dans $\mathbb{R}^m$. La fonction d'inclusion naturelle $[\mathbf{f}]$ de **f** s'obtient en remplaçant chaque variable réelle x_i par sa variable intervalle correspondante $[x_i]$ et chaque opération arithmétique par son équivalente intervalle. Cette fonction d'inclusion est convergente si la fonction **f** comporte seulement des fonctions élémentaires continues ainsi que des opérations continues. La fonction d'inclusion naturelle est minimale si **f** est continue et chaque variable n'apparaît qu'une seule fois.

Exemple 4 : On considère une fonction *f* écrite en utilisant ces quatre expressions :

$$f_1(x) = x^2 + 2x$$

$$f_2(x) = x(x+2)$$

$$f_3(x) = x \cdot x + 2x$$

$$f_4(x) = (x+1)^2 - 1$$

L'évaluation des fonctions d'inclusions naturelles de ces quatre expressions sur l'intervalle $[x] = [-1, 1]$ donne :

$$[f_1]([x])=[x]^2+2[x]=[-2,3]$$

$$[f_2]([x])=[x]([x]+2)=[-3,3]$$

$$[f_3]([x])=[x][x]+2[x]=[-3,3]$$

$$[f_4]([x])=([x]+1)^2-1=[-1,3]$$

On remarque que la taille des intervalles obtenus par ces quatre fonctions d'inclusion dépend de l'expression utilisée pour l'écriture de *f*. Ceci est dû au phénomène de dépendance expliqué dans la section 2.2.1. Comme la fonction *f* est continue, la fonction d'inclusion naturelle est minimale lorsque le nombre d'occurrences des variables est égal à un. Dans cet exemple, la fonction d'inclusion naturelle $[f_4]$ est minimale ; elle permet de trouver le plus petit intervalle contenant l'image de $[x]$ par *f*.

2.4.3. Fonctions d'inclusion centrées

Soit la fonction $f:\mathbb{R}^n\to\mathbb{R}$ différentiable sur un domaine $D\subset\mathbb{R}^n$. On considère un pavé $[\mathbf{x}]\subset D$ et on note par $\hat{\mathbf{x}}=mid([\mathbf{x}])$ le centre du pavé $[\mathbf{x}]$. En utilisant le théorème de la valeur moyenne [Neumaier, 1990], on obtient :

$$\forall \mathbf{x}\in[\mathbf{x}],\ \exists\xi\in[\mathbf{x}]\mid f(\mathbf{x})=f(\hat{\mathbf{x}})+\mathbf{J}(\xi)(\mathbf{x}-\hat{\mathbf{x}}) \tag{11}$$

où **J** est le Jacobien de la fonction *f*.

On suppose qu'une fonction d'inclusion $[\mathbf{J}]$ de **J** est disponible, alors :

$$\forall \mathbf{x}\in[\mathbf{x}],\ f(\mathbf{x})\in f(\hat{\mathbf{x}})+[\mathbf{J}]([\mathbf{x}])(\mathbf{x}-\hat{\mathbf{x}}) \tag{12}$$

Par suite

$$f([\mathbf{x}])\subseteq f(\hat{\mathbf{x}})+[\mathbf{J}]([\mathbf{x}])([\mathbf{x}]-\hat{\mathbf{x}}) \tag{13}$$

Le terme de droite dans l'expression (13) est appelé fonction d'inclusion centrée de *f* notée par $[f_c]$.

$$f([\mathbf{x}])\subseteq[f_c]([\mathbf{x}])=f(\hat{\mathbf{x}})+[\mathbf{J}]([\mathbf{x}])([\mathbf{x}]-\hat{\mathbf{x}}) \tag{14}$$

La fonction d'inclusion centrée est généralement utilisée lorsque la taille des intervalles manipulés est assez petite (et loin des points stationnaires de la fonction *f*). Dans ce dernier cas, la forme centrée donne des résultats plus précis que la fonction d'inclusion naturelle.

2.4.4. Fonctions d'inclusion de Taylor

Les fonctions d'inclusion de Taylor sont obtenues en utilisant un ordre de dérivation plus élevé [Neu03]. A titre d'exemple, on donne ici la fonction d'inclusion de Taylor du second ordre :

$$[f_T]([\mathbf{x}]) = f(\hat{\mathbf{x}}) + \mathbf{J}([\mathbf{x}])([\mathbf{x}]-\hat{\mathbf{x}}) + \frac{1}{2}([\mathbf{x}]-\hat{\mathbf{x}})^T [\mathbf{H}]([\mathbf{x}])([\mathbf{x}]-\hat{\mathbf{x}}) \tag{15}$$

où $[\mathbf{H}]$ est une fonction d'inclusion du Hessien de la fonction *f*.

Comme nous l'avons noté pour le cas des fonctions d'inclusions centrées, le pessimisme peut être drastiquement réduit en utilisant les fonctions d'inclusion de Taylor pour des intervalles de petites tailles.

2.4.5. Convergence

La convergence des fonctions d'inclusion est étudiée au sens d'un critère proposé par Moore [Moo79]. L'ordre de convergence d'une fonction d'inclusion est le plus grand entier α tel que :

$$\exists \beta \in \mathbb{R}^+ \mid \left(w([f]([x])) - w(f([x])) \right) \leq \beta\, w([x])^{\alpha} \tag{16}$$

L'ordre de convergence d'une fonction d'inclusion minimale est infini, celui de la fonction d'inclusion naturelle est au moins linéaire ($\alpha \geq 1$) alors qu'il est au moins quadratique pour la forme centrée et la forme de Taylor ($\alpha \geq 2$). Ceci montre que l'utilisation de ces dernières fonctions d'inclusion est plus intéressante que l'utilisation de la fonction d'inclusion naturelle. Ce résultat n'est valide que pour des intervalles de petite taille. En effet, la propriété de convergence proposée par Moore [Moo79] est asymptotique. Pour les intervalles de grande taille, il est souvent préférable d'utiliser la fonction d'inclusion naturelle.

3. Inversion ensembliste par arithmétique d'intervalles

3.1. Inversion

On souhaite résoudre l'équation suivante :

$$\mathbf{f}(\mathbf{x}) \in [\mathbf{y}],\ \mathbf{x} \in \mathbb{X} \tag{17}$$

L'ensemble $\mathbb{S}$ des solutions de (17) est donné par :

$$\mathbb{S} = \left\{ \mathbf{x} \in \mathbb{X} \mid \mathbf{f}(\mathbf{x}) \in [\mathbf{y}] \right\} \tag{18}$$

Cet ensemble peut être réécrit sous la forme suivante :

$$\mathbb{S} = \mathbf{f}^{-1}\left([\mathbf{y}]\right) \cap \mathbb{X} \tag{19}$$

La caractérisation de l'ensemble $\mathbb{S}$ défini par (19) est un problème d'inversion ensembliste qui, pour le cas non linéaire, peut être résolu d'une manière garantie en utilisant l'algorithme SIVIA (*Set Inversion Via Interval Analysis*) proposé dans [JW93b]. Cet algorithme permet de trouver un encadrement (lorsque au moins une solution existe) de l'ensemble des solutions. On note respectivement par $\underline{\mathbb{S}}$ et $\overline{\mathbb{S}}$, un encadrement intérieur et un encadrement extérieur de l'ensemble solution $\mathbb{S}$:

$$\underline{\mathbb{S}} \subseteq \mathbb{S} \subseteq \overline{\mathbb{S}} \tag{20}$$

où

$$\overline{\mathbb{S}} = \underline{\mathbb{S}} \cup \Delta\mathbb{S}$$

avec

$\underline{\mathbb{S}}$: est l'ensemble des pavés prouvés solutions

$\Delta\mathbb{S}$: l'ensemble des pavés pour lesquels aucune décision n'a pu être établie

on a donc :

$$\text{vol}(\underline{\mathbb{S}}) \subseteq \text{vol}(\mathbb{S}) \subseteq \text{vol}(\overline{\mathbb{S}})$$

où : vol($\mathbb{S}$) est le volume de l'ensemble $\mathbb{S}$.

Propriété 2 :

- si $\overline{\mathbb{S}} = \varnothing$ le problème (19) ne possède aucune solution
- si $\underline{\mathbb{S}} \neq \varnothing$, l'ensemble $\mathbb{S}$ n'est pas vide ; il existe au moins une solution vérifiant (19)
- $\Delta\mathbb{S}$ contient les pavés pour lesquels aucune conclusion n'a pu être établie. $\Delta\mathbb{S}$ peut être interprété en terme d'incertitude sur la caractérisation de $\mathbb{S}$.
- Un majorant de l'incertitude sur chacun des paramètres est donné par la projection de l'ensemble $\overline{\mathbb{S}}$ sur l'axe correspondant à ce paramètre.

L'algorithme de partitionnement SIVIA [JW93b] permet une caractérisation garantie de ces ensembles de pavés en utilisant un test d'inclusion.

3.2. Test d'inclusion

Un test d'inclusion permet de tester si des points appartenant à un ensemble vérifient une propriété donnée. Dans notre cas, les variables booléennes utilisées sont de type intervalle ; elles sont obtenues par extension de l'ensemble des variables booléennes $\mathbb{B} = \{\varnothing, 0, 1\}$ au cas des intervalles. On note par $\mathbb{IB}$ l'ensemble des intervalles booléens.

$$\mathbb{IB} = \{\varnothing, [0], [1], [0,1]\} \tag{21}$$

où l'intervalle $[0]$ est relatif à la variable *faux*, $[1]$ à *vrai*, $[0,1]$ à *indéterminé* et $\varnothing$ pour *impossible*.

Les opérations sur les variables booléennes sont facilement étendues au cas des intervalles booléens. Soient $[x], [y] \in \mathbb{IB}$, alors :

$$\begin{aligned} [x] \wedge [y] &= \{ x \wedge y \mid x \in [x],\ y \in [y] \} \\ [x] \vee [y] &= \{ x \vee y \mid x \in [x],\ y \in [y] \} \\ \neg [x] &= \{ \neg x \mid x \in [x] \} \end{aligned} \tag{22}$$

Soit $t : \mathbb{B}^n \to \mathbb{B}$ une fonction booléenne ; une fonction $[t] : \mathbb{IB}^n \to \mathbb{B}$ est une fonction d'inclusion de *t* si :

$$\forall [\mathbf{x}] \in \mathbb{IB}^n,\ t([\mathbf{x}]) \subseteq [t]([\mathbf{x}])$$

La fonction d'inclusion naturelle de la fonction booléenne *t* est obtenue en remplaçant chaque variable booléenne et chaque opération élémentaire par son équivalent en intervalles booléens.

3.3. SIVIA

On considère de nouveau le problème (19) et le test suivant :

$$t : \begin{array}{l} \mathbb{X} \to \{0, 1\} \\ \mathbf{x} \mapsto (\mathbf{f}(\mathbf{x}) \in [\mathbf{y}]) \end{array}$$

Le test *t* prend les valeurs suivantes :

$$t(\mathbf{x}) = \begin{cases} 1 \text{ si } \mathbf{f}(\mathbf{x}) \in [\mathbf{y}] \\ 0 \text{ si } \mathbf{f}(\mathbf{x}) \notin [\mathbf{y}] \end{cases}$$

On définit un test d'inclusion de *t* comme suit :

$$[t]([\mathbf{x}]) = \begin{cases} [1] \text{ si } [\mathbf{f}]([\mathbf{x}]) \subseteq [\mathbf{y}] \\ [0] \text{ si } [\mathbf{f}]([\mathbf{x}]) \cap [\mathbf{y}] = \varnothing \\ [0,1] \text{ pour les autres cas} \end{cases}$$

Un pavé $[\mathbf{x}] \in \mathbb{X}$, est dit

- Faisable : $[\mathbf{x}] \in \underline{\mathbb{S}}$, si $[t]([\mathbf{x}]) = [1]$
- non faisable : $[\mathbf{x}] \notin \overline{\mathbb{S}}$, si $[t]([\mathbf{x}]) = [0]$
- indéterminé, si $[t]([\mathbf{x}]) = [0,1]$

Dans ce dernier cas, aucune décision à propos du pavé $[\mathbf{x}]$ n'est possible. Si sa taille est supérieure à une certaine tolérance η fixée par l'utilisateur, il sera partitionné en deux afin d'essayer d'avoir une décision sur les deux pavés générés. On présentera plus loin dans ce chapitre plusieurs stratégies de bissections afin d'accélérer cet algorithme. L'algorithme SIVIA avec le test d'inclusion présenté est le suivant :

Algorithme SIVIA(entrées : $[t], [\mathbf{x}], \eta$; sorties : $\underline{\mathbb{S}}, \overline{\mathbb{S}}$)

1 Si $[t]([\mathbf{x}]) = [0]$, rejeter $[\mathbf{x}]$;

2 Si $[t]([\mathbf{x}]) = [1]$, $\underline{\mathbb{S}} := \underline{\mathbb{S}} \cup [\mathbf{x}]$; $\overline{\mathbb{S}} := \overline{\mathbb{S}} \cup [\mathbf{x}]$;

3 Si $w([\mathbf{x}]) \leq \eta$, $\overline{\mathbb{S}} := \overline{\mathbb{S}} \cup [\mathbf{x}]$;

4 bissecter $[\mathbf{x}]$ en $([\mathbf{x}_1], [\mathbf{x}_2])$;

SIVIA(e : $[t], [\mathbf{x}_1], \eta$; s : $\underline{\mathbb{S}}, \overline{\mathbb{S}}$) ; SIVIA(e : $[t], [\mathbf{x}_2], \eta$; s : $\underline{\mathbb{S}}, \overline{\mathbb{S}}$) ;

L'ensemble de pavés $\Delta\mathbb{S} = \overline{\mathbb{S}} \setminus \underline{\mathbb{S}}$ représente l'incertitude sur la caractérisation de l'ensemble solution $\mathbb{S}$, il contient les pavés de tailles plus petites que η. Le volume de cet ensemble décroît quand ε diminue.

Exemple 5 : Soit $\mathbb{S}$ l'ensemble des vecteurs $\mathbf{x} = (x_1, x_2)^T$ de $\mathbb{R}^2$ vérifiant :

$$f(\mathbf{x}) = x_1^2 + x_2^2 \in [0, 0.1]$$

La caractérisation de l'ensemble $\mathbb{S}$ est un problème d'inversion ensembliste. Soit $\mathbb{X} = [-10, 10]^2$ le domaine initial de recherche. SIVIA génère en 10ms sur un Pentium IV, 2 GHz l'ensemble de pavés tracés sur la figure 2. Les pavés gris clair sont faisables, ceux en gris foncé sont indéterminés.

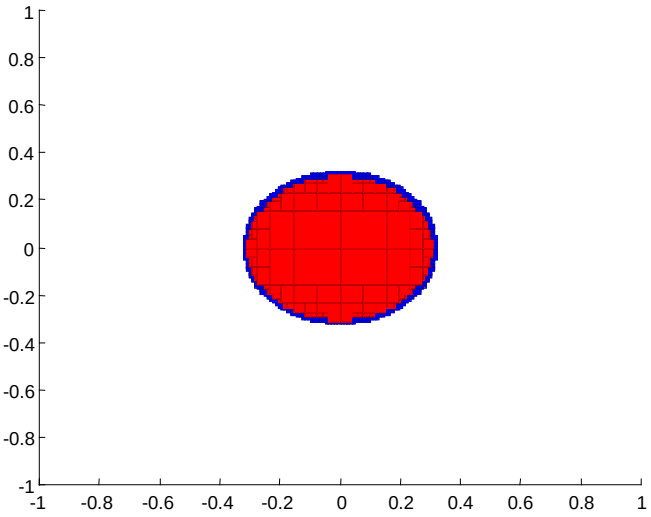

Figure 2 : Ensemble de pavés générés par SIVIA pour l'exemple 5.

La complexité de SIVIA est exponentielle par rapport à la dimension du vecteur des variables. Le nombre de bissections effectuées [JW93b] est inférieur à :

$$Nb = \left(\frac{w([\mathbf{x}_0])}{\eta} + 1 \right)^n \tag{23}$$

où $[\mathbf{x}_0]$ est le pavé initial de recherche et n est la dimension du vecteur $\mathbf{x}$. On verra dans la section 4 que l'utilisation de contracteurs permet de réduire le nombre de bissections nécessaires à la résolution d'un problème d'inversion ensembliste.

3.4. Stratégies de bissection

On considère le problème d'inversion ensembliste suivant :

$$\mathbb{S} = \{\mathbf{x} \in \mathbb{X} \mid f(\mathbf{x}) \in \mathbb{Y}\} = f^{-1}(\mathbb{Y}) \cap \mathbb{X} \tag{24}$$

avec $f : \mathbb{X} \subset \mathbb{R}^n \to \mathbb{R}$ et $\mathbb{Y} \subset \mathbb{R}$. Comme il a été signalé dans les sections précédentes, la résolution du problème d'inversion ensembliste (24), lorsque la fonction f est non linéaire, peut être effectuée en utilisant l'algorithme SIVIA [JW93b]. Cet algorithme, de type *branch-and-bound* est basé sur le partitionnement de l'ensemble initial de recherche $\mathbb{X}$. Dans ce

paragraphe seront présentées les principales stratégies de bissection permettant de bissecter efficacement les pavés étudiés. Dans la suite, on supposera que le pavé $[\mathbf{x}] \subset \mathbb{X}$ est à bissecter.

Les quatre stratégies les plus utilisées sont basées sur le choix d'une fonction de mérite $D(i)$ [CR97]. La direction de bissection k est alors donnée par :

$$k = \arg\left(\max_{i \in \{1,\ldots,n\}} \left(D(i) \right) \right) \tag{25}$$

Stratégie A : Le choix de la direction de bissection est basé sur la taille du pavé à partitionner [Moo66] [RR88].

$$D(i) = w\left([x_i]\right) \tag{26}$$

Ce choix de bissection est motivé par le fait que lorsqu'un pavé est partitionné d'une manière uniforme, la taille des intervalles générés converge uniformément vers zéro.

Stratégie B : Cette stratégie a été proposée par Hansen [Han92]. Elle consiste à choisir la direction qui maximise la quantité suivante :

$$\begin{aligned} q_i = \Bigg\{ & \max_{t \in [x_i]} \left(f\left(m([x_1]), \ldots, m([x_{i-1}]), t, m([x_{i+1}]), \ldots, m([x_n])\right)\right) \\ & - \min_{t \in [x_i]} \left(f\left(m([x_1]), \ldots, m([x_{i-1}]), t, m([x_{i+1}]), \ldots, m([x_n])\right)\right) \Bigg\} \end{aligned} \tag{27}$$

La quantité q_i donne une indication sur la variation de la fonction f lorsque t varie dans $[x_i]$. Le calcul de q_i est en général très compliqué en l'absence de propriétés de monotonicité. Pour faciliter ces calculs, on utilise souvent la fonction de mérite suivante :

$$D(i) = w\left(\left[f_i' \right]([\mathbf{x}])\right) \cdot w\left([x_i]\right) \tag{28}$$

où $\left[f_i' \right]$ est le $i^{ème}$ élément de la fonction d'inclusion du gradient de la fonction f. On trouve une excellente étude des propriétés de convergence de cette méthode dans [Han92].

Stratégie C : Cette stratégie a été proposée par Ratz [Rat92], Elle consiste à trouver la direction de bissection permettant de minimiser la taille de $[f]([\mathbf{x}])$. En utilisant la fonction d'inclusion centrée définie dans la section 2.3.3., on obtient :

$$\forall\ \mathbf{x} \in [\mathbf{x}] \ \Rightarrow\ f(\mathbf{x}) \in [f_c]([\mathbf{x}]) = [f](\hat{\mathbf{x}}) + \left(\left[\mathbf{f}'\right]([\mathbf{x}])\right)([\mathbf{x}] - \hat{\mathbf{x}})$$

avec $\hat{\mathbf{x}} \in [\mathbf{x}]$, dans la suite de ce paragraphe on choisit $\hat{\mathbf{x}} = m([\mathbf{x}])$. La taille de l'image de $[\mathbf{x}]$ en utilisant la forme centrée est :

$$\begin{aligned} w([f_c]([\mathbf{x}])) &= w\left([f](\hat{\mathbf{x}}_0) + \left([\mathbf{f}']([\mathbf{x}])\right)([\mathbf{x}] - \hat{\mathbf{x}})\right) \\ &\approx w\left(\left([\mathbf{f}']([\mathbf{x}])\right)([\mathbf{x}] - \hat{\mathbf{x}})\right) \\ &= \sum_{i=1}^{n} w\left(\left([f_i']([\mathbf{x}])\right)([x_i] - \hat{x}_i)\right) \end{aligned}$$

En minimisant le pessimisme de $\left([f_i']([\mathbf{x}])\right)([x_i] - \hat{x}_i)$, on arrive à minimiser celui de la fonction d'inclusion centrée. Par la suite, la stratégie de bissection **C** est basée sur l'utilisation de la fonction de mérite suivante :

$$D(i) = w\left[\left([f_i']([\mathbf{x}])\right)([x_i] - \hat{x}_i)\right] \tag{29}$$

Stratégie D : elle est analogue à la stratégie **A**. La fonction de mérite est donnée par :

$$D(i) = \begin{cases} w([x_i]) & \text{si } 0 \in [x_i] \\ w([x_i]) / \min\{|x_i| \mid x_i \in [x_i]\} & \text{si } 0 \notin [x_i] \end{cases} \tag{30}$$

L'utilisation de cette stratégie de bissection permet dans certains cas de réduire le pessimisme de $([f]([\mathbf{x}]) - f([\mathbf{x}]))$ dû aux erreurs d'arrondi [CR97]. Dans la littérature, elle est très rarement utilisée.

Dans la littérature, les stratégies **A** et **C** sont les plus utilisées ; généralement, elles sont plus efficaces que la stratégie **B** [CR97]. La stratégie **A** nécessite seulement le calcul de la taille du pavé [**x**] et sa convergence est en général plus rapide que les autres. Les stratégies **B** et **C** nécessitent le calcul du gradient de la fonction *f*.

4. Contracteurs

4.1. Problèmes de satisfaction de contraintes

Définition 5 : On considère le système d'équations à résoudre :

$$\mathbf{f}(\mathbf{x}) = \mathbf{0} \Leftrightarrow \begin{cases} f_1(x_1, x_2, \ldots, x_n) = 0 \\ \vdots \\ f_j(x_1, x_2, \ldots, x_n) = 0, \\ \vdots \\ f_m(x_1, x_2, \ldots, x_n) = 0 \end{cases} \qquad n \le m \tag{31}$$

avec $\mathbf{f}: \mathbb{X} \subset \mathbb{R}^n \to \mathbb{R}^m$, $\mathbf{x} = (x_1, x_2 \ldots x_N)^T$ est le vecteur des variables et enfin $\mathbb{X}$ est le domaine initial des variables.

On appelle *problème de satisfaction de contraintes* (CSP) *H*, la recherche de l'ensemble de toutes les solutions du système d'équations (31) contenues dans l'espace de recherche initial $\mathbb{X}$. Dans la suite de ce chapitre, le (CSP) sera noté par :

$$H: \left(\mathbf{f}(\mathbf{x}) = 0,\ \mathbf{x} \in \mathbb{X}\right) \tag{32}$$

Le CSP (31) est composé d'un ensemble de contraintes ($f_j(x_1, x_2, \ldots, x_n) = 0$) ; on note par $\mathbb{S}_j$ l'ensemble solution de la contrainte $f_j(x_1, x_2, \ldots, x_n) = 0$. La solution générale du CSP (31) est donnée par l'intersection des ensembles solutions de chaque contrainte composant le CSP :

$$\mathbb{S} = \bigcap_{j=1}^{n} \mathbb{S}_j \tag{33}$$

La résolution de (19) est un problème dit NP-difficile, c'est-à-dire qu'il n'existe aucun algorithme de complexité polynomiale permettant de le résoudre. La résolution de ce problème par l'intermédiaire de l'algorithme SIVIA associé aux techniques de partitionnement présentées dans la section 3.4. est limitée aux cas où le nombre de variables est assez réduit (deux ou trois) étant donné que sa complexité est exponentielle. Ces limitations ont donné lieu au développement d'autres outils, appelés contracteurs, basés sur des techniques de consistance [Wal75] [Cle87] [BMH94] permettant de réduire les domaines des variables tout en limitant le recours aux bissections, ainsi qu'à l'utilisation d'autres contracteurs à points fixes tels que celui de Newton ou de Krawczyk [Moo79] [Neu90] [Han92].

Dans les sections suivantes, les contracteurs les plus utilisés seront présentés.

4.2. Consistance

On suppose qu'on a une seule contrainte H_j : $f_j(x_1, x_2, \ldots, x_n) = 0$ et que les domaines initiaux des variables $x_1, x_2, \ldots, x_n$ sont respectivement $[x_1], [x_2], \ldots, [x_n]$.

Une valeur $\hat{x}_i \in [x_i]$ est dite consistante avec H_j s'il existe des valeurs $\hat{x}_j \in [x_j],\ j \in \{1, 2, \ldots, i-1, i+1, \ldots, n\}$, telles que la contrainte $f_j(\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_n) = 0$ soit satisfaite. La valeur $\hat{x}_i$ appartient donc à la projection sur x_i de $\mathbb{S}_j \cap [\mathbf{x}]$ notée par $\pi_i(\mathbb{S}_j \cap [\mathbf{x}])$ [Bra02] [JKBW01]. L'intervalle $[x_i]$ est consistant avec H_j si toute valeur de $[x_i]$ est consistante avec H_j, donc :

$$[x_i] = \pi_i(\mathbb{S}_j \cap [\mathbf{x}]) \tag{34}$$

Considérons maintenant un ensemble de contraintes données par le CSP (32), alors $[x_i]$ est dit globalement consistant avec le CSP (32) si et seulement si toute valeur de $[x_i]$ est consistante avec toutes les contraintes du CSP, alors :

$$[x_i] = \pi_i(\mathbb{S} \cap [\mathbf{x}]) = \pi_i\left(\bigcap_{j \in \{1,\ldots,n\}} \mathbb{S}_j \cap [\mathbf{x}] \right) \tag{35}$$

En général, il est très difficile de vérifier la consistance globale et il est plus facile d'étudier la consistance locale. L'intervalle $[x_i]$ est alors dit localement consistant avec le CSP (32) si :

$$[x_i] = \bigcap_{j \in \{1,\ldots,n\}} \pi_i\left(\mathbb{S}_j \cap [\mathbf{x}]\right) \tag{36}$$

Sur la figure 3, nous illustrons le principe de consistance sur un exemple comprenant, deux contraintes dont les solutions sont respectivement $\mathbb{S}_1$ et $\mathbb{S}_2$, et deux variables x_1 et x_2.

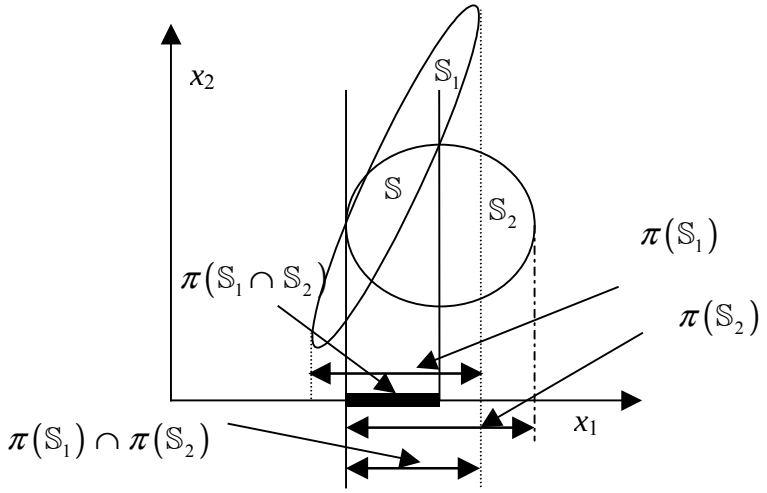

Figure 3 : Principe de la consistance locale et globale

4.3. Définition et propriétés des contracteurs

Un contracteur est un opérateur $\mathcal{C}$ permettant de contracter le domaine de recherche initial du vecteur **x** sans perdre de solution et sans faire aucune bissection. ♦

On considère le CSP (32), alors :

$$\mathbb{S} \subseteq \mathcal{C}(\mathbb{X}) \subseteq \mathbb{X} \tag{37}$$

où $\mathbb{X}$ est le vecteur des variables, $\mathbb{X}$ est le domaine initial de **x** et $\mathbb{S}$ est l'ensemble des solutions. L'intérêt de l'utilisation d'un contracteur est alors d'éliminer certaines parties

inconsistantes du domaine initial $\mathbb{X}$. Ainsi, le contracteur est dit optimal si toutes les parties inconsistantes sont éliminées [Bra02] ; dans ce cas, le pavé généré est minimal.

Un contracteur possède les propriétés suivantes [BG97] :

$$\forall [\mathbf{x}] \subseteq \mathbb{X}, \quad \mathcal{C}([\mathbf{x}]) \subseteq [\mathbf{x}] \quad \text{(contractance)} \tag{38}$$

$$\forall [\mathbf{x}] \subseteq \mathbb{X}, \quad [\mathbf{x}] \cap \mathbb{S} \subseteq \mathcal{C}([\mathbf{x}]) \quad \text{(complétude)} \tag{39}$$

Un contracteur est dit monotone si :

$$\forall([\mathbf{x}], [\mathbf{y}]) \text{ tel que : } [\mathbf{x}] \subseteq [\mathbf{y}] \Rightarrow \mathcal{C}([\mathbf{x}]) \subseteq \mathcal{C}([\mathbf{y}]) \tag{40}$$

4.4. Cas linéaire

On considère un système d'équations donné par :

$$[\mathbf{A}]\mathbf{x} = [\mathbf{b}] \tag{41}$$

avec $[\mathbf{A}] \in \mathbb{IR}^{n \times n}$ une matrice carrée dont les éléments sont des intervalles et $[\mathbf{b}] \in \mathbb{IR}^{n}$.

L'ensemble de toutes les solutions de (41) est donné par

$$\mathbb{S} = \left\{ \mathbf{x} \in \mathbb{R}^n \mid \exists \mathbf{A} \in [\mathbf{A}], \exists \mathbf{b} \in [\mathbf{b}], \ \mathbf{A}\mathbf{x} = \mathbf{b} \right\} \tag{42}$$

On présente dans la suite plusieurs méthodes permettant de donner une approximation extérieure de l'ensemble $\mathbb{S}$. Ces méthodes sont basées sur l'extension des algorithmes classiques comme celui d'élimination de Gauss ou de Gauss-Seidel.

4.4.1. Matrice $[\mathbf{A}]$ ponctuelle

On considère le cas où la matrice $[\mathbf{A}]$ est ponctuelle, par suite :

$$\mathbf{A}\mathbf{x} = [\mathbf{b}]$$

Lorsque la matrice **A** est non singulière, la contraction maximale est donnée par :

$$[\mathbf{x}] = \mathbf{A}^{-1}[\mathbf{b}]$$

Dans ce cas un contracteur optimal est donné par l'expression suivante :

$$C_{\mathrm{d}} : [\mathbf{x}] \mapsto \left(\mathbf{A}^{-1}[\mathbf{b}]\right) \cap [\mathbf{x}] \tag{43}$$

En général, la condition d'inversibilité de la matrice ponctuelle **A** n'est pas suffisante. Le choix d'une matrice **A** mal conditionnée conduit en général à des problèmes numériques donnant lieu à un pavé [**x**] très large. Il est donc nécessaire de bien conditionner le problème.

Exemple 6 : on considère l'équation $\mathbf{Ax}=[\mathbf{b}]$ avec

$$\mathbf{A}=\begin{pmatrix}1 & 2\\ 2 & 1\end{pmatrix},\ \mathbf{x}=\begin{pmatrix}x_1\\ x_2\end{pmatrix},\ \mathbf{b}=\begin{pmatrix}[0.3,1.9]\\ [0.9,2]\end{pmatrix}\ \text{et}\ [\mathbf{x}]=\begin{pmatrix}[-5,5]\\ [-5,5]\end{pmatrix}$$

d'où

$$C_{\mathrm{d}}([\mathbf{x}])=\left(\mathbf{A}^{-1}[\mathbf{b}]\right)\cap[\mathbf{x}]=\begin{pmatrix}[0.5,0.7]\\ [-0.1,0.6]\end{pmatrix}$$

Dans cet exemple, étant donné que A est inversible, la contraction est maximale. Le pavé $[\mathbf{x}]$ constitue une approximation extérieure de l'ensemble solution tracé sur la figure 4.

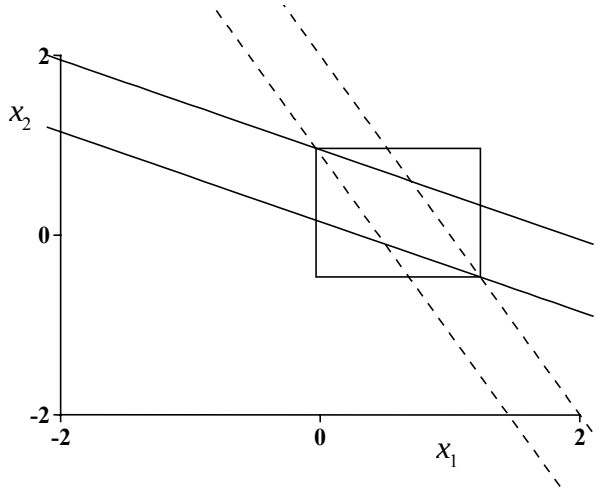

Figure 4 : Encadrement de l'ensemble des solutions faisables de l'exemple 6

Remarque 2 : Lorsque la matrice **A** est singulière, l'ensemble des valeurs faisables peut être non borné ou vide. Ceci est dû à la présence de deux ou plusieurs régions linéairement dépendantes dans l'espace des solutions. Lorsque ces régions sont strictement parallèles, l'ensemble des solutions est vide, il est non borné lorsqu'il y a intersection.

4.4.2. Méthode d'élimination de Gauss

La méthode d'élimination de Gauss par intervalles est une extension de la méthode classique se basant sur la décomposition LU. Le principe de cette méthode est de décomposer (lorsqu'il est possible) la matrice **A** en un produit de deux matrices **L** (matrice triangulaire inférieure avec tous les éléments diagonaux sont égaux à 1) et **U** (matrice triangulaire supérieure).

En utilisant cette décomposition, l'équation $\mathbf{Ax}=\mathbf{b}$ peut s'écrire comme suit :

$$\mathbf{LUx}=\mathbf{b} \tag{44}$$

Le contracteur d'élimination de Gauss $\mathcal{C}_{\mathrm{EG}}$ est obtenu en remplaçant dans la méthode de Gauss les variables ponctuelles par des variables intervalles correspondantes. Il est alors donné par les expressions (45) et (46) [JKDW01].

$$\begin{aligned} &[y_1]=[b_1] \\ &[y_i]=[b_i]-\sum_{j=1}^{i-1}[l_{ij}][y_j], \text{ pour } i=2,\ldots,n \end{aligned} \tag{45}$$

et

$$\begin{aligned} &[x_n]=([y_n]/[u_{nn}])\cap[x_n] \\ &[x_i]=\left(\left([y_i]-\sum_{j=i+1}^{n}[u_{ij}][x_j]\right)/[u_{ii}]\right)\cap[x_i], \text{ pour } i=n-1,\ldots,1 \end{aligned} \tag{46}$$

Dans les relations récurrentes (45) et (46), les coefficients l_{ij} et a_{ij} sont utilisés plusieurs fois pour calculer les éléments y_k et x_k $\{k = 1, 2,\ldots, n\}$. La multi-occurrence de ces coefficients engendre un pessimisme lors de la contraction du domaine du vecteur **x**. Le contracteur C_{EG} n'est efficace que sous certaines conditions : existence de la décomposition LU, la taille des éléments de la matrice $[\mathbf{A}]$ doit être petite, les éléments diagonaux de la matrice $[\mathbf{U}]$ ne doivent pas contenir de zéro.

Remarque 3 : Le pré-conditionnement du système d'équations (41) permet généralement d'obtenir de meilleurs résultats. Il se fait en multipliant les deux membres de (41) par la matrice ponctuelle $\hat{\mathbf{A}}=\left(mid\left([\mathbf{A}]\right)\right)^{-1}$. Le nouveau système à résoudre est le suivant :

$$\left[\hat{\mathbf{A}}\right]\mathbf{x}=\left[\hat{\mathbf{b}}\right] \tag{47}$$

avec :

$$\left[\hat{\mathbf{A}}\right]=\hat{\mathbf{A}}[\mathbf{A}] \text{ et } \left[\hat{\mathbf{b}}\right]=\left[\hat{\mathbf{A}}\right][\mathbf{b}]$$

4.4.3. Méthode de Gauss-Seidel

L'extension de la méthode itérative de Gauss-Seidel constitue une autre alternative [Han92]. Considérons le système préconditionné $\left[\hat{\mathbf{A}}\right]\mathbf{x}=\left[\hat{\mathbf{b}}\right]$, l'extension de la méthode de Gauss-Seidel aux intervalles est basée sur la relation itérative suivante [Han92] [KHN91] [Neu90] :

$$\left[x_i^{(k)}\right] = \left(\left(\left[\hat{b}_i\right] - \sum_{j=1}^{i-1}\left[\hat{a}_{ij}\right]\left[x_j^{(k-1)}\right] - \sum_{j=i+1}^{n}\left[\hat{a}_{ij}\right]\left[x_j^{(k-1)}\right]\right)/\left[\hat{a}_{ii}\right]\right) \cap \left[x_i^{(k-1)}\right] \tag{48}$$

Le contracteur de Gauss-Seidel est donc donné par :

$$\mathcal{C}_{\mathrm{GS}}: [x_i] \mapsto \left(\left(\left[\hat{b}_i\right] - \sum_{j=1}^{i-1}\left[\hat{a}_{ij}\right]\left[x_j\right] - \sum_{j=i+1}^{n}\left[\hat{a}_{ij}\right]\left[x_j\right]\right)/\left[\hat{a}_{ii}\right]\right) \cap [x_i] \tag{49}$$

A noter que comme pour les autres contracteurs, $\mathcal{C}_{\mathrm{GS}}$ est appliqué tant que les réductions du domaine $[\mathbf{x}]$ sont significatives.

4.5. Cas non linéaire

On considère la résolution du CSP suivant:

$$\mathrm{CSP}: \mathbf{f}(\mathbf{x}) = \mathbf{0}, \tag{50}$$

où $\mathbf{f}: \mathbb{R}^n \to \mathbb{R}^m$ est une fonction non linéaire et $\mathbf{x} \in [\mathbf{x}]$.

Plusieurs contracteurs ont été proposés pour résoudre des CSP non-linéaires. Les contracteurs de Krawczyk [Neu90] et de Newton [Han92] sont basés sur des linéarisations garanties du CSP (50). Ils sont applicables lorsque le CSP (50) est composé d'autant de contraintes que de variables. Généralement ces deux contracteurs sont efficaces seulement lorsque les domaines de recherche sont assez petits. Le contracteur *propagation rétropropagation* permet en général de résoudre (50) indépendamment de ces dernières conditions.

4.5.1. Contracteur de Krawczyk

Ce contracteur a été développé indépendamment dans [Kah68] et [Kra69] et a été étudié d'une manière approfondie dans [Moo79] et [Neu90]. Supposons que la fonction **f** est différentiable, le CSP (50) peut alors s'écrire sous la forme :

$$\Psi(\mathbf{x}) = \mathbf{x} - \mathbf{M}\mathbf{f}(\mathbf{x}) = \mathbf{x} \tag{51}$$

où M est une matrice ponctuelle de pré-conditionnement. Généralement, on choisit $\mathbf{M} = \mathbf{J}_{\mathbf{f}}^{-1}(\hat{\mathbf{x}})$, où $\hat{\mathbf{x}}$ est en général le centre du pavé $[\mathbf{x}]$ et $\mathbf{J}_{\mathbf{f}}$ est le Jacobien de **f**. La fonction d'inclusion centrée de Ψ est donnée par :

$$[\Psi]([\mathbf{x}]) = \Psi(\hat{\mathbf{x}}) + [\mathbf{J}_{\Psi}]([\mathbf{x}])([\mathbf{x}] - \hat{\mathbf{x}}) \tag{52}$$

où $[\mathbf{J}_{\Psi}]$ est une fonction d'inclusion du Jacobien de la fonction Ψ.

Le contracteur de Krawczyk est donc défini par :

$$\mathcal{C}_K : [\mathbf{x}] \mapsto \left(\Psi(\hat{\mathbf{x}}) + [\mathbf{J}_\Psi]([\mathbf{x}])([\mathbf{x}] - \hat{\mathbf{x}})\right) \cap [\mathbf{x}] \tag{53}$$

En remplaçant $\Psi(\mathbf{x})$ par $\mathbf{x} - \mathbf{M}\mathbf{f}(\mathbf{x})$, on obtient :

$$\mathcal{C}_K : [\mathbf{x}] \mapsto \left(\hat{\mathbf{x}} - \mathbf{M}\mathbf{f}(\hat{\mathbf{x}}) + \left(\mathbf{I} - \mathbf{M}[\mathbf{J}_\mathbf{f}]([\mathbf{x}])\right)([\mathbf{x}] - \hat{\mathbf{x}})\right) \cap [\mathbf{x}] \tag{54}$$

4.5.2. Contracteur de Newton

Le contracteur de Newton a été étudié et présenté dans plusieurs publications (voir par exemple [Moo79 ; Han92]). On suppose que la fonction **f** est différentiable, le CSP (50) peut être réécrit sous la forme suivante :

$$\Psi(\mathbf{x}) = \mathbf{x} - \mathbf{J}_\mathbf{f}^{-1}\mathbf{f}(\mathbf{x}) = \mathbf{x} \tag{55}$$

Une première version du contracteur de Newton est alors donnée par :

$$\mathcal{C}_N : [\mathbf{x}] \mapsto \left([\mathbf{x}] - [\mathbf{J}_\mathbf{f}]^{-1}\mathbf{f}([\mathbf{x}])\right) \cap [\mathbf{x}] \tag{56}$$

Une seconde version du contracteur de Newton plus efficace lorsque les domaines des variables sont petits [Han92] est obtenue en utilisant le théorème de la valeur moyenne:

$$\begin{aligned} &\mathbf{f}(\hat{\mathbf{x}}) + \mathbf{J}_\mathbf{f}(\xi)(\mathbf{x} - \hat{\mathbf{x}}) = \mathbf{0} \\ &\mathbf{x} \in [\mathbf{x}],\ \xi \in [\mathbf{x}] \end{aligned} \tag{57}$$

La fonction d'inclusion de la forme moyenne de **f** donnée par (57) est :

$$\mathbf{f}(\hat{\mathbf{x}}) + [\mathbf{J}_\mathbf{f}]([\mathbf{x}])([\mathbf{x}] - \hat{\mathbf{x}}) = \mathbf{0} \tag{58}$$

Le contracteur de Newton consiste donc à résoudre le CSP linéarisé (58) en utilisant le contracteur de Gauss-Seidel présenté pour le cas des CSPs linéaires.

4.5.3. Contraction par projection

4.5.3.1. Sous-résolveurs et contraction par inversion explicite

On considère un CSP $H : (\mathbf{f}(\mathbf{x}) = 0,\ \mathbf{x} \in [\mathbf{x}])$, on appelle sous-résolveur, noté ϕ_i, un algorithme permettant de calculer la valeur d'une composante x_i du vecteur **x** en fonction des autres composantes supposées connues [Bra02].

Théorème 1 [Bra02] : supposons qu'il existe un sous-résolveur ϕ_i d'entrée ${}^i\mathbf{x}=(x_1 \dots x_{i-1}, x_{i+1}, \dots x_n)$ et de sortie x_i associé au CSP $H:(\mathbf{f}(\mathbf{x}) = 0, \mathbf{x} \in [\mathbf{x}])$. Soit $[\phi_i]$ une fonction d'inclusion pour ϕ_i, alors

$$\pi_i(\mathbb{S} \cap [\mathbf{x}]) \subseteq [\phi_i]([{}^i\mathbf{x}]) \cap [x_i] \tag{59}$$

où π_i est l'opérateur de projection sur $[x_i]$ défini dans la section 4.2 ; on trouve la démonstration dans [Bra02]. Si en plus, la fonction d'inclusion $[\phi_i]$ est minimale, alors l'inclusion (59) devient une égalité.

Supposons maintenant qu'il est possible de trouver, à l'aide du calcul formel, une expression explicite $x_i = \phi_i({}^i\mathbf{x})$ pour chacune des variables x_i, on obtient alors un ensemble de n sous-résolveurs. Si on dispose d'une fonction d'inclusion pour chacune des fonctions ϕ_i, alors on peut construire le contracteur suivant :

$$\mathcal{C} : [\mathbf{x}] \mapsto \begin{pmatrix} [\phi_1]([{}^1\mathbf{x}]) \\ [\phi_2]([{}^2\mathbf{x}]) \\ \vdots \\ [\phi_n]([{}^n\mathbf{x}]) \end{pmatrix} \cap [\mathbf{x}] \tag{60}$$

Le contracteur défini par (60) est alors optimal si les fonctions d'inclusions $[\phi_i]$ sont toutes minimales, i.e. le nombre d'occurrences de chacune des variables ne dépasse pas un.

Exemple 7 : On considère la contrainte suivante

$$H : \left(x_1 \exp(x_2) + x_3 \exp(x_4) + \log(x_5) = 0, \ \mathbf{x} \in [\mathbf{x}]\right)$$

On peut alors facilement obtenir les sous-résolveurs suivants :

$$\begin{cases} \phi_1 : x_1 \mapsto -\dfrac{x_3 \exp(x_4) + \log(x_5)}{\exp(x_2)} \\ \phi_2 : x_2 \mapsto \log\left(-\dfrac{x_3 \exp(x_4) + \log(x_5)}{x_1}\right) \\ \phi_3 : x_3 \mapsto -\dfrac{x_1 \exp(x_2) + \log(x_5)}{\exp(x_4)} \\ \phi_4 : x_4 \mapsto \log\left(-\dfrac{x_1 \exp(x_2) + \log(x_5)}{x_3}\right) \\ \phi_5 : x_5 \mapsto \exp\left(-x_1 \exp(x_2) - x_3 \exp(x_4)\right) \end{cases}$$

On obtient alors dans ce cas un contracteur optimal en utilisant les fonctions d'inclusions naturelles des sous-résolveurs ϕ_i qui sont minimales. En revanche, les fonctions d'inclusions naturelles sont très souvent pessimistes ; ce pessimisme est dû principalement à l'effet de dépendance, il est alors souhaitable d'utiliser d'autres fonctions d'inclusions comme la forme centrée ou les fonctions de Taylor d'ordres élevés.

4.5.3.2. Décomposition en contraintes primitives

Nous avons vu dans le paragraphe précédent que le contracteur obtenu par projection des contraintes sur chacune des variables x_i est optimal lorsque les fonctions d'inclusions des sous-résolveurs utilisés sont minimales. Néanmoins, la construction de ce contracteur nécessite l'inversion explicite de la contrainte principale du CSP par rapport à chacune des variables x_i ; dans certains cas, le calcul formel ne nous permet pas de disposer de ces sous-résolveurs (ou bien les sous-résolveurs obtenus sont très complexes et leurs fonctions d'inclusions sont alors trop pessimistes).

Afin d'éviter de calculer ces sous-résolveurs, un autre contracteur, appelé *Propagation – Rétropropagation*, noté par $\mathcal{C}_{\downarrow\uparrow}$, a été développé [JKBW01] [Bra02]. Le principe de ce contracteur est inspiré de l'algorithme de Waltz [Wal75] [Cle87][Dav87]. Pour ce contracteur, on ne calcule plus des solutions explicites mais on décompose la contrainte (ou les contraintes) du CSP en un ensemble de contraintes élémentaires. Une contrainte élémentaire ne contient qu'une opération arithmétique telle que $\{+, -, *, /\}$ entre deux variables ou une fonction élémentaire comme {exp, log…}. Un ensemble de variables intermédiaires est alors introduit.

Exemple 8 :

On considère de nouveau l'exemple défini dans la section précédente :

$$H : \left(x_1 \exp(x_2) + x_3 \exp(x_4) + \log(x_5) = 0, \ \mathbf{x} \in [\mathbf{x}]\right)$$

La décomposition de H en un ensemble de contraintes primitives se fait en introduisant des variables intermédiaires dont les domaines initiaux sont $]-\infty, +\infty[$. L'ensemble de contraintes résultantes sont données par le système suivant :

$$\begin{cases} a_1 = \exp(x_2) \\ a_2 = x_1 a_1 \\ a_3 = \exp(x_4) \\ a_4 = x_3 a_3 \\ a_5 = a_2 + a_4 \\ a_6 = \log(x_5) \\ a_7 = a_5 + a_6 = 0 \end{cases}$$

Il s'agit donc de projeter le système de contraintes primitives sur chacune des variables. Cet ensemble de contraintes est représenté par le graphe suivant (figure 5) où chaque variable est représentée par un nœud et chaque contrainte primitive par un arc.

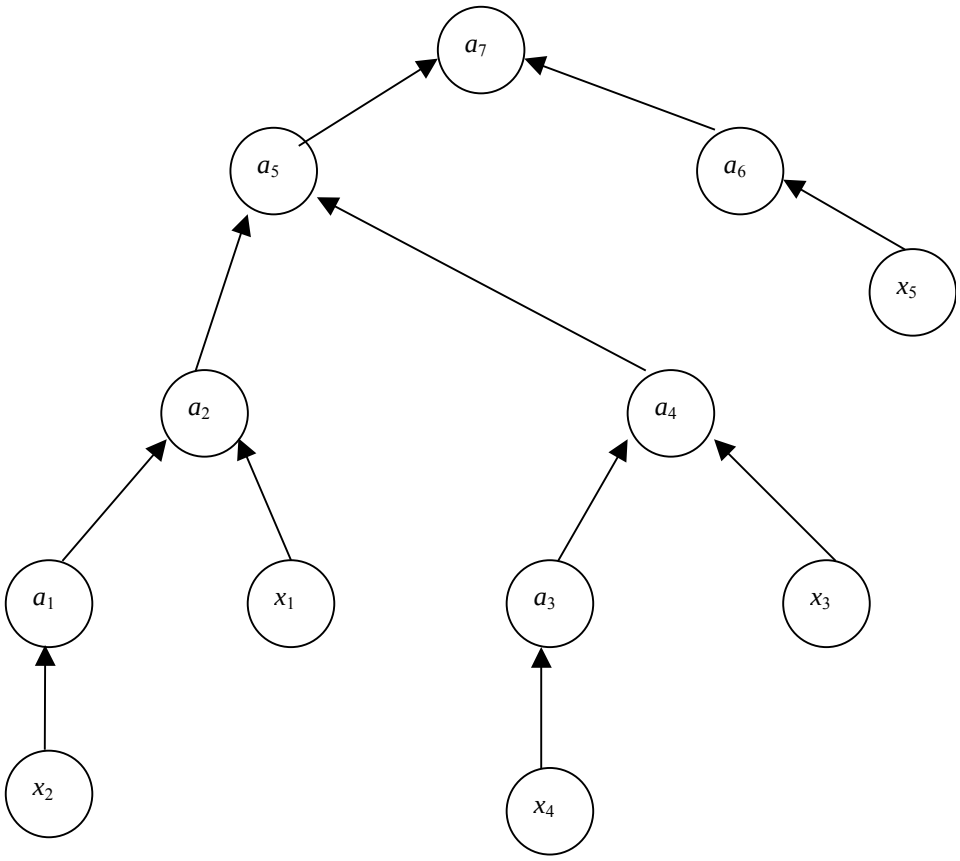

Figure 5 : Graphe associé au système de contraintes primitives de l'exemple 8

Lorsque le graphe est connexe et ne comporte pas de cycle, comme c'est le cas dans notre exemple, il est alors dit acyclique : le graphe est un arbre.

Le contracteur $\mathcal{C}_{\downarrow\uparrow}$ comporte deux phases : la première, dite *propagation*, consiste à projeter les contraintes primitives une à une tout en parcourant le graphe des feuilles à la racine (pour un arbre). La deuxième phase, dite *rétropropagation*, consiste à parcourir le graphe dans le sens contraire en projetant à tour de rôle les contraintes primitives.

Le principe de ce contracteur est illustré par l'exemple défini précédemment. Soient alors $[x_1]$, $[x_2]$, $[x_3]$, $[x_4]$, $[x_5]$, $[a_1]$, $[a_2]$, $[a_3]$, $[a_3]$, $[a_4]$, $[a_5]$, $[a_6]$ et $[a_7]$ les domaines respectifs des variables x_1, x_2, x_3, x_4, x_5, a_1, a_2, a_3, a_4, a_5, a_6 et a_7,. La phase de *propagation* se fait comme suit :

$$\begin{cases} [a_1] := \exp([x_2]) \cap [a_1] \\ [a_2] := ([x_1][a_1]) \cap [a_2] \\ [a_3] := \exp([x_4]) \cap [a_3] \\ [a_4] := ([x_3][a_3]) \cap [a_4] \\ [a_5] := ([a_2] + [a_4]) \cap [a_5] \\ [a_6] := \log([x_5]) \cap [a_6] \\ [a_7] := [a_5] + [a_6][a_7] \cap [a_7] \end{cases}$$

La phase de *propagation* permet de calculer le domaine de la variable intermédiaire a_7 consistant avec les domaines des autres variables. D'un autre coté a_7 est égal au singleton $\{0\}$, ce qui signifie que si l'intervalle $[a_7]$ ne contient pas zéro, le CSP n'aurait pas de solution dans le domaine de recherche initial. Si $[a_7]$ contient zéro, il sera alors remplacé par $\{0\}$.

La seconde phase, appelée *rétropropagation*, consiste à éliminer les parties des domaines des variables qui ne sont pas consistantes avec $[a_7] = 0$. La *rétropropagation* s'effectue comme suit :

$$\begin{cases} [a_7] := [a_7] \cap \{0\} \\ [a_6] := ([a_7] - [a_5]) \cap [a_6] \\ [a_5] := ([a_7] - [a_6]) \cap [a_5] \\ [x_5] := \exp([a_6]) \cap [x_5] \\ [a_2] := ([a_5] - [a_4]) \cap [a_2] \\ [a_4] := ([a_5] - [a_2]) \cap [a_4] \\ [x_3] := ([a_4] / [a_3]) \cap [x_3] \\ [a_3] := ([a_4] / [x_3]) \cap [a_3] \\ [x_4] := \log([a_3]) \cap [x_4] \\ [a_1] := ([a_2] / [x_1]) \cap [a_1] \\ [x_1] := ([a_2] / [a_1]) \cap [x_1] \\ [x_2] := \log([a_1]) \end{cases}$$

Dans l'exemple présenté, une *propagation* et une *rétropropagation* sont suffisantes pour avoir une contraction optimale du CSP. Ceci est toujours possible lorsque le graphe correspondant au CSP est un arbre. Néanmoins, lorsque le graphe n'est pas un arbre i.e. présence d'une variable multi-occurrente, la contraction n'est pas optimale et il est nécessaire d'effectuer le processus de *propagation – rétropropagation* autant de fois qu'une contraction significative est possible.

4.6. SIVIA avec contracteur

Une version de l'algorithme SIVIA avec contracteur [JKDW01] est donnée dans cette section. L'algorithme, noté SIVIAP, sera utilisé dans la suite de cette thèse étant donné ses propriétés intéressantes.

On considère le CSP suivant :

$$\text{CSP}: \mathbf{f}(\mathbf{x}) \in [\mathbf{y}]$$

Ce CSP peut être résolu en utilisant conjointement SIVIA (voir section 3.3) et un contracteur [JKDW01]. Ce dernier permet de réduire le domaine de recherche des variables et ainsi de limiter le nombre de bissections. La méthode est donnée par l'algorithme suivant :

Algorithme SIVIAP$\left(\text{e} : C, [\mathbf{x}], [\mathbf{f}], [\mathbf{y}], \eta \ ; \ \text{s}: \ \underline{\mathbb{S}}, \overline{\mathbb{S}} \right)$

1 $[\mathbf{x}] := \mathcal{C}([\mathbf{x}]);$

2 Si $[\mathbf{x}] = \varnothing$, rejeter $[\mathbf{x}]$;

3 Si $[\mathbf{f}]([\mathbf{x}]) \subseteq [\mathbf{y}]$, $\underline{\mathbb{S}} := \underline{\mathbb{S}} \cup [\mathbf{x}]$; $\overline{\mathbb{S}} := \overline{\mathbb{S}} \cup [\mathbf{x}]$;

4 Si $w([\mathbf{x}]) \leq \eta$, $\overline{\mathbb{S}} := \overline{\mathbb{S}} \cup [\mathbf{x}]$;

5 bissecter $[\mathbf{x}]$ en $([\mathbf{x}_1], [\mathbf{x}_2])$;

6 SIVIAP$\left(\text{e} : C, [\mathbf{x}_1], [\mathbf{f}], [\mathbf{y}], \eta \ ; \ \text{s}: \ \underline{\mathbb{S}}, \overline{\mathbb{S}} \right)$;

SIVIAP $\left(\text{e} : C, [\mathbf{x}_2], [\mathbf{f}], [\mathbf{y}], \eta \ ; \ \text{s}: \ \underline{\mathbb{S}}, \overline{\mathbb{S}} \right)$;

La différence entre SIVIA et SIVIAP vient du fait que le test d'inclusion sur un pavé $[\mathbf{x}]$ est remplacé par une phase de contraction qui consiste à éliminer des parties inconsistantes de $[\mathbf{x}]$. Si la contraction engendre un ensemble vide, alors $[\mathbf{x}]$ est rejeté.

Exemple 9 : Nous présentons dans cette section un exemple simple mettant en évidence l'intérêt de l'utilisation combinée de l'algorithme SIVIA avec un contracteur et également les limites de cette méthode dans un cas d'étude où les paramètres à estimer sont multi-occurrents. On considère le modèle suivant :

$$\varepsilon(\omega) = \varepsilon_{\infty} + \frac{\Delta\varepsilon}{\left(1+(j\omega\tau)^{\alpha}\right)^{\beta}}, \quad \text{avec } j^2 = -1 \tag{61}$$

où on propose d'estimer les paramètres τ, α, β, $\Delta\varepsilon$ et ε_{∞} en utilisant des mesures de la sortie ε. Pour évaluer une fonction d'inclusion de la sortie du modèle, on décompose ε en une partie réelle ε' et une partie imaginaire ε'' ; leurs expressions analytiques sont données par :

$$\varepsilon'(\omega) = \varepsilon_\infty + \Delta\varepsilon \frac{\cos(\beta\varphi)}{\left(1 + 2(\omega\tau)\sin\left(\frac{\pi(1-\alpha)}{2}\right) + (\omega\tau)^{2\alpha}\right)^{\beta/2}} \tag{62}$$

et

$$\varepsilon''(\omega) = \Delta\varepsilon \frac{\sin(\beta\varphi)}{\left(1 + 2(\omega\tau)^{\alpha}\sin\left(\frac{\pi(1-\alpha)}{2}\right) + (\omega\tau)^{2\alpha}\right)^{\beta/2}} \tag{63}$$

avec

$$\varphi = \arctan\left[\frac{(\omega\tau)^{\alpha}\cos\left(\frac{\pi(1-\alpha)}{2}\right)}{1 + (\omega\tau)^{\alpha}\sin\left(\frac{\pi(1-\alpha)}{2}\right)}\right]$$

D'autre part, on suppose que les paramètres à estimer doivent satisfaire les contraintes suivantes :

$$\begin{cases} \varepsilon_\infty > 1 \\ \Delta\varepsilon > 0 \\ \tau > 0 \\ \alpha \in]0;1] \\ \beta \in]0;1] \end{cases}$$

On remarque que les paramètres α, β et τ sont multi-occurrents, ce qui engendre donc, d'une part un pessimisme lors de l'évaluation de la fonction d'inclusion naturelle de (62) et (63), et d'autre part, le contracteur propagation-rétropropagation n'est pas optimal.

Les données utilisées sont obtenues en simulant le modèle (61) avec $\alpha = 1$, $\beta = 1$, $\Delta\varepsilon = 6$, $\varepsilon_\infty = 3$ et log(τ) = -6.443. La borne d'erreur *a priori* est prise égale à 1%. Les pseudo-mesures sont alors données par les intervalles suivants :

$$\varepsilon_j' = \left[0.99 \cdot \hat{\varepsilon}_j', 1.01 \cdot \hat{\varepsilon}_j'\right] \text{ et } \varepsilon_j'' = \left[0.99 \cdot \hat{\varepsilon}_j'', 1.01 \cdot \hat{\varepsilon}_j''\right]$$

En utilisant SIVIAP et SIVIA, on génère deux ensembles de pavés qui contiennent toutes les solutions compatibles avec le modèle et avec les hypothèses sur le bruit de mesure. La projection de ces ensembles par rapport aux différents paramètres donne une approximation extérieure de toutes les solutions ; pour chaque paramètre on trouve un intervalle qui contient d'une manière garantie sa valeur exacte. Les deux versions de SIVIA avec et sans contracteur donnent, pour $\eta = 0.01$, les intervalles suivants :

$$\alpha_1 \in [0.997,1],\ \beta_1 \in [0.9968,1],\ \tau_1 \in [15.742,\ 16.172] \times 10^{-4}\ \text{s},\ \Delta\varepsilon_1 \in [5.925,6.06],$$

$$\varepsilon_\infty \in [2.97,3.03].$$

Cependant, le temps de calcul nécessaire pour SIVIAP est de 21.9 secondes sur un Celeron 1GHz alors que SIVIA sans contracteur nécessite plusieurs heures ; ceci montre l'intérêt de SIVIAP lorsque le nombre de paramètres à estimer est élevé.

Par ailleurs, les fonctions d'inclusions naturelles des parties réelle et imaginaire de ε sont très pessimistes étant donné le nombre d'occurrences élevé pour les variables à estimer. Il est donc nécessaire d'utiliser d'autres fonctions d'inclusions moins pessimistes comme les formes centrées. Ces dernières nécessitent de calculer des Jacobiens (voire des dérivées d'ordres supérieurs). Ces dérivées seront calculées à l'aide de la différentiation automatique.

Enfin, le nombre d'occurrences assez élevé dans les parties réelle et imaginaire nous a conduit à nous diriger vers les intervalles complexes afin de travailler directement avec l'expression (61). Le chapitre 2 sera dédié aux intervalles complexes.

5. Différentiation automatique

Comme on l'a constaté dans ce chapitre, le calcul des dérivées de fonctions est souvent nécessaire. Par exemple, pour évaluer une fonction en utilisant une forme centrée, il est indispensable de calculer d'abord son Jacobien. L'évaluation du Jacobien est aussi nécessaire pour mettre en œuvre les contracteurs de Newton et de Krawczyk. La dérivée peut être calculée en utilisant des outils de calcul symbolique ; lorsque la fonction à dériver est de forme complexe, on obtient généralement une forme non utilisable à cause des nombres d'occurrences. Généralement, pour éviter le calcul symbolique, on calcule la dérivée d'une fonction en utilisant la différence centrée :

$$\frac{\partial f}{\partial x} \approx \frac{f(x+\Delta x) - f(x-\Delta x)}{2\Delta x} = \frac{\partial f}{\partial x} + O(\Delta x^2) \tag{64}$$

où $f : \mathbb{R} \to \mathbb{R}$ et $O(\Delta x^2)$ représente l'erreur de troncature.

Dans certains cas, le résultat obtenu n'est pas précis à cause de la propagation des erreurs de troncature. La différentiation automatique, développée indépendamment par Moore [Moo62] et Wengert [Wen64] représente une alternative fiable aux méthodes citées ci-dessus. Cette méthode est efficace lorsque les fonctions étudiées comportent des variables données par des intervalles. Dans la plupart des cas, l'évaluation de la dérivée en utilisant la différentiation automatique est moins pessimiste que celle obtenue par des calculs symboliques.

La différentiation automatique (DA) est basée sur les propriétés élémentaires de dérivation. Soient u, v et $w : \mathbb{R}^n \to \mathbb{R}$ et notons par $\nabla u, \nabla v$ et ∇w leurs gradients, on définit les propriétés élémentaires de différentiation comme suit :

$$w = u + v \qquad \Rightarrow \qquad \nabla w = \nabla u + \nabla v$$

$$w = u - v \quad \Rightarrow \quad \nabla w = \nabla u - \nabla v$$

$$w = u \cdot v \quad \Rightarrow \quad \nabla w = \nabla u \cdot v + \nabla v \cdot u$$

$$w = u / v \quad \Rightarrow \quad \nabla w = \left(\nabla u \cdot v - \nabla v \cdot u\right) / v^2$$

Soit $\phi : \mathbb{R} \to \mathbb{R}$, alors

$$w = \phi(u) \quad \Rightarrow \quad \nabla w = \nabla u \cdot \phi'(u)$$

5.1. Code

Soit une fonction $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^m$ ne contenant que des opérations et des fonctions élémentaires (par exemple $\{+, -, *, /, \cos, \sin, \ldots\}$). Nous pouvons décomposer la fonction **f** en une liste d'opérations comme suit :

$$\begin{aligned} &\tau_i(\mathbf{x}) = g_i(\mathbf{x}) = x_i, && \text{pour } i = 1, 2, \ldots, n \\ &\tau_i = g_i\left(\tau_1(\mathbf{x}), \tau_2(\mathbf{x}), \ldots, \tau_{i-1}(\mathbf{x})\right), && \text{pour } i = n+1, \ldots, l \end{aligned} \tag{65}$$

où les τ_i et les g_i sont des fonctions scalaires contenant une seule opération arithmétique. La suite de fonctions définie par (65) est appelée *code*[1]. On note par a_i l'*arité* (nombre de dépendances) de la fonction g_i ; dans la suite a_i prendra une des valeurs $\{0,1,2\}$, i.e. g_i peut dépendre au maximum de deux variables .

Exemple 10 : On considère la fonction $f : \mathbb{R}^3 \to \mathbb{R}$ définie par :

$$f(x, y, z) = (xyz + \sin(z) + z)$$

La fonction *f* peut être réécrite en utilisant la suite des fonctions donnée par (65). On obtient alors :

$$\begin{aligned} \tau_1 &= x \\ \tau_2 &= y \\ \tau_3 &= z \\ \tau_4 &= \tau_1 \tau_2 \\ \tau_5 &= \tau_4 \tau_3 \\ \tau_6 &= \sin(\tau_3) \\ \tau_7 &= \tau_5 + \tau_6 \\ \tau_8 &= \tau_7 + \tau_3 \end{aligned}$$

Ainsi l'évaluation de la fonction en (x, y, z) est donnée par τ_8.

[1] Traduction de « code-list »

5.2. Différentiation

On considère une fonction différentiable $\mathbf{f} : \mathbb{R}^n \rightarrow \mathbb{R}^m$. On suppose que **f** est décomposable en une suite de fonction élémentaires (supposées dérivables) de la forme (65). En utilisant les propriétés de dérivation élémentaires, on obtient :

$$\frac{\partial \tau_i}{\partial \tau_j} = \delta_{ij} + \sum_{k=j}^{i-1} \frac{\partial g_i}{\partial \tau_k} \frac{\partial \tau_k}{\partial \tau_j} \tag{66}$$

où

$$\delta_{ij} = \begin{cases} 1 & \text{si } i = j \\ 0 & \text{si } i \neq j \end{cases}$$

On doit noter que, dans l'équation (66), le dernier terme de l'expression de droite n'est calculé que lorsque $i \neq j$.

La différentiation automatique est basée sur l'expression (66) et sur les règles élémentaires de dérivation. Son principe est donné dans l'exemple 11.

Exemple 11 : On considère la fonction *f* définie dans l'exemple 10 ; en effectuant la différentiation de chacune des variables intermédiaires, on obtient :

$$\begin{aligned}
t_1 &= x & \nabla t_1 &= [1,0,0] \\
t_2 &= y & \nabla t_2 &= [0,1,0] \\
t_3 &= z & \nabla t_3 &= [0,0,1] \\
t_4 &= t_1 t_2 & \nabla t_4 &= t_1 \nabla t_2 + t_2 \nabla t_1 = [t_2, t_1, 0] \\
t_5 &= t_4 t_3 & \nabla t_5 &= t_3 \nabla t_4 + t_4 \nabla t_3 = [t_3 t_2, t_3 t_1, t_4] \\
t_6 &= \sin(t_3) & \nabla t_6 &= \nabla t_3 \cos(t_3) = [0,0,\cos(t_3)] \\
t_7 &= t_5 + t_6 & \nabla t_7 &= \nabla t_5 + \nabla t_6 = [t_3 t_2, t_3 t_1, t_4 + \cos(t_3)] \\
t_8 &= t_7 + t_3 & \nabla t_8 &= \nabla t_7 + \nabla t_3 = [t_3 t_2, t_3 t_1, t_4 + \cos(t_3) + 1]
\end{aligned}$$

Le gradient de la fonction *f* est donné par ∇t_8. On remarque que l'évaluation des variables intermédiaires et de leurs gradients s'effectue simultanément. Ce mode de dérivation est appelé *différentiation directe* ; on trouve dans la littérature d'autres modes i.e. *différentiation indirecte* et *différentiation de Taylor* [RC96] [Ral81].

Enfin, on doit noter qu'en utilisant la différentiation automatique, aucune approximation n'est faite. De plus, on ne cherche pas à calculer symboliquement la dérivée.

6. Conclusion

Ce chapitre a été consacré aux différents outils dont l'utilisation est devenue classique dans le cadre de l'analyse par intervalles. Dans la première partie, nous avons présenté un bref rappel de l'arithmétique d'intervalles ainsi que des principaux problèmes rencontrés lors de la manipulation des intervalles. Les exemples présentés montrent que l'utilisation de fonctions d'inclusion naturelles, en dépit de leurs implémentations aisées, est souvent déconseillée voir inefficace. Ceci est principalement lié au phénomène de dépendance dû aux variables multi-occurrentes. Il est alors plus utile d'utiliser d'autres fonctions d'inclusions qui sont généralement plus précises (au moins pour des intervalles de tailles assez petites) ; en général on utilise des formes centrées. Ces formes nécessitent l'évaluation de dérivées, ceci est fait d'une manière garantie et optimale en utilisant la différentiation automatique. Le deuxième problème souvent rencontré est lié au phénomène d'enveloppement, rencontré lorsqu'on est amené à représenter un ensemble de forme géométrique quelconque par un pavé ; un pessimisme est alors introduit. Pour réduire ce pessimisme, on fait recours à l'utilisation de fonctions d'inclusion centrées et de sous-pavages.

La deuxième partie de ce chapitre a été consacrée aux méthodes d'inversion ensemblistes. En particulier, on a détaillé l'algorithme de partitionnement SIVIA, utilisé pour des modèles non linéaires ; il sera utilisé tout au long de cette thèse. Comme on l'a noté dans la section consacrée à SIVIA, la complexité de cet algorithme est exponentielle vis-à-vis du nombre des variables ; il est donc utilisable lorsque le nombre de variables ne dépasse pas 2 ou 3. Le cas échéant et pour un grand nombre de paramètres, SIVIA est alors associé aux contracteurs permettant de limiter le nombre de bissections. Dans les chapitres suivants, on utilisera le contracteur *propagation rétropropagation* étant donné son efficacité lorsque les domaines des variables sont de tailles assez importantes.

Comme on l'a montré dans l'exemple 9, on est souvent amené à étudier des modèles à variables complexes. On est donc obligé de décomposer la sortie en une partie réelle et une partie imaginaire afin de pouvoir utiliser l'arithmétique des intervalles réels. Cette décomposition provoque un accroissement du nombre d'occurrences des variables à estimer. Pour éviter ce problème, on propose d'utiliser des intervalles complexes ; le chapitre suivant sera alors consacré aux intervalles complexes et en particulier, on étendra la représentation polaire des nombres complexes aux intervalles ; cette forme étant particulièrement souhaitable pour des modèles fortement non linéaires.

Chapitre 2

Intervalles Complexes

1. Introduction

Les données utilisées pour l'estimation de paramètres sont en général de nature incertaine, ces incertitudes peuvent être représentées à l'aide des intervalles. Dans le domaine de l'estimation de paramètres, cette approche permet de propager les erreurs (de modélisation et/ou de mesure) afin de caractériser l'incertitude sur les paramètres estimés. Nous verrons dans le chapitre suivant que dans certaines applications, le modèle utilisé peut être décrit par une fonction dont les variables sont de type complexe. Dans ce cas, les incertitudes sont définies par des intervalles complexes, d'où la nécessité de définir l'arithmétique des intervalles complexes.

Dans la littérature, on trouve principalement deux représentations des intervalles complexes : intervalles *circulaires* et *rectangulaires* [AH83] [PP98]. Pour la représentation circulaire, on définit un intervalle complexe circulaire par son centre et son rayon. En revanche, un intervalle rectangulaire est défini par deux intervalles réels : un pour la partie réelle et un pour la partie imaginaire.

Nous verrons dans ce chapitre que l'addition et la soustraction de deux intervalles complexes représentés par les formes circulaires et rectangulaires sont des opérations exactes, i.e. la somme (ainsi que la différence) de deux intervalles complexes est encore un intervalle complexe, alors que la multiplication et la division ne le sont pas. Ces deux représentations sont donc très efficaces lorsque le modèle utilisé est linéaire.

Dans ce chapitre, nous allons étendre la représentation polaire des nombres complexes pour le cas des intervalles complexes. Nous verrons alors que la multiplication et la division sont exactes et sont obtenues facilement, mais l'addition et la soustraction ne le sont plus. Nous proposerons alors des algorithmes permettant de trouver le plus petit intervalle polaire contenant la somme (ou la différence) de deux intervalles complexes. Dans la suite de ce chapitre, nous utiliserons l'expression *intervalle complexe* indépendamment de la représentation utilisée.

Ce chapitre est structuré comme suit, nous allons commencer d'abord par donner quelques rappels sur les représentations des intervalles complexes les plus utilisées i.e. forme rectangulaire et forme circulaire [AH83] [PP98]. Dans la section 4, on va étendre la représentation polaire des nombres complexes aux intervalles. Les opérations arithmétiques

élémentaires sont alors définies par des algorithmes spécifiques dans la section 5. Enfin, nous allons présenter quelques exemples illustrant l'utilité des algorithmes proposés.

2. Représentation rectangulaire

Cette représentation est la plus facile à mettre en œuvre, mais comme on va le voir, elle est souvent très pessimiste.

2.1. Définition

Soit $[x_1]$ et $[x_2]$ deux intervalles de $\mathbb{IR}$. L'ensemble des nombres complexes donné par

$$[X] = \{x_1 + ix_2 \mid x_1 \in [x_1],\ x_2 \in [x_2]\} = [x_1] + i[x_2]; \qquad i^2 = -1 \tag{1}$$

est appelé intervalle complexe. L'ensemble des intervalles complexes rectangulaires est noté par $\mathbb{R}(\mathbb{C})$.

L'intervalle complexe $[X]$ défini par (1) est un rectangle dont les cotés sont parallèles aux axes réel et imaginaire. A titre d'exemple, l'intervalle complexe $[1,2] + i[0,2]$ est représenté sur la figure 1.

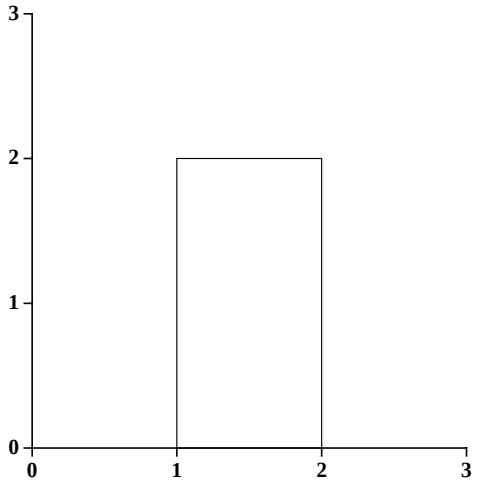

Figure 1 : Représentation de l'intervalle complexe rectangulaire [1,2] + i[0,2]

Remarque 1 : Soient $[X], [Y] \in \mathbb{R}(\mathbb{C})$ tels que : $[X] = [x_1] + i[x_2]$ et $[Y] = [y_1] + i[y_2]$, alors $[X]$ et $[Y]$ sont égaux si et seulement si $[x_1] = [y_1]$ et $[x_2] = [y_2]$.

2.2. Opérations élémentaires

Soient $[X], [Y] \in \mathbb{R}(\mathbb{C})$ tels que : $[X] = [x_1] + i[x_2]$ et $[Y] = [y_1] + i[y_2]$. Les opérations $\{+, -, \cdot, /\}$ sont définies comme suit :

$$[X] + [Y] = [x_1] + [y_1] + i([x_2] + [y_2])$$

$$[X] - [Y] = [x_1] - [y_1] + i([x_2] - [y_2])$$

$$[X] \cdot [Y] = ([x_1] \cdot [y_1] - [x_2] \cdot [y_2]) + i([x_1] \cdot [y_2] + [x_2] \cdot [y_1])$$

$$[X] / [Y] = (([x_1] \cdot [y_1] + [x_2] \cdot [y_2]) + i([x_2] \cdot [y_1] - [x_1] \cdot [y_2])) / ([y_1]^2 + [y_2]^2)$$

Remarque 2 : La division de deux intervalles complexes n'est pas possible lorsque $0 \in [y_1]^2 + [y_2]^2$. Ceci est le cas lorsque $0 \in [y_1]$ et $0 \in [y_2]$, mais on peut aussi tomber sur ce cas lorsqu'on calcule $[y_1]^2$ par $[y_1] \cdot [y_1]$ et $[y_2]^2$ par $[y_2] \cdot [y_2]$. ♦

On remarque d'après la définition de l'addition (ainsi que de la soustraction) que l'égalité

$$[X] \pm [Y] = \{x \pm y \mid x \in [X],\ y \in [Y]\}$$

est toujours vraie, i.e. l'ensemble des nombres complexes obtenu par le terme à droite est égal à la somme des intervalles complexes $[X]$ et $[Y]$. Ceci n'est pas le cas pour la multiplication.

Exemple 1 : Soit $[X] = [1,2] + i[0,0]$ et $[Y] = [1,1] + i[1,1]$ deux intervalles rectangulaires, alors en utilisant la définition de la multiplication on obtient :

$$[X] \cdot [Y] = [1, 2] + i[1, 2]$$

D'autre part on a

$$\{x \cdot y \mid x \in [X],\ y \in [Y]\} = \{t + ti \mid t \in [1,2]\}$$

On remarque que la multiplication définie ci-dessus n'est pas exacte : elle est pessimiste, dans le sens où le résultat obtenu contient mais n'est pas égal à l'ensemble des résultats des multiplications des nombres complexes contenus respectivement dans les intervalles $[X]$ et $[Y]$. ♦

Comme la multiplication, l'opération de division définie ci-dessus n'est pas exacte. En général, l'utilisation de cette définition génère un résultat très pessimiste. Une autre définition moins pessimiste consiste à calculer la division en utilisant la formule suivante [RL71] [PP98]:

$$[X]/[Y] = [X] \cdot \frac{1}{[Y]}$$

où

$$\frac{1}{[Y]} := \inf\left\{[A] \in \mathbb{R}(\mathbb{C}) \mid \{1/a \mid a \in [Y]\} \subseteq [A]\right\}$$

Cette méthode consiste donc à trouver le plus petit intervalle complexe contenant $1/[Y]$ puis à le multiplier par $[X]$. Elle requiert un temps de calcul important. Une autre méthode [LG85], plus efficace, consiste à calculer le minimum et le maximum de la partie réelle et de la partie imaginaire de :

$$\frac{x_1 + ix_2}{y_1 + iy_2} = \frac{x_1y_1 + x_2y_2}{y_1^2 + y_2^2} + \frac{x_2y_1 - x_1y_2}{y_1^2 + y_2^2}$$

avec

$$x_1 + ix_2 \in [X], \quad y_1 + iy_2 \in [Y] \text{ et } y_1^2 + y_2^2 > 0.$$

Remarque 3 : On aura besoin dans le chapitre suivant de calculer l'intersection de deux intervalles complexes. La représentation rectangulaire des complexes représente alors un avantage puisque l'intersection de deux (ou plusieurs) intervalles complexes est un intervalle complexe. ♦

2.3. Fonctions d'inclusions

Soit $\mathbf{f} : \mathbb{C}^n \to \mathbb{C}^m$ une fonction vectorielle d'argument complexe contenant un nombre fini d'opérations arithmétiques et de fonctions élémentaires. Une fonction d'inclusion de **f**, notée [**f**] est une fonction de $\mathbb{R}(\mathbb{C}^n)$ dans $\mathbb{R}(\mathbb{C}^m)$ vérifiant

$$\mathbf{f}([\mathbf{X}]) = \{\mathbf{f}(\mathbf{x}) \mid \mathbf{x} \in [\mathbf{X}]\} \subseteq [\mathbf{f}]([\mathbf{X}]) \tag{2}$$

En général, la fonction d'inclusion n'est pas unique et elle dépend de l'écriture de **f**. Etant donné que la multiplication et la division ne sont pas exactes, les fonctions d'inclusions sont en général pessimistes lorsqu'on utilise la représentation rectangulaire. Le pessimisme est alors dû à l'effet de dépendance, à l'effet d'enveloppement et à la représentation des intervalles complexes. Souvent, on utilise la fonction d'inclusion naturelle qui consiste à remplacer chaque variable ponctuelle par son intervalle complexe correspondant et chaque opération élémentaire par son extension aux intervalles complexes. On a vu dans le chapitre précédent que la fonction d'inclusion naturelle est minimale lorsque toutes les variables sont mono-occurrentes. Ceci n'est pas le cas pour les intervalles complexes, il suffit alors d'avoir une division ou une multiplication pour que la fonction soit non minimale.

Exemple 2 : On considère la fonction :

$$f : \mathbb{C} \to \mathbb{C}$$
$$x \mapsto \exp(x)$$

Soit $[X] = [0,1] + i[0,\pi/2]$, alors :

$$f([X]) = \exp([0,1] + i[0,\pi/2]) = \exp([0,1]) \cdot \exp(i[0,\pi/2])$$
$$= [\exp(0), \exp(1)] \cdot (\cos([0,\pi/2]) + i\sin([0,\pi/2]))$$
$$= [0,1] \cdot ([1,e] + i[0,1]) = [0,e] + i[0,e]$$

3. Forme circulaire

Dans la littérature, on trouve plus d'études concernant la représentation circulaire que pour la forme rectangulaire [PP99], ceci s'explique par le fait qu'elle est en général moins pessimiste.

3.1. Définition

Soit $c \in \mathbb{C}$ et $r \in \mathbb{R}$, $r \geq 0$. L'ensemble Z tel que :

$$Z = \{z \in \mathbb{C} \mid |z - c| \leq r\} \tag{3}$$

est un *intervalle complexe circulaire* (appelé aussi *disque*) [Nic69] [GH72] [AH83] [PP98]. Un disque noté $\{c; r\}$ est alors caractérisé par son centre c et par son rayon r. Sur la figure 2, on a tracé le disque $\{1+i; 1\}$.

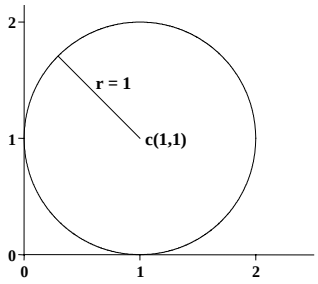

Figure 2 : Représentation du disque $\{1+i; 1\}$

L'ensemble des disques de $\mathbb{C}$ est noté par $\mathbb{K}(\mathbb{C})$.

Remarque 4 : Deux disques $Z_1 = \{c_1; r_1\}$ et $Z_2 = \{c_2; r_2\}$ sont dits égaux si et seulement si $c_1 = c_2$ et $r_1 = r_2$. ♦

3.2. Opérations arithmétiques

Les opérations arithmétiques élémentaires sont étendues sur $\mathbb{K}(\mathbb{C})$ en utilisant quelques propriétés des nombres complexes [PP98].

Soient $Z_1, Z_2 \in \mathbb{K}(\mathbb{C})$ tels que $Z_1 = \{c_1 ; r_1\}$ et $Z_2 = \{c_2 ; r_2\}$, alors :

$$Z_1 \pm Z_2 = \{c_1 \pm c_2 ; r_1 + r_2\}$$

L'addition et la soustraction de deux disques est alors une opération exacte, *i.e. :*

$$\{c_1 \pm c_2 ; r_1 + r_2\} = \{x \pm y \mid x \in Z_1, y \in Z_2\}$$

De même pour l'opération $1/Z_2$ définie par :

$$\frac{1}{Z_2} = \left\{ \frac{\overline{c}_2}{|c_2|^2 - r_2^2} ; \frac{r_2}{|c_2|^2 - r_2^2} \right\}$$

où $\overline{c}_2$ et $|c_2|$ représentent le conjugué et le module de c_2. Cette opération n'est définie que lorsque $0 \notin Z_2$, *i.e.* $|c_2|^2 - r_2^2 \neq 0$.
La multiplication est définie comme suit :

$$Z_1 \cdot Z_2 = \{c_1 \cdot c_2 ; |c_1| r_2 + |c_2| r_1 + r_1 r_2\}$$

On démontre (voir par exemple [PP98]) que l'inclusion suivante est vérifiée

$$Z_1 \cdot Z_2 \supseteq \{x \cdot y \mid x \in Z_1, y \in Z_2\}$$

mais l'égalité est rarement vraie. Ceci implique que l'évaluation de la multiplication est en général pessimiste.

La division de Z_1 par Z_2 est donnée par

$$\frac{Z_1}{Z_2} = Z_1 . \frac{1}{Z_2}, \quad 0 \notin Z_2$$

La division nécessite alors d'effectuer une multiplication, d'où une opération non exacte.

Remarque 5 : Nous avons vu que la multiplication est une opération non exacte, *i.e.* le produit de deux disques n'est pas un disque. En plus, en utilisant la définition présentée ci-dessus, on n'obtient généralement pas le plus petit disque contenant le produit. Une autre définition de la multiplication a été alors proposée dans [Kri74]. Cette dernière forme est minimale i.e. on obtient le plus petit disque contenant l'ensemble $\{x \cdot y \mid x \in Z_1, y \in Z_2\}$. Le diamètre de ce disque est égal au diamètre de l'ensemble $\{x \cdot y \mid x \in Z_1, y \in Z_2\}$. Cette

définition permet de limiter le pessimisme lors de l'évaluation de la multiplication et ainsi que de la division. Mais cette méthode présente deux inconvénients : sa complexité est importante et la propriété de distributivité au sens de l'inclusion n'est pas vraie ($Z_1 \subseteq T_1$, $Z_2 \subseteq T_2$ n'implique pas que $Z_1 \cdot Z_2 \subseteq T_1 \cdot T_2$) [PP98].

Remarque 6 : Une fonction d'inclusion circulaire est une extension d'une fonction d'argument complexe aux intervalles circulaires. Une fonction d'inclusion est obtenue en utilisant les extensions des fonctions élémentaires et en utilisant les opérations définies ci-dessus. L'extension des fonctions élémentaires se fait en cherchant le plus petit disque contenant l'évaluation exacte de la fonction.

Exemple 3 : Soit $Z = \{c; r\} \in \mathbb{K}(\mathbb{C})$, on démontre dans [PP98] que le plus petit disque contenant le Logarithme de Z est donné par :

$$\mathrm{Log}(Z) = \left\{ \mathrm{Log}\left(\sqrt{|c^2| - r^2}\right) + i \arg c \; ; \frac{1}{2} Log\left(\frac{|c| + r}{|c| - r}\right) \right\}$$

On pourra consulter [Kri73] [PT93] [PP98] pour plus de détails sur les fonctions d'inclusions des fonctions élémentaires.

4. Forme Polaire

Nous avons vu dans les paragraphes précédents que les opérations de multiplication et de division de deux intervalles complexes (rectangulaires et circulaires) ne sont pas exactes. De plus, il existe plusieurs définitions de la multiplication dans le cas des intervalles circulaires. Lors de l'évaluation de ces opérations, un pessimisme est dès lors introduit. Ces représentations des intervalles complexes ne sont pas toujours efficaces lorsqu'on a à étudier une fonction fortement non linéaire. Dans ce paragraphe, nous nous proposons d'étendre la représentation polaire des nombres complexes au cas des intervalles.

4.1. Définition

Soient $[\rho] \subseteq \mathbb{R}^+$ et $[\theta] \subseteq \mathbb{R}$. L'ensemble *Z* tel que :

$$Z = \{ z \in \mathbb{C} \mid |z| \in [\rho], \arg(z) \in [\theta] \}$$

est un intervalle complexe polaire (ou secteur) noté $\{[\rho];[\theta]\}$. L'ensemble des intervalles complexes polaires est noté par $\mathbb{S}(\mathbb{C})$.

Exemple 4 : Sur la figure 3 on représente le secteur $\{[1,2];[\pi/4,\pi/3]\}$.

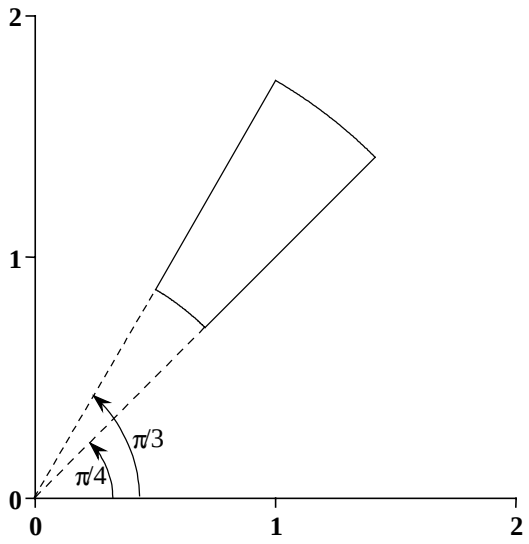

Figure 3 : Exemple de secteur dans le plan complexe

Un secteur est alors caractérisé par deux paramètres : son module $[\rho] = \left[\rho^-, \rho^+\right]$ ($\rho^- \geq 0$) et son argument $[\theta]$ (de taille inférieure ou égale à 2π).

Remarque 7 : On considère le secteur représenté sur la figure 4, les extrémités de son argument sont $3\pi/4$ et $-3\pi/4$ (on prend bien sûr le sens trigonométrique direct). La représentation de l'argument de ce secteur pose donc un problème ; en effet, le minimum de $[\theta]$ est supérieur à son maximum (à cause du caractère cyclique de l'argument). Pour lever cette ambiguïté, on choisit :

$$0 \leq \theta^- < 2\pi \ , \ 0 \leq \theta^+ < 4\pi \ \text{ et } \ 0 \leq \theta^+ - \theta^- \leq 2\pi$$

où : $[\theta] = \left[\theta^-, \theta^+\right]$.

On prendra donc ici $[\theta] = \left[\frac{3\pi}{4}, \frac{5\pi}{4}\right]$ en lieu et place de $[\theta] = \left[\frac{3\pi}{4}, -\frac{\pi}{4}\right]$.

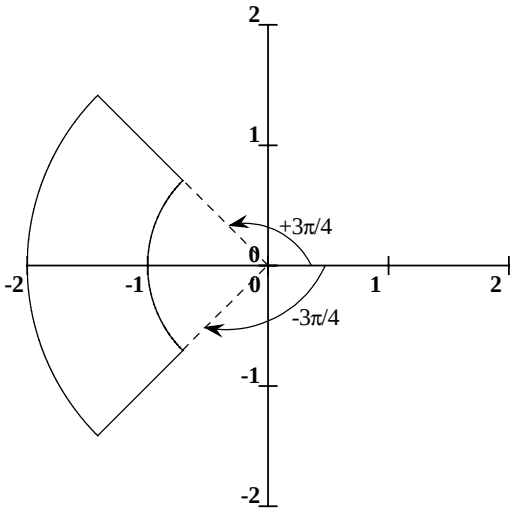

Figure 4 : Choix de la représentation de l'argument d'un secteur

4.2. Opérations arithmétiques

Soit $Z_1 = \{[\rho_1];[\theta_1]\}$ et $Z_2 = \{[\rho_2];[\theta_2]\}$ deux secteurs ; la multiplication de ces deux secteurs est définie par :

$$\begin{aligned} Z_1 \cdot Z_2 &= \{\{\rho_1;\theta_1\}\cdot\{\rho_2;\theta_2\} \mid \{\rho_1;\theta_1\} \in Z_1, \{\rho_2;\theta_2\} \in Z_2\} \\ &= \{\rho_1 \cdot \rho_2 \exp(j(\theta_1+\theta_2)) \mid \rho_1 \in [\rho_1], \rho_2 \in [\rho_2], \theta_1 \in [\theta_1], \theta_2 \in [\theta_2]\} \\ &= \{[\rho_1]\cdot[\rho_2] ; [\theta_1]+[\theta_2]\} = \{[\rho] ; [\theta]\} \end{aligned}$$

Etant donné que le produit et la somme de deux intervalles réels sont des opérations exactes, le produit de deux secteurs est alors un secteur. De même pour la division définie par :

$$\begin{aligned} Z_1 / Z_2 &= \{\{\rho_1;\theta_1\}/\{\rho_2;\theta_2\} \mid \{\rho_1;\theta_1\} \in Z_1, \{\rho_2;\theta_2\} \in Z_2\} \\ &= \{[\rho_1]/[\rho_2] ; [\theta_1]-[\theta_2]\} = \{[\rho] ; [\theta]\} \end{aligned}$$

avec $0 \notin [\rho_2]$. On doit noter que lorsqu'on effectue une opération quelconque entre deux secteurs, les bornes de l'argument du secteur résultat peuvent dépasser les limites imposées : $0 \le \theta^- < 2\pi$ et $0 \le \theta^+ < 4\pi$. Dans ce cas, on peut ajuster ces bornes en leur ajoutant $2k\pi,\ (k \in \mathbb{Z})$ de telle sorte à avoir $0 \le \theta^- < 2\pi$, $0 \le \theta^+ < 4\pi$ et $0 \le \theta^+ - \theta^- \le 2\pi$.

L'addition de Z_1 et Z_2 est définie par :

$$Z_1 + Z_2 = \{\{\rho_1;\theta_1\} + \{\rho_2;\theta_2\} \mid \{\rho_1;\theta_1\} \in Z_1, \{\rho_2;\theta_2\} \in Z_2\}$$

L'ensemble exact, résultat de $Z_1 + Z_2$, n'est donc pas un secteur du plan complexe. Une première idée pour trouver un secteur contenant l'ensemble défini ci-dessus consiste à passer par la forme rectangulaire. Dans ce cas, on cherche d'abord les deux plus petits rectangles $\left[Z_1^{'}\right]$ et $\left[Z_2^{'}\right]$ contenant respectivement les secteurs Z_1 et Z_2 ; ensuite, on additionne $\left[Z_1^{'}\right]$ et $\left[Z_2^{'}\right]$ pour obtenir un rectangle $[Z]$. Le secteur résultat est alors le plus petit secteur contenant $[Z]$. Cette méthode est très facile à mettre en œuvre, en contre-partie le résultat obtenu est généralement très pessimiste.

Dans la suite de ce chapitre, nous proposons une méthode permettant de trouver le plus petit secteur $\hat{Z} = \left\{[\hat{\rho}];\left[\hat{\theta}\right]\right\}$ contenant la somme de deux secteurs sans passer par la forme rectangulaire. Le secteur $\hat{Z}$ vérifie alors :

$$\forall\ Z \supseteq \left\{\{\rho_1;\theta_1\}\ +\ \{\rho_2;\theta_2\}\ \mid\ \{\rho_1;\theta_1\} \in Z_1, \{\rho_2;\theta_2\} \in Z_2\right\},\ \ \hat{Z} \subseteq Z$$

On dira alors que $\hat{Z}$ est *minimal*.

Remarque 8 : La soustraction est obtenue directement en utilisant l'addition

$$Z_1 - Z_2 = \left\{\ [\rho_1]\ ;\ [\theta_1]\right\} - \left\{\ [\rho_2]\ ;\ [\theta_2]\right\} = \left\{\ [\rho_1]\ ;\ [\theta_1]\right\} + \left\{\ [\rho_2]\ ;\ [\theta_2] + \pi\ \right\}$$

5. Addition de deux secteurs

Soient $Z_1 = \left\{[\rho_1];[\theta_1]\right\}$ et $Z_2 = \left\{[\rho_2];[\theta_2]\right\}$ deux secteurs du plan complexe. Dans cette partie nous proposons de calculer le plus petit secteur $\hat{Z}$ contenant la somme de Z_1 et Z_2, ce qui revient donc à calculer son module et son argument.

Soient $[\rho_1]=\left[\rho_1^-,\rho_1^+\right]$, $[\theta_1]=\left[\theta_1^-,\theta_1^+\right]$,$[\rho_2]=\left[\rho_2^-,\rho_2^+\right]$ et $[\theta_2]=\left[\theta_2^-,\theta_2^+\right]$. Alors :

$$\forall\ z_1(\rho_1,\theta_1) \in Z_1 \text{ et } z_2(\rho_2,\theta_2) \in Z_2, \exists\ z(\rho,\varphi) \in \hat{Z} \mid z = z_1 + z_2$$

où :

$$\rho^2 = \rho_1^2 + \rho_2^2 + 2\rho_1\rho_2\cos(\theta_1 - \theta_2) \text{ et } \varphi = \operatorname{atan}\left(\frac{\rho_1\sin\theta_1 + \rho_2\sin\theta_2}{\rho_1\cos\theta_1 + \rho_2\cos\theta_2}\right)$$

avec $\rho_1\cos\theta_1 + \rho_2\cos\theta_2 \neq 0$. Nous allons étudier plus loin le cas où $\rho_1\cos\theta_1 + \rho_2\cos\theta_2 = 0$.

Pour déterminer $\hat{Z}$, il suffit alors de déterminer le plus petit intervalle réel $[\rho]$ contenant l'ensemble :

$$\begin{aligned}\{\rho \in \mathbb{R}^+ \mid \rho^2 = \rho_1^2 + \rho_2^2 + 2\rho_1\rho_2\cos(\theta_1 - \theta_2),\\ \text{où } \rho_1 \in [\rho_1], \rho_2 \in [\rho_2] \text{ et } \theta_1 - \theta_2 \in [\theta_1] - [\theta_2]\}\end{aligned} \tag{4}$$

et le plus petit intervalle $[\varphi]$ contenant l'ensemble :

$$\begin{aligned}\left\{\varphi \in \mathbb{R} \mid \varphi = \operatorname{atan}\left(\frac{\rho_1\sin\theta_1 + \rho_2\sin\theta_2}{\rho_1\cos\theta_1 + \rho_2\cos\theta_2}\right),\right.\\ \left.\text{où } \theta_1 \in [\theta_1],\ \theta_2 \in [\theta_2], \rho_1 \in [\rho_1] \text{ et } \rho_2 \in [\rho_2]\right\}\end{aligned} \tag{5}$$

5.1. Problèmes d'optimisation

Calculer les bornes inférieures et supérieures de $[\rho]$ et $[\varphi]$ est un problème d'optimisation qui peut être résolu d'une manière garantie en utilisant des méthodes basées sur l'analyse par intervalles [Neu90] [Han92] [Kea96]. Ces méthodes permettent de trouver un ensemble contenant la solution ponctuelle du problème d'optimisation. Le minimum (respectivement le maximum) de la fonction étudiée correspond alors à celui de l'ensemble calculé par l'algorithme d'optimisation garantie. Néanmoins, on n'obtient qu'un encadrement extérieur ; un pessimisme est alors introduit. Dans la suite de ce chapitre, les problèmes d'optimisation définis par (4) et (5) seront résolus analytiquement [CRRI04]. Ceci est possible puisque le nombres des variables est assez réduit et que les fonctions à optimiser sont assez simples.

Les bornes inférieure et supérieure de $[\rho]$ sont alors données par les racines carrées des bornes de la fonction définie par :

$$f(\rho_1, \rho_2, \theta) = \rho_1^2 + \rho_2^2 + 2\rho_1\rho_2\cos(\theta) \tag{6}$$

avec $\theta = \theta_1 - \theta_2$.

Les bornes inférieure et supérieure de $[\varphi]$ correspondent aux bornes de la fonction *g* définie par :

$$g(x, \theta_1, \theta_2) = \operatorname{atan}\left(\frac{x\sin\theta_1 + \sin\theta_2}{x\cos\theta_1 + \cos\theta_2}\right) \tag{7}$$

avec $x = \rho_1/\rho_2$ $(\rho_2^- > 0)$.

Pour trouver $[\rho]$ et $[\varphi]$, il suffit alors de résoudre les quatre problèmes d'optimisation définis par :

$$\min_{\Omega_1} f(\rho_1, \rho_2, \theta) \tag{8}$$

et

$$\max_{\Omega_1} f(\rho_1, \rho_2, \theta) \tag{9}$$

avec $$\Omega_1 = [\rho_1] \times [\rho_2] \times ([\theta_1] - [\theta_2]) \subseteq \mathbb{R}^+ \times \mathbb{R}^+ \times [0, 2\pi] \tag{10}$$

et

$$\min_{\Omega_2} g(x, \theta_1, \theta_2) \tag{11}$$

$$\max_{\Omega_2} g(x, \theta_1, \theta_2) \tag{12}$$

avec $$\Omega_2 = ([\rho_1] / [\rho_2]) \times [\theta_1] \times [\theta_2] \subseteq \mathbb{R}^+ \times [0, 2\pi] \times [0, 2\pi] \tag{13}$$

5.2. Conditions d'optimalité

Les problèmes d'optimisation donnés par les équations (8) à (13) sont résolus en exploitant les propriétés de monotonie des fonctions *f* et *g* définies par (6) et (7) par rapport à chacune des variables.

Soient $(u_1, u_2, u_3) \in [\rho_1] \times [\rho_2] \times [\theta] = \Omega_1$, la fonction *f* atteint son maximum pour le vecteur $\mathbf{u}^* = (u_1^*, u_2^*, u_3^*) \in \Omega_1$, si pour chaque indice *i*, l'élément u_i^* vérifie l'une des conditions suivantes :

$$\frac{\partial f}{\partial u_i}(\mathbf{u}^*) = 0 \quad \text{et} \quad \frac{\partial^2 f}{\partial u_i^2}(\mathbf{u}^*) \leq 0 \tag{14}$$

$$u_i^* = u_i^- \quad \text{et} \quad \frac{\partial f}{\partial u_i}(\mathbf{u}^*) < 0 \tag{15}$$

$$u_i^* = u_i^+ \quad \text{et} \quad \frac{\partial f}{\partial u_i}(\mathbf{u}^*) > 0 \tag{16}$$

Dans le cas de la minimisation de la fonction *f*, les trois conditions (14), (15), (16) sont applicables mais le sens des inégalités change.

Nous remarquons que chacune de ces conditions est composée d'une première partie donnée par une équation (condition de premier ordre) et d'une seconde partie définie par une inégalité (condition de second ordre). Un ensemble de 3 conditions de premier ordre, une pour chaque indice *i*, est un système d'équations qui a généralement au plus une solution.

Dans la suite, nous appellerons *candidat* tout point de Ω_1 qui vérifie, pour chacun de ces éléments, une des conditions du premier ordre. Si en plus, les conditions du second ordre sont satisfaites, le candidat est dit *acceptable* il correspond donc à un minimum local.

La stratégie suivie dans la suite de ce chapitre pour ce problème d'optimisation consiste à déterminer analytiquement tous les candidats en examinant toutes les combinaisons des conditions du premier ordre, et ensuite, éliminer les candidats qui ne peuvent jamais être acceptables en examinant les conditions du second ordre. Le minimum parmi les candidats acceptables est le minimum global.

Dans la suite de ce chapitre, nous allons donner des algorithmes permettant de résoudre les problèmes d'optimisation (8) à (12) en utilisant le raisonnement donné ci-dessus [CRRI04].

5.3. Maximum du module : ρ^+

Pour trouver le maximum du module de $\hat{Z}$, il suffit de résoudre le problème d'optimisation (9) en faisant un raisonnement sur chacun des paramètres de la fonction *f*. On note alors par ρ_1^*, ρ_2^* et θ^* les valeurs de ρ_1, ρ_2 et θ permettant de calculer ρ^+ en utilisant la formule (6).

5.3.1. Maximisation par rapport à ρ_1

Dans ce paragraphe on fixe $\rho_2 = \rho_2^\circ$ et $\theta = \theta^\circ$ afin d'étudier l'évolution de *f* en fonction de ρ_1. On a alors :

$$\frac{\partial f}{\partial \rho_1} = 2\left(\rho_1 + \rho_2^\circ \cos\theta^\circ\right) \quad \text{et} \quad \frac{\partial^2 f}{\partial \rho_1^2} = 2$$

La dérivée seconde de *f* par rapport à ρ_1 est toujours positive, ce qui signifie que la fonction *f* n'atteint pas son maximum pour une valeur ρ_1 à l'intérieur de l'intervalle $[\rho_1]=[\rho_1^-, \rho_1^+]$; le maximum est alors nécessairement atteint à une borne, soit $\rho_1^* = \rho_1^-$ ou $\rho_1^* = \rho_1^+$ selon le signe de $\frac{\partial f}{\partial \rho_1}$. Il y a donc deux valeurs possibles pour ρ_1^*.

5.3.2. Maximisation par rapport à ρ_2

Maintenant on fixe $\rho_1 = \rho_1^\circ$ et $\theta = \theta^\circ$ afin d'étudier l'évolution de *f* en fonction de ρ_2. La dérivée première et la dérivée seconde de *f* par rapport à ρ_2 sont données par :

$$\frac{\partial f}{\partial \rho_2} = 2\left(\rho_2 + \rho_1^\circ \cos\theta^\circ\right) \quad \text{et} \quad \frac{\partial^2 f}{\partial \rho_2^2} = 2$$

Comme pour la maximisation par rapport à ρ_1, la fonction *f* atteint alors son maximum pour $\rho_2^* = \rho_2^-$ ou pour $\rho_2^* = \rho_2^+$ selon le signe de $\frac{\partial f}{\partial \rho_2}$.

5.3.3. Maximisation par rapport à θ

On fixe les deux variables $\rho_1 = \rho_1^\circ$ et $\rho_2 = \rho_2^\circ$. Les dérivées première et seconde de *f* par rapport à θ sont alors données par :

$$\frac{\partial f}{\partial \theta} = -2\rho_1^\circ \rho_2^\circ \sin\theta \text{ et } \frac{\partial^2 f}{\partial \theta^2} = -2\rho_1^\circ \rho_2^\circ \cos\theta$$

Il y a alors quatre cas à étudier :

- Si $\sin\theta = 0$, alors :

  • $\theta = 0$ (ou $\theta = 2\pi$) $\Rightarrow$ $\cos\theta = 1$, la valeur $\theta^* = 0$ correspond bien à un maximum (la dérivée seconde est négative).

  • $\theta = \pi$ $\Rightarrow$ la dérivée seconde est positive, cette valeur ne correspond donc pas à un maximum.

- Si $\sin\theta^+ < 0$, le maximum correspond à $\theta^* = \theta^+$

- Si $\sin\theta^- > 0$, le maximum correspond à $\theta^* = \theta^-$

- Si les deux conditions $\sin\theta^+ < 0$ et $\sin\theta^- > 0$ sont valides, alors le maximum peut correspondre à θ^+ ou θ^- ; les deux cas doivent alors être examinés.

5.3.4. Synthèse du calcul du maximum du module

Pour calculer le maximum de la fonction *f*, il suffit alors de déterminer le triplet (ρ_1^*, ρ_2^*, θ^*). Ces dernières valeurs sont choisies en examinant les conditions suivantes :

- $\rho_1^* = \rho_1^+$, si $\rho_1^+ + \rho_2^* \cos\theta^* > 0$ (A1)

- $\rho_1^* = \rho_1^-$, si $\rho_1^- + \rho_2^* \cos\theta^* < 0$ (A2)

- $\rho_2^* = \rho_2^+$, si $\rho_2^+ + \rho_1^* \cos\theta^* > 0$ (B1)

- $\rho_2^* = \rho_2^-$, si $\rho_2^- + \rho_1^* \cos\theta^* < 0$ (B2)

- $\theta^* = 0$, si $0 \in [\theta]$ (C1)

- $\theta^* = \theta^+$, si $\sin\theta^+ < 0$ (C2)

- $\theta^* = \theta^-$, si $\sin\theta^- > 0$ (C3)

On rappelle que les conditions (C2) et (C3) peuvent exister simultanément, il est donc indispensable de déterminer laquelle des valeurs $\theta^* = \theta^-$ ou $\theta^* = \theta^+$ correspond au maximum.

On note alors par « *ijk* » (i = 1, 2 ; j = 1, 2; k = 1, 2, 3) une combinaison des paramètres ρ_1, ρ_2 et θ permettant de calculer le maximum de ρ.

Exemple 5 :

A titre d'exemple, le candidat 123 correspond aux conditions (A1), (B2) et (C3) vraies, donc le module est calculé en utilisant (6) avec $\rho_1^* = \rho_1^+$, $\rho_2^* = \rho_2^-$ et $\theta^* = \theta^-$. Ceci suppose que les trois conditions (A1), (B2) et (C3) sont vérifiées :

- $\rho_1^+ + \rho_2^- \cos\theta^- > 0$ (A1)

- $\rho_2^- + \rho_1^+ \cos\theta^- < 0$ (B2)

- $\sin\theta^- > 0$ (C3)

Le module calculé en utilisant 123 est alors donné par :

$$\rho(123) = \sqrt{\left(\rho_1^+\right)^2 + \left(\rho_2^-\right)^2 + 2\rho_1^+\rho_2^-\cos\theta^-} \qquad \blacklozenge$$

Candidats: Pour trouver le triplet (ρ_1^*, ρ_2^*, θ^*) permettant de calculer le maximum de la fonction *f*, il suffit de trouver toutes les combinaisons (*ijk*) satisfaisant les conditions (A_i et B_j et C_k).

- supposons que la condition C1 est satisfaite (i.e. $0 \in [\theta]$), alors le maximum global de ρ est donné par $\rho(111)$ (puisque $\rho^+ \le \rho_1^+ + \rho_2^+$).
- Supposons maintenant que $0 \notin [\theta]$ et que C2 est satisfaite (mais pas C3), alors les candidats sont :

  $$\rho(112),\ \rho(122),\ \rho(212),\ \rho(222)$$

  Le maximum de ρ^+ correspond alors au maximum de ($\rho(112)$, $\rho(122)$, $\rho(212)$, $\rho(222)$).
- Maintenant supposons que $0 \notin [\theta]$ et que C3 est vérifiée (mais pas C2), les candidats sont alors :

  $$\rho(113),\ \rho(123),\ \rho(213),\ \rho(223)$$

Remarque 5: Supposons que les conditions A2 et B2 sont satisfaites, alors on obtient la contradiction suivante: $\rho_1^- / \rho_2^- < -\cos\theta^*$ et $\rho_2^- / \rho_1^- < -\cos\theta^*$ ce qui conduit à $\cos^2\theta^* > 1$, ainsi les candidats $\rho(222)$ et $\rho(223)$ ne sont pas acceptables, ils sont donc rejetés.

Le maximum de $[\rho]$ est alors calculé en utilisant l'algorithme 1.

Notons d'abord que l'une des conditions (C1) ou (C2) ou (C3) doit être vérifiée. Lorsque la condition (C1) est vérifiée *i.e.* $0 \in [\theta]$, le maximum correspond forcément à (111). Par contre, lorsque (C1) n'est pas valide, alors on peut avoir en même temps (C2) et (C3), dans ce cas, ces deux cas doivent être traités. L'étape 3 (respectivement l'étape 4) de l'algorithme est exécutée seulement lorsque (C2) (resp. C3) est vraie et C3 (resp. C2) est fausse, ceci permettant de calculer seulement trois modules (et non pas 6 comme pour l'étape 2).

Algorithme 1 : MaxMod(Z_1, Z_2)

1. Si $0 \in [\theta]$, retourner $\rho(111)$;

2. Si $\sin\theta^+ < 0$ et $\sin\theta^- > 0$

   $\rho_{m1} = \max(\rho(112), \rho(122), \rho(212))$;

   $\rho_{m2} = \max(\rho(113), \rho(123), \rho(213))$;

   retourner $\max(\rho_{m1}, \rho_{m2})$; fin ;

3. Si $\sin\theta^+ < 0$

   retourner $\max(\rho(112), \rho(122), \rho(212))$; fin ;

4. Si $\sin\theta^- > 0$

   retourner $\max(\rho(113), \rho(123), \rho(213))$; fin ;

Dans l'étape 2 de l'algorithme 1, au pire des cas, on aura à calculer six amplitudes, i.e. $\rho(112)$, $\rho(122)$, $\rho(212)$, $\rho(113)$, $\rho(123)$ et $\rho(213)$. Par ailleurs, rappelons qu'une combinaison $A_iB_jC_k$ est un candidat si et seulement si les trois conditions A_i, B_j et C_k sont vérifiées en même temps. En pratique, il est rare que les six combinaisons citées ci-dessus représentent tous des candidats. Par conséquent, il est plus judicieux de ne calculer que les combinaisons pouvant représenter le maximum. En pratique, nous testons pour chaque combinaison les conditions du maximum avant de calculer le module correspondant. Par souci de clarté, nous n'avons pas indiqué dans l'algorithme que les conditions du maximum étaient systématiquement vérifiées.

5.4. Minimum du Module ρ^-

La même démarche que celle utilisée pour le calcul du maximum du module est appliquée au calcul du minimum du module. On considère donc de nouveau la fonction *f* définie par (6). Les différentes combinaisons permettant d'avoir des candidats sont étudiées dans la suite de cette section.

5.4.1 Minimisation par rapport à ρ_1 et ρ_2

On rappelle que pour $\rho_2 = \rho_2^\circ$ et $\theta = \theta^\circ$, les dérivées première et seconde de la fonction *f* par rapport à ρ_1 sont :

$$\frac{\partial f}{\partial \rho_1} = 2\left(\rho_1 + \rho_2^\circ \cos\theta^\circ\right) \text{ et } \frac{\partial^2 f}{\partial \rho_1^2} = 2$$

La dérivée seconde de f est positive, la valeur de ρ_1^* permettant d'annuler $\frac{\partial f}{\partial \rho_1}$, i.e. $\rho_1^* = -\rho_2^\circ \cos\theta^\circ$, correspond au minimum global de ρ (en prenant bien entendu les valeurs adéquates de ρ_2 et de θ). Par contre, si la dérivée ne s'annule pas, alors le minimum correspond forcément à une des extrémités de $[\rho_1]$. La même démarche peut être suivie pour ρ_2.

5.4.2 Minimisation par rapport à θ

On fixe les deux variables $\rho_1 = \rho_1^\circ$ et $\rho_2 = \rho_2^\circ$. On rappelle que les dérivées première et seconde de *f* par rapport à θ sont données par

$$\frac{\partial f}{\partial \theta} = -2\rho_1^\circ \rho_2^\circ \sin\theta \text{ et } \frac{\partial^2 f}{\partial \theta^2} = -2\rho_1^\circ \rho_2^\circ \cos\theta$$

Il y a alors trois cas à étudier :

- Si $\frac{\partial f}{\partial \theta}$ s'annule, alors $\sin\theta = 0$:

  • $\theta^* = 0$ (ou $\theta^* = 2\pi$) $\Rightarrow$ $\cos\theta^* = 1$, cette valeur ne correspond pas à un minimum (la dérivée seconde est négative).

  • $\theta^* = \pi$ $\Rightarrow$ la dérivée seconde est positive, $\theta^* = \pi$ correspond alors à un minimum.

- Si $\sin\theta^+ > 0$, le minimum correspond à $\theta^* = \theta^+$

- Si $\sin\theta^- < 0$, le minimum correspond à $\theta^* = \theta^-$

5.4.3 Calcul du minimum de ρ

En regroupant les conditions présentées dans les sections 5.4.1 et 5.4.2 on obtient l'ensemble des conditions suivantes :

- $\rho_1^* = -\rho_2^* \cos\theta^*$, si $\rho_1^* + \rho_2^* \cos\theta^* = 0$ (A1)
- $\rho_1^* = \rho_1^+$, si $\rho_1^+ + \rho_2^* \cos\theta^* < 0$ (A2)
- $\rho_1^* = \rho_1^-$, si $\rho_1^- + \rho_2^* \cos\theta^* > 0$ (A3)
- $\rho_2^* = -\rho_1^\circ \cos\theta^\circ$, si $\rho_2^* + \rho_1^* \cos\theta^* = 0$ (B1)
- $\rho_2^* = \rho_2^+$, si $\rho_2^+ + \rho_1^* \cos\theta^* < 0$ (B2)
- $\rho_2^* = \rho_2^-$, si $\rho_2^- + \rho_1^* \cos\theta^* > 0$ (B3)
- $\theta^* = \pi$, si $\pi \in [\theta]$ (C1)
- $\theta^* = \theta^+$, si $\sin\theta^+ > 0$ (C2)
- $\theta^* = \theta^-$, si $\sin\theta^- < 0$ (C3)

Si on considère toutes les combinaisons possibles entre les $A_iB_jC_k$ on obtiendrait alors 27 cas. Néanmoins, plusieurs combinaisons sont impossibles à satisfaire.

- Supposons que A1 et B1 sont vraies, alors $\theta^* = \pi$ et $\rho_1^* = \rho_2^*$; par la suite seule C1 est vraie et les combinaisons 112 et 113 sont alors impossibles.
- De même, si on a A1 et C1, alors forcément B1 est vraie, on élimine alors les combinaisons 121, 131, (de même si on a B1 et C1 valides, alors seulement A1 peut donner lieu à une combinaison acceptable, on élimine aussi 211 et 311).
- Si on a A1 et B2, alors $\rho_2^+ \left(1 - \left(\cos\theta^*\right)^2\right) < 0 \Rightarrow$ impossible, les cas 122 et 123 sont exclus. Le même raisonnement sur A2 et B1 nous permet d'éliminer les cas 212 et 213.
- Si on a A2 et B2, alors on obtiendrait $\cos\theta^* < -\frac{\rho_1^+}{\rho_2^+}$ et $\cos\theta^* < -\frac{\rho_2^+}{\rho_1^+}$, ce qui est impossible, on exclut alors 221, 222 et 223.

- La combinaison A3 et B3 et C1 implique que $\rho_1^- - \rho_2^- > 0$ et $\rho_1^- - \rho_2^- < 0$, ce qui est impossible ; on élimine alors 331.

- Sans perte de généralité, on peut toujours permuter Z_1 et Z_2 afin d'obtenir $\rho_1^+ > \rho_2^-$ (le cas où les modules sont ponctuels et égaux est étudié plus loin), ce choix nous permet alors d'éliminer les cas 231, 232 et 233.

Les cas restants sont alors 111, 132, 133, 321, 312, 322, 332, 313, 323 et 333. Ils sont regroupés dans l'algorithme 2.

Exemple 6 : Le module correspondant à une combinaison (*ijk*) est calculé en utilisant la formule

$$\rho(ijk) = \sqrt{\left(\rho_1^*\right)^2 + \left(\rho_2^*\right)^2 + 2\rho_1^*\rho_2^*\cos\left(\theta^*\right)}$$

Par exemple :

- pour 111, on a $\rho_1^* = \rho_2^*$ et $\theta^* = \pi$, on obtient alors $\rho(111) = 0$

- pour 312, on a $\rho_1^* = \rho_1^-$, $\theta^* = \theta^+$ et $\rho_2^* = -\rho_1^- \cos\theta^+$, on obtient alors

$$\rho(312) = \rho_1^- \sin\theta^+ \qquad \blacklozenge$$

Nous allons maintenant regrouper les 10 candidats restant en testant les conditions C1, C2 et C3, alors :

- Si C1 est vraie i.e. $\pi \in [\theta]$; étant donné que $[\theta]$ est un intervalle de taille n'excédant pas 2π, alors C2 et C3 ne peuvent pas être valides. Les candidats restants sont alors (111) et (321) ; si (111) est acceptable (il suffit de tester si $\exists \rho \in [\rho_1] \cap [\rho_2]$), alors il correspond au minimum global ; sinon, le minimum correspond à (321).

- Si C2 n'est pas vraie, alors (312), (322), (332) et (132) sont éliminés.

- Si C3 n'est pas vraie, alors (313), (323), (333) et (133) sont éliminés.

L'algorithme 2 permet alors de calculer ρ^-. Comme on l'a noté pour l'algorithme 1, il faut à chaque fois vérifier que les trois conditions du minimum sont vérifiées pour chaque combinaison avant de calculer le module correspondant, ceci permettra de réduire le temps de calcul.

Algorithme 2 : MinMod (Z_1, Z_2)

1. Si $\pi \in [\theta_2]$

1.1 Si $[\rho_1] \cap [\rho_2] \neq \varnothing$

retourner $\rho(111)$; fin ; $(\rho(111) = 0)$

1.2 Si $\rho_1^- > \rho_2^+$

retourner $\rho(321)$; fin ;

2. Si $\sin(\theta^+) > 0$ et $\sin(\theta^-) < 0$

$\rho_{i1} = \min(\rho(312), \rho(322), \rho(332), \rho(132))$;

$\rho_{i2} = \min(\rho(313), \rho(323), \rho(333), \rho(133))$;

retourner $\min(\rho_{i1}, \rho_{i2})$; fin;

3. Si $\sin(\theta^+) > 0$

retourner $\min(\rho(312), \rho(322), \rho(332), \rho(132))$; fin ;

4. Si $\sin(\theta^-) < 0$

retourner $\min(\rho(313), \rho(323), \rho(333), \rho(133))$; fin ;

5.5 Minimum de l'argument φ^-

Pour calculer l'argument, il suffit de résoudre le problème d'optimisation (11). Ceci revient à chercher les paramètres x^*, θ_1^* et θ_2^* permettant de minimiser la fonction *g* définie par l'équation (7). Etant donné que la fonction atan(.) est strictement croissante, il suffit de minimiser la fonction *h* suivante :

$$h(x, \theta_1, \theta_2) = \frac{x \sin\theta_1 + \sin\theta_2}{x \cos\theta_1 + \cos\theta_2} \tag{17}$$

Remarque : Nous avons posé $[x] = [\rho_1] / [\rho_2]$ afin de réduire le nombre de variables à 3 ; il est possible d'avoir $\rho_2^- = 0$ et ainsi $x = +\infty$. Néanmoins, ceci ne pose aucun problème étant donné que la fonction *h* et ses dérivées ont des limites finies en $x = +\infty$; les conditions d'optimalité peuvent donc être vérifiées en $x = +\infty$.

5.5.1 Minimisation par rapport à *x*

Les dérivées première et seconde de la fonction *h* par rapport à *x* sont données par :

$$\frac{\partial h}{\partial x} = \frac{\sin(\theta_1^\circ - \theta_2^\circ)}{(x \cos\theta_1^\circ + \cos\theta_2^\circ)^2} \tag{18}$$

et

$$\frac{\partial^2 h}{\partial x^2} = \frac{-2\sin\left(\theta_1^\circ - \theta_2^\circ\right)\cos\theta_1^\circ}{\left(x\cos\theta_1^\circ + \cos\theta_2^\circ\right)^3} \tag{19}$$

La dérivée de *h* par rapport à *x* s'annule lorsque $\sin\left(\theta_1^\circ - \theta_2^\circ\right) = 0$, le minimum est alors indépendant de *x*. Ce résultat est valable pour deux situations : $\theta_1^\circ = \theta_2^\circ$ ou $\theta_1^\circ = \theta_2^\circ + \pi$ et on obtient :

$$h\left(x, \theta_1^\circ, \theta_2^\circ\right) = \left(\frac{x\sin\theta_1^\circ \pm \sin\theta_1^\circ}{x\cos\theta_1^\circ \pm \cos\theta_1^\circ}\right) = \tan\theta_1^\circ$$

pour $x \neq 1$.

Maintenant, si $\sin\left(\theta_1^\circ - \theta_2^\circ\right) \neq 0$, deux cas peuvent se présenter :

- Si $\sin\left(\theta_1^\circ - \theta_2^\circ\right) > 0$, le minimum de l'argument de $\hat{Z}$ correspond à $x^* = x^-$.
- Si $\sin\left(\theta_1^\circ - \theta_2^\circ\right) < 0$, le minimum de l'argument de $\hat{Z}$ correspond à $x^* = x^+$.

5.5.2 Minimisation par rapport à θ_1

Les dérivées première et seconde de la fonction *h* par rapport à θ_1 sont données par :

$$\frac{\partial h}{\partial \theta_1} = \frac{x^\circ + \cos\left(\theta_1 - \theta_2^\circ\right)}{\left(x^\circ\cos\theta_1 + \cos\theta_2^\circ\right)^2} \tag{20}$$

et

$$\frac{\partial^2 h}{\partial \theta_1^2}\left(x^\circ, \theta_1, \theta_2^\circ\right) = \frac{-x^\circ\sin\left(\theta_1 - \theta_2^\circ\right)}{\left(x^\circ\cos\theta_1 + \cos\theta_2^\circ\right)^2} + 2\frac{\left(x^\circ\right)^2\left(x^\circ + \cos\left(\theta_1 - \theta_2^\circ\right)\right)\sin\theta_1}{\left(x^\circ\cos\theta_1 + \cos\theta_2^\circ\right)^3} \tag{21}$$

Lorsque $\frac{\partial h}{\partial \theta_1} = 0$, le deuxième terme de (21) est nul ; dans ce cas la dérivée seconde de *h* à étudier est donnée par :

$$\frac{\partial^2 h}{\partial \theta_1^2}\left(x^\circ, \theta_1, \theta_2^\circ\right) = \frac{-x^\circ\sin\left(\theta_1 - \theta_2^\circ\right)}{\left(x^\circ\cos\theta_1 + \cos\theta_2^\circ\right)^2} \tag{22}$$

Donc si

$$\exists \theta_1^* \in [\theta_1] \ | \ x^\circ + \cos\left(\theta_1^* - \theta_2^\circ\right) = 0 \text{ et } \sin\left(\theta_1^* - \theta_2^\circ\right) < 0$$

alors le minimum est atteint pour $\theta_1 = \theta_1^*$.

Dans le cas contraire, le minimum est atteint pour :

- $\theta_1^* = \theta_1^+$, si $x^\circ + \cos\left(\theta_1^+ - \theta_2^\circ\right) < 0$
- $\theta_1^* = \theta_1^-$, si $x^\circ + \cos\left(\theta_1^- - \theta_2^\circ\right) > 0$

5.5.3 Minimisation par rapport à θ_2

Les dérivées première et seconde de *h* par rapport à θ_2 avec $\theta_1 = \theta_1^\circ$ et $x = x^\circ$ sont données par :

$$\frac{\partial h}{\partial \theta_2} = \frac{x^\circ \cos\left(\theta_1^\circ - \theta_2\right) + 1}{\left(x \cos\theta_1^\circ + \cos\theta_2\right)^2} \tag{23}$$

et

$$\frac{\partial^2 h}{\partial \theta_2^{\,2}}\left(x^\circ, \theta_1^\circ, \theta_2\right) = \frac{x^\circ \sin\left(\theta_1^\circ - \theta_2\right)}{\left(x^\circ \cos\theta_1^\circ + \cos\theta_2\right)^2} + 2\frac{\left(1 + x^\circ \cos\left(\theta_1^\circ - \theta_2\right)\right)\sin\theta_2}{\left(x^\circ \cos\theta_1^\circ + \cos\theta_2\right)^3} \tag{24}$$

Lorsque $\frac{\partial h}{\partial \theta_2} = 0$, la dérivée seconde de *h* devient :

$$\frac{\partial^2 h}{\partial \theta_2^{\,2}}\left(x^\circ, \theta_1^\circ, \theta_2\right) = \frac{x^\circ \sin\left(\theta_1^\circ - \theta_2\right)}{\left(x^\circ \cos\theta_1^\circ + \cos\theta_2\right)^2} \tag{25}$$

La dérivée de *h* par rapport à θ_2 s'annule si :

$$\exists \theta_2^* \in [\theta_2] \ | \ x^\circ \cos\left(\theta_1^\circ - \theta_2^*\right) + 1 = 0$$

Si en plus θ_2^* vérifie

$$\sin\left(\theta_1^\circ - \theta_2^*\right) > 0$$

alors $\frac{\partial^2 h}{\partial \theta_2^{\,2}}\left(x^\circ, \theta_1^\circ, \theta_2^*\right) > 0$, la valeur $\theta_2 = \theta_2^*$ correspond au minimum global.

Dans le cas contraire, le minimum correspond à l'une des extrémités de $[\theta_2]$. Deux cas peuvent alors se présenter :

- Le minimum correspond à $\theta_2^* = \theta_2^-$, si $x^\circ \cos\left(\theta_1^\circ - \theta_2^-\right) + 1 > 0$

- Il correspond à $\theta_2^* = \theta_2^+$, si $x^\circ \cos\left(\theta_1^\circ - \theta_2^+\right) + 1 < 0$

5.5.4 Calcul du minimum de φ

Dans les paragraphes précédents, les conditions d'obtention du minimum de φ par rapport à chacune des trois variables *x*, θ_1 et θ_2 ont été étudiées. Dans ce paragraphe ces conditions seront examinées afin de calculer le minimum de la fonction *h*. D'abord, on rappelle que pour calculer le minimum, les variables *x*, θ_1 et θ_2 sont choisies comme suit :

- $x^* = x^+$, si $\sin\left(\theta_1^* - \theta_2^*\right) < 0$ (A1)

- $x^* = x^-$, si $\sin\left(\theta_1^* - \theta_2^*\right) > 0$ (A2)

- indépendant de *x*, si $\sin\left(\theta_1^* - \theta_2^*\right) = 0$ (A3)

- $\theta_1^* = \theta_1^+$, si $x^* + \cos\left(\theta_1^+ - \theta_2^*\right) < 0$ (B1)

- $\theta_1^* = \theta_1^-$, si $x^* + \cos\left(\theta_1^* - \theta_2^*\right) > 0$ (B2)

- $\theta_1^* = \theta_2^* + \cos^{-1}\left(-x^*\right)$, si $x^* < 1$ et $\sin\left(\theta_1^* - \theta_2^*\right) < 0$ (B3)

- $\theta_2^* = \theta_2^+$, si $x^* \cos\left(\theta_1^* - \theta_2^+\right) + 1 < 0$ (C1)

- $\theta_2^* = \theta_2^-$, si $x^* \cos\left(\theta_1^* - \theta_2^-\right) + 1 > 0$ (C2)

- $\theta_2^* = \theta_1^* - \cos^{-1}\left(-1/x^*\right)$, si $1/x^* < 1$ et $\sin\left(\theta_1^* - \theta_2^*\right) > 0$ (C3)

Pour trouver le minimum de l'angle φ, il suffit alors de trouver les angles correspondant aux 27 cas résultant des différentes combinaisons $A_iB_jC_k$. Néanmoins, plusieurs cas sont impossibles et ceci réduira donc le nombre de combinaisons à étudier.

Chapitre 2

On commence d'abord par permuter les secteurs Z_1 et Z_2 de manière à avoir $x^+ > 1$. Cette dernière condition implique que B3 et B1 ne peuvent pas être satisfaites, ceci permet alors d'éliminer les combinaisons 111, 112, 113, 131, 132 et 133. D'autre part, B3 et C3 sont contradictoires, donc les cas 233, 333 sont éliminés. Si en plus le cas A3 est valide, alors les cas 311, 313, 323, 331 et 332 sont impossibles. D'autre part, les cas A2 et B3 sont contradictoires, on exclut 231 et 232. En outre, les cas A1 et C3 sont contradictoires, on supprime 123. De plus, les cas B1 et C1 sont contradictoires ($x^* > 1$ et $x^* < 1$), on élimine alors 211. Enfin, la combinaison 213 donne lieu à une contradiction ($x^- < 1$ et $x^- > 1$), elle est alors éliminée.

Les combinaisons restantes sont alors 121, 122, 212, 213, 221, 222, 312, 321 et 322. Les différentes conditions correspondantes sont données dans le tableau 1.

<u>Analyse de (312), (321) et (322) :</u>

Ces 3 candidats correspondent à une indétermination en x°.

Le candidat (312) correspond à l'ensemble des conditions suivantes :

$$\begin{cases} \sin\left(\theta_1^+ - \theta_2^-\right) = 0 \\ x^\circ + \cos\left(\theta_1^+ - \theta_2^-\right) < 0 \\ x^\circ \cos\left(\theta_1^+ - \theta_2^-\right) + 1 > 0 \end{cases} \tag{26}$$

Ces conditions imposent que $\cos\left(\theta_1^+ - \theta_2^-\right) = -1$ et $x^\circ < 1$; donc on peut choisir $x^\circ = x^-$, ainsi ce candidat est un cas de « bord » du candidat (212), si pour ce dernier on étend la condition d'acceptabilité $\sin\left(\theta_1^+ - \theta_2^-\right) > 0$ à $\sin\left(\theta_1^+ - \theta_2^-\right) \geq 0$. ♦

De la même manière le candidat (321) est un cas limite de (121) et (322) de (122).

Les candidats restants sont alors : (121), (122), (212), (221), (222) et (223) ; leurs conditions d'optimalité sont données par le tableau 1.

Pour calculer φ^-, il suffit alors de choisir parmi les candidats présentés dans le tableau 1 vérifiant les conditions du minimum. Chacun des candidats permet de calculer un angle, le minimum des angles calculées est alors φ^-. Néanmoins, plusieurs combinaisons ne peuvent pas être satisfaites simultanément.

On remarque qu'on ne peut pas avoir simultanément 121 et 221 ($\sin\left(\theta_1^- - \theta_2^+\right) < 0$ et $\sin\left(\theta_1^- - \theta_2^+\right) > 0$) ; de même, les combinaisons 122 et 222 ne peuvent pas être faisables en

même temps ($\sin\left(\theta_1^- - \theta_2^-\right) < 0$ et $\sin\left(\theta_1^- - \theta_2^-\right) > 0$). Le minimum correspond alors au minimum de l'un des quadruplets suivants :

- $\varphi(121)$, $\varphi(122)$, $\varphi(212)$, $\varphi(223)$
- ou, $\varphi(121)$, $\varphi(212)$, $\varphi(222)$, $\varphi(223)$
- ou, $\varphi(221)$, $\varphi(122)$, $\varphi(212)$, $\varphi(223)$
- ou, $\varphi(221)$, $\varphi(212)$, $\varphi(222)$, $\varphi(223)$

Candidats	**Variables**	**Conditions**
121	$x^+,\ \theta_1^-,\ \theta_2^+$	$\sin\left(\theta_1^- - \theta_2^+\right) \le 0$ $x^+ + \cos\left(\theta_1^- - \theta_2^+\right) > 0$ $x^+ \cos\left(\theta_1^- - \theta_2^+\right) + 1 < 0$
122	$x^+,\ \theta_1^-,\ \theta_2^-$	$\sin\left(\theta_1^- - \theta_2^-\right) \le 0$ $x^+ + \cos\left(\theta_1^- - \theta_2^-\right) > 0$ $x^+ \cos\left(\theta_1^- - \theta_2^-\right) + 1 > 0$
212	$x^-,\ \theta_1^+,\ \theta_2^-$	$\sin\left(\theta_1^+ - \theta_2^-\right) \ge 0$ $x^- + \cos\left(\theta_1^+ - \theta_2^-\right) < 0$ $x^- \cos\left(\theta_1^+ - \theta_2^-\right) + 1 > 0$
213	$x^-,\ \theta_1^+,\ \theta_2^*$ $\theta_2^* = \theta_1^+ - \cos^{-1}\left(-1/x^-\right)$	$\sin\left(\theta_1^+ - \theta_2^*\right) > 0$ $x^- + \cos\left(\theta_1^+ - \theta_2^*\right) < 0$ $x^- > 1$
221	$x^-,\ \theta_1^-,\ \theta_2^+$	$\sin\left(\theta_1^- - \theta_2^+\right) > 0$ $x^- + \cos\left(\theta_1^- - \theta_2^+\right) > 0$ $x^- \cos\left(\theta_1^- - \theta_2^+\right) + 1 < 0$
222	$x^-,\ \theta_1^-,\ \theta_2^-$	$\sin\left(\theta_1^- - \theta_2^-\right) > 0$ $x^- + \cos\left(\theta_1^- - \theta_2^-\right) > 0$ $x^- \cos\left(\theta_1^- - \theta_2^-\right) + 1 > 0$

Tableau 1 : Candidats permettant de calculer le minimum de φ et les conditions à satisfaire.

Le minimum de φ est alors calculé par l'algorithme suivant :

Algorithme 3 : MinArg(Z_1, Z_2)

1. Si $\rho_1^+ \,/\, \rho_2^- \leq 1$

   Permuter (Z_1, Z_2);

2. Si $\pi \in]\theta[$ et $]\rho_1[\cap]\rho_2[\neq \varnothing$

   retourner 0; fin;

3. Si $\sin\left(\theta_1^- - \theta_2^+\right) \leq 0$

   a) Si $\sin\left(\theta_1^- - \theta_2^-\right) \leq 0$

      retourner min(φ(121), φ(122), φ(212), φ(223)) ; fin ;

   b) Si $\sin\left(\theta_1^- - \theta_2^-\right) > 0$

      retourner min(φ(121), φ(212), φ(222), φ(223)) ; fin ;

4. Si $\sin\left(\theta_1^- - \theta_2^+\right) > 0$

   a) Si $\sin\left(\theta_1^- - \theta_2^-\right) \leq 0$

      retourner min(φ(221), φ(122), φ(212), φ(223)) ; fin ;

   b) Si $\sin\left(\theta_1^- - \theta_2^-\right) > 0$

      retourner min(φ(221), φ(212), φ(222), φ(223)) ; fin ;

5.6 Maximum de l'argument φ^+

Pour calculer le maximum de l'argument, il suffit de résoudre le problème d'optimisation (12) sous contraintes (13). Dans cette partie, on va donc trouver les paramètres x^*, θ_1^* et θ_2^* permettant de maximiser la fonction *h* définie par (17).

5.6.1 Maximisation par rapport à *x*

On considère la fonction *h* ainsi que ses dérivées définies dans la section 5.5., on vérifie alors que lorsque $\frac{\partial h}{\partial x} = 0$, le maximum ne dépend pas de la variable *x*. Ceci est vrai pour $\theta_1^\circ = \theta_2^\circ$ et

pour $\theta_1^\circ = \theta_2^\circ + \pi$. Pour ces deux cas, la valeur de x° est indifférente, on obtient alors $\varphi^+ = \theta_1^\circ$ (même raisonnement que pour le calcul du minimum de φ^-).

Maintenant si $\sin\left(\theta_1^\circ - \theta_2^\circ\right) \neq 0$, alors $\frac{\partial h}{\partial x}$ garde un signe constant sur l'intervalle $[x]$; *h* est alors strictement monotone et elle atteint son maximum pour l'une des bornes de l'intervalle $[x]$, on obtient alors deux cas :

- Si $\sin\left(\theta_1^\circ - \theta_2^\circ\right) > 0$, la fonction *h* atteint son maximum pour $x^* = x^+$.

- Si $\sin\left(\theta_1^\circ - \theta_2^\circ\right) < 0$, la fonction *h* atteint son maximum pour $x^* = x^-$.

5.6.2 Maximisation par rapport à θ_1 et θ_2

On considère la fonction *h* définie dans la section 5.5. En examinant les dérivées de *h*, on remarque que le maximum peut être atteint pour $\theta_1^* \in [\theta_1]$ si $x^\circ + \cos\left(\theta_1^* - \theta_2^\circ\right) = 0$ et $\sin\left(\theta_1^* - \theta_2^\circ\right) > 0$. Dans le cas contraire, le maximum est atteint pour :

- $\theta_1^* = \theta_1^+$, si $x^\circ + \cos\left(\theta_1^+ - \theta_2^\circ\right) > 0$
- $\theta_1^* = \theta_1^-$, si $x^\circ + \cos\left(\theta_1^- - \theta_2^\circ\right) < 0$

Le même raisonnement sur θ_2 permet de montrer que le maximum est obtenu avec :

- θ_2^*, si $\exists \theta_2^* \in [\theta_2] \mid x^\circ \cos\left(\theta_1^\circ - \theta_2^*\right) + 1 = 0$ et $\sin\left(\theta_1^\circ - \theta_2^*\right) < 0$
- $\theta_2^* = \theta_2^-$, si $x^\circ \cos\left(\theta_1^\circ - \theta_2^-\right) + 1 < 0$
- $\theta_2^* = \theta_2^+$, si $x^\circ \cos\left(\theta_1^\circ - \theta_2^+\right) + 1 > 0$

5.6.3 Synthèse

Les conditions de maximisation de φ sur les paramètres *x*, θ_1, θ_2 sont

- $x^* = x^+$, si $\sin\left(\theta_1^* - \theta_2^*\right) > 0$ (A1)

- $x^* = x^-$, si $\sin\left(\theta_1^* - \theta_2^*\right) < 0$ (A2)

- indépendant de *x*, si $\sin\left(\theta_1^* - \theta_2^*\right) = 0$ (A3)

- $\theta_1^* = \theta_1^+$, si $x^* + \cos\left(\theta_1^+ - \theta_2^*\right) > 0$ (B1)

- $\theta_1^* = \theta_1^-$, si $x^* + \cos\left(\theta_1^- - \theta_2^*\right) < 0$ (B2)
- $\theta_1^* = \theta_2^* + \cos^{-1}\left(-x^*\right)$, si $x^* < 1$ et $\sin\left(\theta_1^* - \theta_2^*\right) > 0$ (B3)
- $\theta_2^* = \theta_2^+$, si $x^* \cos\left(\theta_1^* - \theta_2^+\right) + 1 > 0$ (C1)
- $\theta_2^* = \theta_2^-$, si $x^* \cos\left(\theta_1^* - \theta_2^-\right) + 1 < 0$ (C2)
- $\theta_2^* = \theta_1^* - \cos^{-1}\left(-1/x^*\right)$, si $1/x^* < 1$ et $\sin\left(\theta_1^* - \theta_2^*\right) < 0$ (C3)

La combinaison de ces différentes conditions engendre 27 possibilités. Néanmoins, plusieurs combinaisons ne sont pas possibles comme pour le minimum de φ.

Supposons que A3 est vraie, donc $\theta^* = 0$ ou $\theta^* = \pi$, alors

- Si $\theta^* = 0$, les conditions B2, B3, C2 et C3 ne sont pas possibles ; seule la combinaison 311 est alors possible.
- Si $\theta^* = \pi$, seules les combinaisons 312 et 321 sont possibles.

Donc, les combinaisons 313, 322, 323, 331, 332 et 333 sont écartées.

D'autre part, on remarque que les conditions A1 et C3 sont incompatibles, les cas 113, 123 et 133 sont alors exclus. De même, A2 est incompatible avec B3, les combinaisons 231, 232 et 233 ne peuvent pas être admissibles. De plus, on peut, sans perte de généralité, supposer que $x^+ > 1$ ceci permet d'exclure 121, 122, 132 et 131. Les conditions B2 et C3 sont contradictoires, en effet, on obtient $x^- < 1$ et $x^- > 1$, ce qui permet alors d'éliminer 223. Enfin le même raisonnement permet d'exclure 222. Les cas restants sont alors : 111, 112, 211, 212, 213, 221, 311, 312 et 321.

Remarque : Les candidats 311, 312 et 321 correspondent à une indétermination en x° ; comme on l'a montré dans le cas de φ^-, ces cas peuvent être fusionnés respectivement avec 111, 112 et 221.

Les candidats restant ainsi que les conditions à satisfaire sont donnés dans le tableau 2 :

Candidats	Variables	Conditions
111	x^+, θ_1^+, θ_2^+	$\sin\left(\theta_1^+ - \theta_2^+\right) \geq 0$ $x^+ + \cos\left(\theta_1^+ - \theta_2^+\right) > 0$ $x^+ \cos\left(\theta_1^+ - \theta_2^+\right) + 1 > 0$
112	x^+, θ_1^+, θ_2^-	$\sin\left(\theta_1^+ - \theta_2^-\right) \geq 0$ $x^+ + \cos\left(\theta_1^+ - \theta_2^-\right) > 0$ $x^+ \cos\left(\theta_1^+ - \theta_2^-\right) + 1 < 0$
211	x^-, θ_1^+, θ_2^+	$\sin\left(\theta_1^+ - \theta_2^+\right) < 0$ $x^- + \cos\left(\theta_1^+ - \theta_2^+\right) > 0$ $x^- \cos\left(\theta_1^+ - \theta_2^+\right) + 1 > 0$
212	x^-, θ_1^+, θ_2^-	$\sin\left(\theta_1^+ - \theta_2^-\right) < 0$ $x^- + \cos\left(\theta_1^+ - \theta_2^-\right) > 0$ $x^- \cos\left(\theta_1^+ - \theta_2^-\right) + 1 < 0$
213	x^- , θ_1^+, θ_2^* $\theta_2^* = \theta_1^+ - \cos^{-1}(-1/x^-)$	$\sin\left(\theta_1^+ - \theta_2^*\right) < 0$ $x^- + \cos\left(\theta_1^+ - \theta_2^*\right) > 0$ $x^- > 1$
221	x^- , θ_1^-, θ_2^+	$\sin\left(\theta_1^- - \theta_2^+\right) \leq 0$ $x^- + \cos\left(\theta_1^- - \theta_2^+\right) < 0$ $x^- \cos\left(\theta_1^- - \theta_2^+\right) + 1 > 0$

Tableau 2 : Candidats permettant de calculer le maximum de φ et les conditions à satisfaire.

Le maximum de l'angle φ est obtenu en calculant les angles définis par les candidats présentés dans le tableau précédent, il est alors calculé par l'algorithme suivant :

Algorithme 4 : MaxArg(Z_1, Z_2)

1. Si $\rho_1^+ \,/\, \rho_2^- \leq 1$

   Permuter (Z_1, Z_2);

2. Si $\pi \in]\theta[$ et $]\rho_1[\cap]\rho_2[\neq \varnothing$

Retourner 2π ; fin;

3. Si $\sin\left(\theta_1^+ - \theta_2^-\right) \geq 0$

   a) Si $\sin\left(\theta_1^+ - \theta_2^+\right) \geq 0$

      Retourner max(φ(112), φ(111), φ(221), φ(213)) ; fin ;

   b) Si $\sin\left(\theta_1^+ - \theta_2^+\right) < 0$

      Retourner min(φ(112), φ(211), φ(221), φ(213)) ; fin ;

4. Si $\sin\left(\theta_1^+ - \theta_2^-\right) < 0$

   a) Si $\sin\left(\theta_1^+ - \theta_2^+\right) \geq 0$

      Retourner min(φ(212), φ(111), φ(221), φ(213)) ; fin ;

   b) Si $\sin\left(\theta_1^+ - \theta_2^+\right) < 0$

      Retourner min(φ(212), φ(211), φ(221), φ(213)) ; fin ;

6 Algorithme général

L'algorithme général permettant de calculer la somme minimale de deux secteurs Z1 et Z2 est donné par :

Algorithme 5 : Somme(Z_1, Z_2)

1. $\rho^- =$ MinMod(Z_1, Z_2);
2. $\rho^+ =$ MaxMod(Z_1, Z_2);
3. $\varphi^- =$ MinArg(Z_1, Z_2);
4. $\varphi^+ =$ MaxArg(Z_1, Z_2);
5. Retourner $\left\{\left[\rho^-, \rho^+\right], \left[\varphi^-, \varphi^+\right]\right\}$;

7. **Exemples numériques**

Dans cette section, nous allons illustrer l'intérêt de l'arithmétique des intervalles complexes polaires sur quelques exemples numériques.

Exemple 1: Considérons les deux secteurs Z_1 et Z_2 définis par

$$Z_1 = \{[1, 2] ; [\pi/6, \pi/3]\} \text{ et } Z_2 = \{[3, 5] ; [5\pi/4, 11\pi/8]\}$$

Soient Z, le résultat de la somme de Z_1 et Z_2 en utilisant les algorithmes proposés dans ce chapitre et Z', le résultat de la somme obtenue en passant par la forme rectangulaire. Ces deux secteurs sont représentés sur la figure 5. Pour calculer Z', il faut d'abord trouver les plus petits rectangles contenant Z_1 et Z_2, puis effectuer l'addition, enfin on retourne à la forme polaire afin d'obtenir le secteur Z' (plus petit secteur incluant le résultat $Z_1 + Z_2$ en représentation rectangulaire). Il faut noter alors qu'à chaque passage d'une représentation à une autre, un pessimisme est introduit. Dans notre exemple, l'utilisation de la représentation polaire permet d'obtenir un secteur Z dont la surface est de 58% plus petite que celle de Z'.

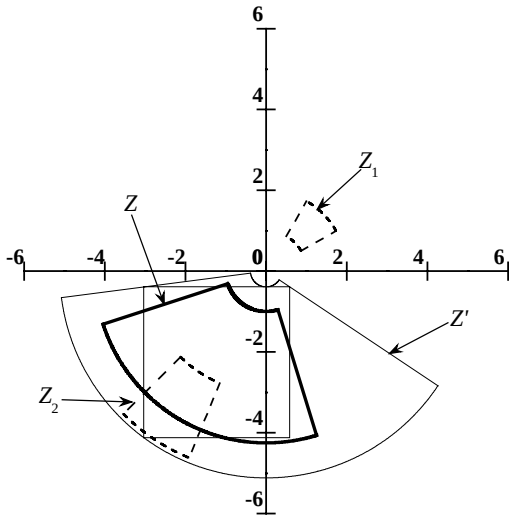

Figure 5 : La somme des secteurs Z_1 et Z_2 donnés dans l'exemple 1; comparaison des représentations rectangulaire et polaire

Exemple 2: Considérons maintenant les secteurs suivants :

$$Z_1 = \{[3, 3.5] ; [0, 2\pi]\},\ Z_2 = \{[4.5, 5] ; [\pi/2, 3\pi/2]\} \text{ et } Z = Z_1 + Z_2$$

En utilisant les algorithmes proposés dans ce chapitre, on obtient

$$Z = \{[1, 8.5] ; [0.679674, 5.60352]\}$$

Par contre, en passant par la forme rectangulaire, on obtient

$$Z^{'} = \{[0, 12.0209] ; [0, 2\pi]\}$$

Les deux secteurs Z et Z' sont représentés sur la figure 6. Le secteur *Z*' est alors un disque, ceci est dû au pessimisme introduit à chaque conversion. Dans notre exemple, l'utilisation de la représentation polaire nous a permis de réduire le pessimisme de 61%. D'autre part, étant donné que $0 \in Z$, la fonction suivante

$$l\colon \mathbb{S}(\mathbb{C})\times\mathbb{S}(\mathbb{C}) \to \mathbb{S}(\mathbb{C})$$

$$(X, Y) \mapsto \log(X + Y)$$

ne peut pas être évaluée sur les deux secteurs Z_1 et Z_2 en passant par la représentation rectangulaire. Par contre, en utilisant les algorithmes présentés dans ce chapitre, la fonction peut être évaluée (puisque $0 \notin Z^{'}$).

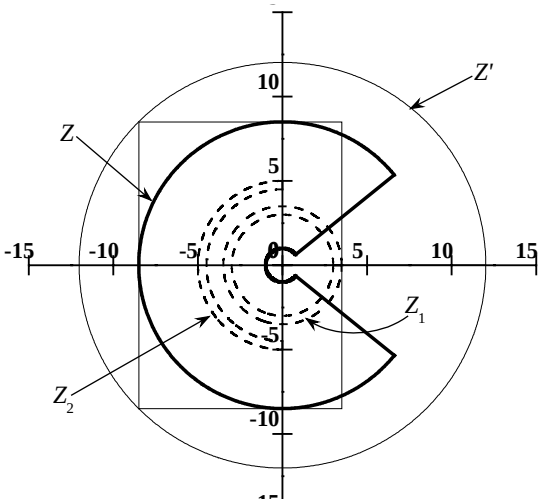

Figure 6: La somme des secteurs Z_1 et Z_2 donnés dans l'exemple 2; avantage de la forme polaire

Exemple 3: Considérons la fonction complexe suivante [RIRC04], appelée fonction de Havriliak-Negami

$$\varepsilon(\omega) = \varepsilon_\infty + \frac{\Delta\varepsilon}{\left(1+(j\omega\tau)^{\alpha}\right)^{\beta}}, \qquad j^2 = -1 \tag{27}$$

où les paramètres ε_∞, $\Delta\varepsilon$, τ, α et β peuvent être incertains. Supposons que l'on ait :

$$\varepsilon_\infty = 3,\ \Delta_\varepsilon = 6,\ \alpha = [0.1, 1],\ \beta = 1,\ \mathrm{Ln}(\tau) = [-7, 5] \text{ et } \omega = 2\pi 10^{-2}\text{rad/s}.$$

- L'évaluation de la fonction de Havriliak-Négami en utilisant les algorithmes présentés dans ce chapitre donne :

$$\varepsilon(\omega) = \{[3.11619, 9] ; [5.26021, 6.28319]\}$$
$$\equiv [1.62299, 9] + j[-7.68298, -4.5361\times10^{-6}]$$

- L'évaluation de cette fonction en passant par la représentation rectangulaire donne

$$\varepsilon(\omega) = [3.00169, 7300.49] + j[-7255.89, -1.38519\times10^{-8}].$$

Maintenant, supposons que $\omega = 2\pi10^6$rad/s. Dans ce nouveau cas :

- en utilisant les algorithmes présentés dans ce chapitre, on obtient :

$$\begin{aligned}\varepsilon(\omega) &= \{ [3.00001, 5.31976] ; [5.6233, 6.2832]\} \\ &\equiv [2.37019, 5.31976] + j[-3.26116, -7.1055\times10^{-10}]\end{aligned}$$

- en utilisant la représentation rectangulaire, on obtient

$$\varepsilon(\omega) = [3, 5.2553\times10^{27}]+ j[-5.2553\times10^{27}, -7.529\times10^{-28}]$$

Nous remarquons donc encore qu'en utilisant uniquement la représentation polaire, on obtient des résultats bien moins pessimistes. Le passage par la forme rectangulaire pour effectuer l'addition de deux secteurs peut générer des résultats numériquement dramatiques.

8. Conclusion

Ce chapitre a été consacré aux intervalles complexes ; dans la littérature, les représentations rectangulaires et circulaires sont les plus utilisées ; pour ces deux formes, l'addition et la soustraction sont définies d'une manière exacte. Néanmoins, la multiplication et la division sont obtenues par des approximations extérieures ; du pessimisme est alors introduit à chaque évaluation. Ces formes sont spécialement efficaces pour l'évaluation de fonctions linéaires. Dans plusieurs domaines d'applications, comme par exemple pour l'analyse de propriétés diélectriques (cas du modèle d'Havriliak-Négami), les modèles sont fortement non-linéaires et l'utilisation des représentations rectangulaire ou circulaire est trop pessimiste.

Dans ce chapitre, nous avons étendu la représentation polaire des nombres complexes aux cas des intervalles, on parle alors de secteurs L'avantage majeur de cette représentation est que la multiplication et la division sont définies d'une manière exacte ; d'où l'intérêt de l'employer lors de l'évaluation de fonctions fortement non-linéaires. Mais l'inconvénient vient du fait que l'addition et la soustraction ne sont plus exactes.

La contribution majeure de ce chapitre est de présenter des algorithmes permettant d'effectuer l'addition de deux secteurs sans utiliser une autre représentation. En particulier, le résultat obtenu est minimal i.e. il n'existe pas de secteur plus petit contenant le résultat. L'opération d'addition a été traitée comme un problème d'optimisation sous contraintes. Dans ce chapitre, le problème d'optimisation a été résolu analytiquement ; ceci est possible puisque le nombre de variables est assez réduit et les fonctions à minimiser sont assez simples. Les algorithmes proposés sont assez simples à mettre en œuvre et quelques tests sur des exemples illustratifs ont montré l'utilité de ces algorithmes. Ils seront alors utilisés dans le chapitre suivant afin de procéder à l'estimation de modèles dans le cadre de l'analyse de spectres de relaxations diélectriques et de l'évaluation de propriétés thermiques de matériaux.

Chapitre 3

Estimation de paramètres physiques dans un contexte à erreurs bornées

1. Introduction

Ce chapitre est consacré à l'application des techniques d'inversion ensembliste par arithmétique d'intervalles à l'estimation de paramètres physiques dans un contexte à erreurs bornées. Les modèles utilisés dans les deux applications considérées sont décrits par des fonctions explicites.

Le but de la première partie du chapitre est d'estimer les paramètres diélectriques du modèle d'Havriliak-Negami [HH97]. Ce modèle est défini comme une somme de plusieurs termes, appelés *relaxations*, et la difficulté rencontrée lors de l'analyse de propriétés diélectriques vient du fait que le nombre de ces relaxations n'est généralement pas connu *a priori*. Nous verrons que l'intérêt des méthodes ensemblistes est de rendre possible la sélection du meilleur modèle permettant d'expliquer les mesures et de rejeter les autres.

Le modèle d'analyse de propriétés diélectriques utilisé dans ce chapitre est non linéaire et à variables complexes. Nous allons considérer deux approches, la première consiste à décomposer la sortie du modèle en une partie réelle et une partie imaginaire ; ainsi l'arithmétique des intervalles réels sera utilisée. La seconde approche consiste à utiliser les intervalles complexes polaires définis au chapitre 2.

Dans le cadre de cette première application, nous allons montrer qu'un choix judicieux de la stratégie de bissection utilisée avec l'algorithme de partitionnement SIVIAP [Jau00] [JKDW01] permet dans certains cas de réduire le temps de calcul. Enfin, nous allons considérer plusieurs cas expérimentaux afin d'évaluer la performance de la méthode d'estimation utilisée.

Dans la deuxième partie du chapitre, nous allons nous intéresser à une deuxième application concernant l'estimation des propriétés thermo-physiques (conductivité et diffusivité thermique) de matériaux par inversion ensembliste. Le modèle utilisé a été développé dans des travaux précédents au sein du laboratoire [TK98] [Bou03]. Ce banc d'essai a déjà fait l'objet d'une caractérisation garantie et des algorithmes d'inversion et de projection ensemblistes ont été développés et appliqués aux données expérimentales issues de ce banc [Bra02] [BRKW03] [BJKRW03]. Dans le cadre de ces travaux, la représentation

rectangulaire des intervalles complexes a été utilisée. Néanmoins, cette représentation est inadaptée pour le modèle utilisé étant donné sa nature fortement non linéaire. Dans ce chapitre, nous allons réaliser une première évaluation de la représentation polaire des intervalles complexes développée dans le chapitre 2 en l'utilisant avec l'algorithme d'inversion ensembliste SIVIAP.

2. Estimation de paramètres de modèles de relaxations diélectriques

Dans la première partie de ce chapitre, nous allons estimer les paramètres inconnus figurant dans le modèle semi-empirique d'Havriliak-Negami [HH97] utilisé pour l'analyse des spectres de relaxation diélectrique. Cette application nous permet de montrer un intérêt majeur des méthodes ensemblistes : elles permettent d'invalider le modèle lorsqu'au moins une mesure est incompatible avec la sortie correspondante du modèle et avec les hypothèses fixées *a priori* pour les bornes d'erreurs. Comme dans le modèle d'Havriliak-Negami le nombre de modes de relaxations n'est pas connu *a priori* nous allons montrer dans cette partie que les méthodes ensemblistes permettent de sélectionner un nombre de modes de relaxation optimal.

Les grandeurs qui sont utilisées dans cette partie sont données par le tableau 1 :

Grandeur	***Symbole***	***Unité***
Facteur de perte diélectrique	$\tan(\delta)$	
Polarisation d'orientation	P_{or}	Cm^{-2}
Permittivité diélectrique du vide	ε_0	Fm^{-1}
Permittivité diélectrique relative réelle	ε'	
Permittivité diélectrique relative imaginaire	ε''	
Permittivité diélectrique relative haute fréquence	ε_∞	
Permittivité diélectrique relative basse fréquence	ε_s	
Permittivité diélectrique relative complexe	ε	
Conductivité électrique	σ_0	Sm^{-1}
Temps de relaxation	τ	s
Dispersion diélectrique	$\Delta\varepsilon$	
Paramètre relatif à la largeur de distribution du temps de relaxation	α	
Paramètre relatif à la dissymétrie de distribution du temps de relaxation	β	

Tableau 1 : Grandeurs diélectriques

Avant de présenter les résultats obtenus, il est nécessaire d'apporter quelques notions générales concernant les propriétés diélectriques des matériaux (en particulier des polymères), la mesure de ces propriétés ainsi que les modèles couramment utilisés pour l'analyse des spectres de relaxation diélectrique.

2.1. Analyse diélectrique

Les méthodes diélectriques sont aujourd'hui largement utilisées pour l'étude et la caractérisation des matériaux polymères, les céramiques, ainsi que d'autres matériaux. Le principal avantage des méthodes diélectriques par rapport aux méthodes mécaniques est la large gamme de fréquences pouvant être étudiée avec une excellente résolution. Aujourd'hui, cette gamme couvre le domaine allant des très basses fréquences (10^{-5} Hz) jusqu'aux fréquences optiques. Le tableau 2 donne un aperçu rapide des techniques diélectriques existantes en fonction du domaine fréquentiel.

Méthode	Gamme de fréquence	Résolution	Gamme d'impédances
Spectroscopie diélectrique (*Frequency Response Analysis*)	10 mHz < f < 10 MHz	tan δ < 10^{-4} [1]	10 mΩ < Z < 10^{14} Ω
Méthodes par ponts d'impédances (*AC-bridge methods*)	20 Hz < f < 1 MHz	tan δ < 5×10^{-4}	1 mΩ < Z < 100 MΩ
Réflectométrie diélectrique (*Coaxial-line reflectometry*)	1 MHz < f < 10 GHz	tan δ < 10^{-3}	-

Tableau 2. Les méthodes d'analyse diélectrique [Sch94].

2.1.1. Mécanismes de polarisation dans les diélectriques

L'application d'un champ électrique sur un matériau diélectrique induit l'apparition d'une polarisation macroscopique du matériau. Plusieurs mécanismes peuvent contribuer à l'établissement de cette polarisation. Chaque mécanisme apporte une contribution à la permittivité diélectrique du matériau dans diverses plages de fréquence. La figure 1 rend compte de ces différents mécanismes en fonction de la fréquence du champ électrique ou de l'onde électromagnétique.

2.1.1.1. Mécanismes Hautes Fréquences

- Polarisation électronique :

Le champ électromagnétique appliqué au matériau provoque un déplacement du nuage électronique à l'échelle de chaque atome. La polarisation électronique est donc due à un décalage du centre de gravité des charges négatives par rapport aux charges positives.

[1] tan δ correspond au facteur de pertes diélectriques (Cf Equation (6))

- Polarisation ionique :

Elle est due à la distorsion de l'arrangement des noyaux atomiques sous l'influence du champ électrique. Elle induit une variation de longueur des liaisons chimiques.

Ces deux composantes définissent la permittivité hautes fréquences que nous notons par ε_∞.

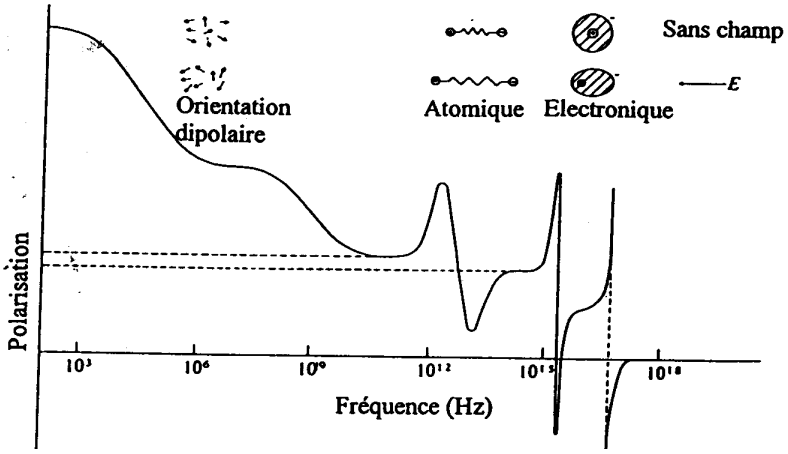

Figure 1. Mécanismes de polarisation dans un diélectrique en fonction de la fréquence [Bly79].

2.1.1.2. Mécanismes Basses Fréquences

- Polarisation d'orientation :

Elle existe lorsque les molécules du matériau possèdent des dipôles permanents. Lors de l'application d'un champ électrique E, les dipôles tendent à s'orienter dans le sens du champ pour donner une polarisation d'orientation P_{or} globale:

$$P_{or} = \varepsilon_0 \left(\varepsilon - \varepsilon_\infty \right) E \tag{1}$$

ε étant la permittivité diélectrique relative du matériau.

- Polarisation par charge d'espace :

Elle trouve son origine dans la migration de charges libres entre des pièges situés à des distances macroscopiques. Elle est observable à très basses fréquences. Elle peut être attribuée à la présence d'hétérogénéités dans le matériau (effet Maxwell-Wagner-Sillars (MWS)), ou bien à l'accumulation de charges par injection entre les électrodes et le diélectrique.

Etant donné que les techniques d'analyse diélectrique dynamique sont des méthodes basses fréquences, seuls les phénomènes de polarisation d'orientation ou par charge d'espace seront observables par ces techniques de mesure.

2.1.2. Principe de l'analyse diélectrique dynamique (ADD)

L'ADD est basée sur la mesure de la réponse électrique d'un matériau à l'application d'un signal électrique alternatif (dans la plupart des cas sinusoïdal).

Lors de l'application d'un champ électrique $E(t)$ variable aux bornes d'un matériau diélectrique, la densité de courant $J(t)$ dans l'échantillon est définie à partir de l'équation de Maxwell :

$$J(t) = \sigma_0 . E(t) + \frac{\partial D(t)}{\partial t} \tag{2}$$

où σ_0 est la conductivité électrique "continue" du matériau et $D(t)$ l'induction électrique [Jon83].

Le premier terme de l'équation (2) définit le courant de conduction et le second terme est le courant de déplacement. En effectuant une Transformée de Fourier de l'équation (2), on obtient l'équation correspondante dans le domaine fréquentiel [Jon83]:

$$J(\omega) = \sigma_0 . E(\omega) + j\omega D(\omega) \tag{3}$$

où ω est la pulsation du signal sinusoïdal et $j^2 = -1$.

En introduisant la permittivité complexe $\varepsilon(\omega) = \varepsilon'(\omega) - j.\varepsilon''(\omega)$ du matériau, exprimée en valeur relative, on obtient :

$$J(\omega) = \left[\sigma_0 + j\omega\varepsilon_0\varepsilon(\omega)\right] E(\omega) \tag{4}$$

où ε_0 est la permittivité diélectrique du vide ($\varepsilon_0 = 8.85\times10^{-12}$ F.m^{-1}).

L'expression (4) devient :

$$J(\omega) = \left[\sigma_0 + \varepsilon_0\varepsilon''\omega + j\omega\varepsilon_0\varepsilon'\right] E(\omega) \tag{5}$$

La partie réelle ε' de la permittivité diélectrique complexe ε rend compte de l'effet capacitif tandis que la partie imaginaire ε'' correspond aux pertes diélectriques. On note par $\tan\delta$, le facteur de pertes diélectriques :

$$\tan\delta = \frac{\varepsilon''}{\varepsilon'} \tag{6}$$

Dans ce cas, le matériau diélectrique peut être schématisé électriquement par une résistance R en parallèle avec une capacité "complexe" C, alimentées par un générateur de tension sinusoïdale $U(\omega)$.

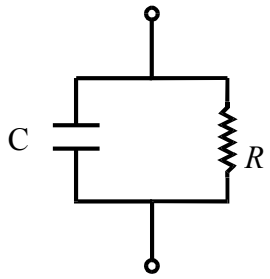

Figure 2. Schéma électrique équivalent d'un matériau diélectrique. [Bly79]

En posant *e* et *S*, comme étant respectivement l'épaisseur et la surface de l'échantillon, on peut écrire:

$$R = \frac{e}{\sigma_0 . S} \qquad \text{et} \qquad C = \varepsilon \frac{\varepsilon_0 . S}{e} = C' - jC'' \tag{7}$$

L'impédance équivalente $Z(\omega)$ du matériau est donnée par:

$$\frac{1}{Z(\omega)} = \frac{1}{R} + j\omega C(\omega) \tag{8}$$

Le courant $I(\omega)$ circulant dans l'échantillon, s'écrit donc:

$$I(\omega) = \left[\left(\frac{1}{R} + \omega C''(\omega) \right) + j\omega C'(\omega) \right] U(\omega) \tag{9}$$

où $C'(\omega)$ et $C''(\omega)$ sont les capacités réelle et imaginaire de l'échantillon.

Le courant se compose donc d'un terme en phase avec la tension appliquée tenant compte des phénomènes de conduction et de pertes diélectriques et d'un courant en quadrature de phase avec le champ, rendant compte de l'effet capacitif.

Expérimentalement, la mesure du courant $I_1(t)=I_1\sin(\omega t)$ en phase avec la tension sinusoïdale appliquée $U(t)=U_0\sin(j\omega t)$ et du courant $I_2(t) = I_2 \sin(\omega t + \pi/2)$ en quadrature de phase avec le champ permettent de remonter aux valeurs des capacités réelle et imaginaire de l'échantillon, donc à la permittivité du matériau (Cf. figure 3).

$$\varepsilon' = \frac{I_2}{U_0 C_0 \omega} \quad \text{et} \quad \varepsilon'' + \frac{\sigma_0}{\varepsilon_0 \omega} = \frac{I_1}{U_0 C_0 \omega} \tag{10}$$

avec C_0, capacité du vide équivalente à l'échantillon:

$$C_0 = \frac{\varepsilon_0 . S}{e} \tag{11}$$

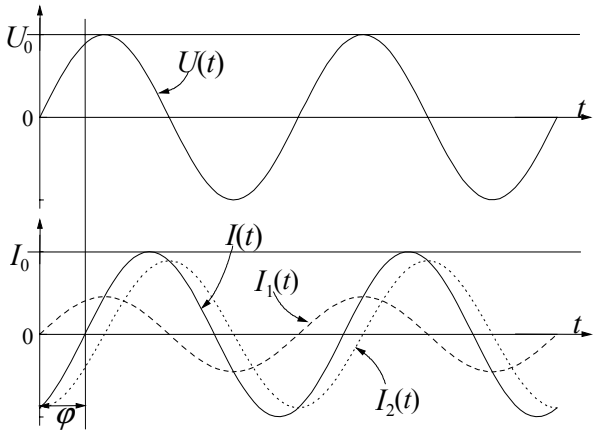

Figure 3. Principe de l'ADD

Nous voyons ici que le terme de conduction $\sigma_0/\varepsilon_0\omega$ ne peut être séparé expérimentalement du terme de pertes diélectriques ε''. Dans le domaine des hautes fréquences et/ou des basses températures, la contribution du terme de conductivité électrique est en général négligeable. En revanche, dans le domaine des basses fréquences et/ou des hautes températures, la contribution du terme de conductivité électrique peut devenir prépondérante, même si les valeurs de σ_0 sont relativement faibles pour les polymères (typiquement $\sigma_0 \in [10^{-16}, 10^{-8}]$ S.m^{-1}).

Dans la suite de ce chapitre, on notera par ε'' la grandeur $\varepsilon'' + \sigma_0/\varepsilon_0\omega$.

2.1.3 Modèles pour l'analyse de spectres de relaxation diélectrique

La relaxation dipolaire au sens de Debye est un processus purement visqueux, sans force de rappel élastique, et est donc du premier ordre [CA93]. La polarisation d'orientation P_{or} d'un ensemble de dipôles en équilibre thermique soumis à un champ électrique E obéit à l'équation:

$$\frac{dP_{or}(t)}{dt} + \frac{P_{or}(t)}{\tau} = \frac{\varepsilon_0(\varepsilon_S - \varepsilon_\infty)}{\tau} E(t) \tag{12}$$

où τ est le temps de relaxation, ε_0 la permittivité diélectrique du vide, ε_S, la permittivité diélectrique relative "statique" (ou basse fréquence) et ε_∞ la permittivité diélectrique relative "instantanée" (ou haute fréquence).

En régime alternatif, le champ électrique et la polarisation d'orientation s'écrivent respectivement:

$$E(t) = E_0.\exp(j\omega t) \quad \text{et} \quad P_{or}(t) = P_{or}.\exp(j\omega t) \tag{13}$$

Ce qui donne pour la polarisation d'orientation (en utilisant la relation (1) et (12)) :

$$P_{or} = \varepsilon_0(\varepsilon - \varepsilon_\infty)E_0 = \frac{\varepsilon_0(\varepsilon_S - \varepsilon_\infty)}{1 + j\omega\tau} E_0 \tag{14}$$

La permittivité complexe relative ε est donc définie par la relation de Debye [CA93] :

$$\varepsilon(\omega) = \varepsilon_\infty + \frac{(\varepsilon_S - \varepsilon_\infty)}{1 + j\omega\tau} = \varepsilon_\infty + \frac{\Delta\varepsilon}{1 + j\omega\tau} \tag{15}$$

$\Delta\varepsilon$ étant la dispersion diélectrique associée à la relaxation. Dans ce cas les parties réelle ε' et imaginaire ε'' de la permittivité complexe s'écrivent :

$$\varepsilon'(\omega) = \varepsilon_\infty + \frac{\Delta\varepsilon}{1 + (\omega\tau)^2} \quad \text{et} \quad \varepsilon''(\omega) = \frac{\Delta\varepsilon}{1 + (\omega\tau)^2}.\omega\tau \tag{16}$$

Le tracé de la variation de ε'' en fonction de ε', appelé diagramme de Cole-Cole (similaire à un diagramme de Nyquist), est un demi-cercle, de centre $[(\varepsilon_S + \varepsilon_\infty)/2;0]$ et de rayon $r = (\varepsilon_S - \varepsilon_\infty)/2$ (figure 4). Le maximum de ε'' est atteint lorsque $\omega\tau = 1$. Les relaxations diélectriques observées dans la plupart des liquides polaires suivent cette courbe théorique. En revanche, dans les polymères, les relaxations diélectriques sont distribuées, c'est à dire qu'elles ne peuvent pas être décrites par un seul temps de relaxation.

Cole et Cole, en 1941, ont proposé une équation semi-empirique rendant compte de cette distribution [Bly79] :

$$\varepsilon(\omega) = \varepsilon_\infty + \frac{\Delta\varepsilon}{1 + (j\omega\tau)^\alpha} \tag{17}$$

Le paramètre α ($0 < \alpha \leq 1$) rend compte de la largeur de la distribution: quand α diminue, la distribution s'élargit; pour $\alpha = 1$, on retrouve l'équation de Debye. Cette équation rend compte du fait que les spectres de relaxation diélectrique, représentés en diagramme de Cole-Cole, sont "aplatis" par rapport à l'équation de Debye (Figure 4).

Afin de tenir compte de la dissymétrie observée sur certains diagrammes de Cole-Cole issus de données expérimentales, Davidson et Cole [DC50] ont proposé une nouvelle équation semi-empirique :

$$\varepsilon(\omega) = \varepsilon_\infty + \frac{\Delta\varepsilon}{(1 + j\omega\tau)^\beta} \tag{18}$$

Le paramètre β ($0 < \beta \leq 1$) rend compte de la dissymétrie de la distribution: quand β diminue, la dissymétrie augmente; pour $\beta = 1$, on retrouve l'équation de Debye.

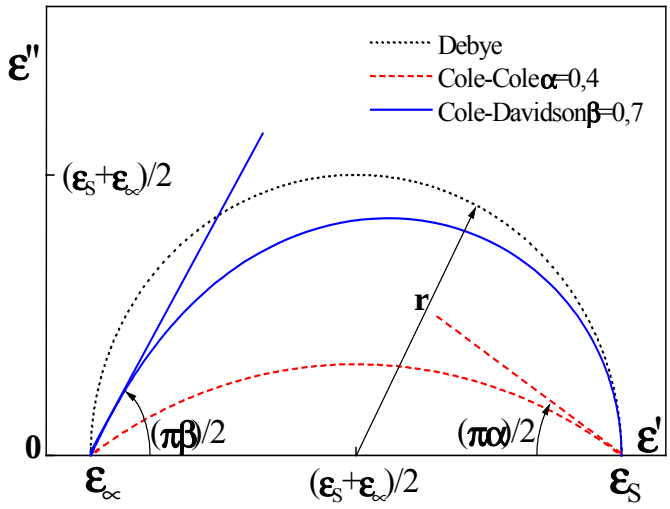

Figure 4. Diagramme de Cole-Cole des équations de Debye, Cole-Cole et Cole-Davidson pour des valeurs de paramètres $\alpha = 0.4$ et $\beta = 0.7$.

La formule proposée par Havriliak et Negami (HN) [HH97] est une combinaison des formules de Cole-Cole et Cole-Davidson. Elle fait intervenir les deux paramètres α et β rendant compte respectivement de la largeur et de la dissymétrie de la fonction de distribution des temps de relaxation [Bly79] :

$$\varepsilon(\omega) = \varepsilon_{\infty} + \frac{\Delta\varepsilon}{\left(1+(j\omega\tau)^{\alpha}\right)^{\beta}} \tag{19}$$

Lorsque plusieurs modes de relaxation se recouvrent mutuellement, un ajustement de la courbe expérimentale par une somme de plusieurs termes d'Havriliak-Negami peut être réalisé pour extraire les contributions respectives des différents processus de relaxation (Equation (20)). De plus, un terme prenant en compte la contribution de la conductivité électrique peut être ajouté. On obtient donc [HH97] :

$$\varepsilon(\omega) = \varepsilon_{\infty} + \sum_{i=1}^{n}\left[\frac{\Delta\varepsilon_i}{\left(1+(j\omega\tau_i)^{\alpha_i}\right)^{\beta_i}}\right] + \frac{\sigma_0}{\varepsilon_0 (j\omega)^{s}} \tag{20}$$

où n est le nombre de modes de relaxations à déterminer (dans la gamme de fréquences considérée, le nombre de modes de relaxation habituellement observées ne dépasse généralement pas 3), et l'exposant s est un paramètre ajustable ($0 < s \leq 1$). A partir de l'expression (20), nous pouvons écrire la permittivité diélectrique complexe sous la forme suivante :

$$\varepsilon(\omega) = \varepsilon'(\omega) - j\varepsilon''(\omega) \;; \quad \text{avec } j^2 = -1 \tag{21}$$

où les parties réelle $\varepsilon'(\omega)$ et imaginaire $\varepsilon''(\omega)$ sont données par :

$$\varepsilon'(\omega) = \varepsilon_\infty + \sum_{i=1}^{n}\left[\Delta\varepsilon_i \frac{\cos(\beta_i\varphi_i)}{\left(1+2(\omega\tau_i)^{\alpha_i}\sin\left(\frac{\pi(1-\alpha_i)}{2}\right)+(\omega\tau_i)^{2\alpha_i}\right)^{\beta_i/2}}\right] + \frac{\sigma_0}{\varepsilon_0}\omega^{-s}\cos\left(\frac{\pi s}{2}\right) \qquad (22)$$

et

$$\varepsilon''(\omega) = \sum_{i=1}^{n}\left[\Delta\varepsilon_i \frac{\sin(\beta_i\varphi_i)}{\left(1+2(\omega\tau_i)^{\alpha_i}\sin\left(\frac{\pi(1-\alpha_i)}{2}\right)+(\omega\tau_i)^{2\alpha_i}\right)^{\beta_i/2}}\right] + \frac{\sigma_0}{\varepsilon_0}\omega^{-s}\sin\left(\frac{\pi s}{2}\right) \qquad (23)$$

avec :

$$\varphi_i = \arctan\left[\frac{(\omega\tau_i)^{\alpha_i}\cos\left(\frac{\pi(1-\alpha_i)}{2}\right)}{1+(\omega\tau_i)^{\alpha_i}\sin\left(\frac{\pi(1-\alpha_i)}{2}\right)}\right] \qquad (24)$$

Dans la suite de ce chapitre nous allons considérer uniquement le modèle d'Havriliak-Negami.

2.2. Estimation dans un contexte à erreurs bornées

Considérons le modèle d'Havriliak-Negami donné par l'expression (20) ; le but est d'estimer les paramètres τ_i, α_i, β_i, $\Delta\varepsilon_i$, ε_∞, σ_0 et s à l'aide des mesures de la permittivité complexe ε obtenues à différentes valeurs de la pulsation du champ électrique appliqué au matériau.

Supposons en premier lieu que l'on est en présence d'un seul mode de relaxation et que la contribution de la conductivité est négligeable. Si en plus, les données expérimentales ne sont pas bruitées, alors le spectre de relaxation est semblable à ceux présentés sur la figure 4. Dans ce dernier cas, les paramètres du modèle d'Havriliak-Negami peuvent même être déterminés géométriquement ; néanmoins, ce cas ne correspond pas à la réalité étant donné que les données mesurées sont généralement bruitées. On peut donc déduire que cette méthode géométrique pourrait produire des résultats éloignés de la réalité même lorsque l'on est en présence d'un seul mode de relaxation.

En général, le nombre de modes de relaxations est supérieur à 1 et dans certains cas ces modes peuvent se recouvrir. Sur la figure 5 est présenté dans un diagramme de Cole-Cole, le spectre de relaxation diélectrique d'un échantillon de Polyamide 11 hydraté, obtenu à la température de 0°C, pour une gamme de fréquence comprise entre 10^{-1} Hz et 10^6 Hz. Ce spectre peut être décomposé en trois modes de relaxation, dénommés successivement α, β_1 et

β_2 (à ne pas confondre avec les paramètres α_i, β_i) pour une échelle de fréquence croissante [Ibo00]. Les mesures ont été obtenues expérimentalement à l'aide d'un spectromètre diélectrique (modèle Novocontrol BDS-4000) [Ibo00].

L'analyse présentée sur la figure 5 a été réalisée en utilisant le modèle d'Havriliak-Negami avec seulement deux modes de relaxation (correspondant aux modes β_1 et β_2). En effet, il n'est pas possible en utilisant une méthode d'estimation de paramètres basée sur une minimisation de critère, par exemple quadratique, d'obtenir une estimation correcte des paramètres relatifs au mode α car celui-ci n'est pas « entièrement » observé dans la gamme de fréquence utilisée. En fait, on n'observe que la partie haute fréquence de ce mode. D'autre part, on remarque un recouvrement important des deux modes β_1 et β_2. L'estimation des paramètres par une méthode classique (le livre [WP94] représente un excellent support pour les personnes désirant s'intéressant aux méthodes classiques d'identification) est dans ce cas également assez difficile et la solution retournée dépend de l'initialisation des paramètres.

Cet exemple illustre assez bien la complexité de l'estimation des paramètres du modèle d'Havriliak-Negami d'une part lorsque plusieurs modes se recouvrent mutuellement et d'autre part lorsque seule la partie haute fréquence ou basse fréquence d'un mode est observée.

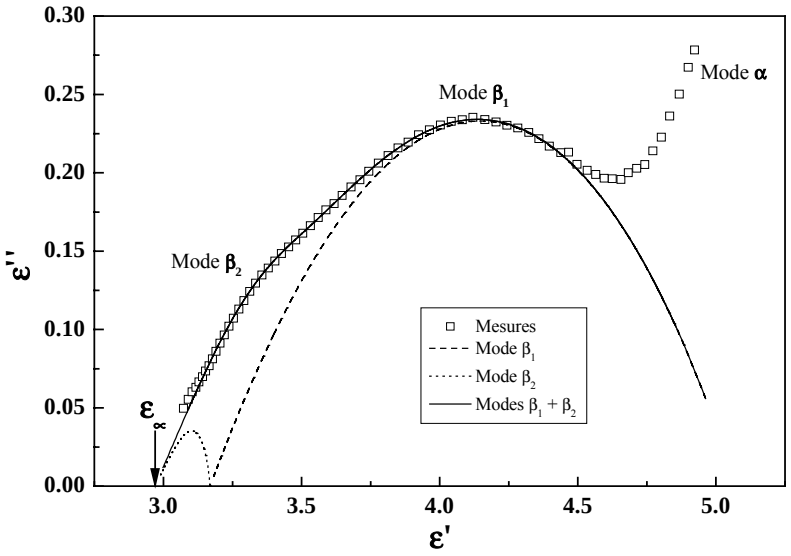

Figure 5 : Diagramme de Cole-Cole du spectre de relaxation diélectrique à 0°C du Polyamide 11 hydraté : mesures et analyse à partir du modèle d'Havriliak-Negami en utilisant deux modes de relaxation [Ibo00]

Dans la suite de ce chapitre, nous allons estimer les paramètres du modèle d'Havriliak-Negami donné par l'expression (20) dans un contexte ensembliste ; l'algorithme SIVIAP présenté dans le chapitre 1, c'est-à-dire l'algorithme SIVIA [JW93a] associé au contracteur p*ropagation-rétropropagation* [JKDW01] [Bra02] est alors utilisé. On suppose donc que la permittivité diélectrique complexe mesurée $\varepsilon(\omega_i)$ à différentes fréquences du champ électrique appliqué appartient à des intervalles connus *a priori* :

$$\varepsilon(\omega_i) = \left[\varepsilon_i^-, \varepsilon_i^+ \right] \quad i = 1 \ldots N \tag{25}$$

Notre but est alors de trouver toutes les valeurs des paramètres du modèle d'Havriliak-Negami telles que la sortie du modèle reste compatible avec les bornes expérimentales définies par (25). Néanmoins, le nombre de modes de relaxations n'étant pas connu *a priori*, la méthodologie suivie dans ce chapitre consiste donc à identifier ces paramètres en augmentant progressivement le nombre de modes de relaxation jusqu'à trouver le nombre optimal.

Pour estimer les paramètres du modèle d'Havriliak-Negami défini par l'expression (20) à l'aide de SIVIAP, il faut d'abord choisir une fonction d'inclusion de la permittivité diélectrique complexe. Deux approches sont envisagées ; la première consiste à décomposer la permittivité diélectrique complexe en parties réelle et imaginaire. La deuxième consiste à représenter la permittivité par un intervalle complexe.

Nous utiliserons dans cette partie, la première approche [RIRC03a] [RIRC03b] [RIRC04].

2.3. Validation de modèle – choix du nombre de relaxations

Dans ce paragraphe, nous allons montrer que les méthodes d'estimation de paramètres dans un contexte à erreurs bornées permettent de trouver le nombre de modes de relaxations optimal.

Premier cas d'étude : un seul mode.

Dans un premier lieu, nous avons simulé le modèle d'Havriliak-Negami pour un seul mode de relaxation en utilisant l'expression (19) avec :

$$(\alpha_1, \beta_1, \tau_1, \Delta\varepsilon_1, \varepsilon_\infty)^T = (1, 1, 0.00159, 6, 3)^T$$

Les domaines *a priori* des pseudo-mesures sont alors données par :

$$\left[\varepsilon_j^{'}\right]=\left[0.99\cdot\hat{\varepsilon}_j^{'},1.01\cdot\hat{\varepsilon}_j^{'}\right] \text{ et } \left[\varepsilon_j^{"}\right]=\left[0.99\cdot\hat{\varepsilon}_j^{"},1.01\cdot\hat{\varepsilon}_j^{"}\right]$$

$\varepsilon_j^{'}$ et $\varepsilon_j^{"}$ étant respectivement la partie réelle et la partie imaginaire de la permittivité diélectrique et $\hat{\varepsilon}_j^{'}$ et $\hat{\varepsilon}_j^{"}$ sont les pseudo-mesures ponctuelles obtenues par simulation du modèle. Les domaines initiaux de recherche des paramètres à estimer sont :

$$[\alpha_1] = [0.1,1],\ [\beta_1] = [0.1,1],\ [\tau_1] = [1.52\cdot10^{-8},\ 6.56\cdot10^{7}]\,s,\ [\Delta\varepsilon_1] = [0.1,1000],$$

$$[\varepsilon_\infty] = [1,\ 1000].$$

Pour trouver le nombre optimal de relaxations, on commence par estimer les paramètres du modèle d'Havriliak-Negami avec un seul mode. En utilisant SIVIAP, on génère, en 21.9 secondes sur un Celeron 1 GHz et avec une précision η = 0.01, un ensemble de pavés contenant toutes les solutions compatibles avec le modèle à un seul mode de relaxation et avec les hypothèses sur le bruit de mesure. La projection de cet ensemble par rapport aux différents paramètres donne une approximation extérieure de toutes les solutions ; pour chaque paramètre on trouve un intervalle qui contient sa valeur exacte.

$$\begin{pmatrix} \alpha_1 \\ \beta_1 \\ \tau_1 \\ \Delta\varepsilon_1 \\ \varepsilon_\infty \end{pmatrix} \in \begin{pmatrix} [0.99702,1] \\ [0.99682,1] \\ [0.00157,\ 0.001616] \\ [5.9251,6.0601] \\ [2.97,3.03] \end{pmatrix}$$

Etant donné que SIVIAP ne trouve pas d'approximations extérieures vides pour les paramètres à estimer, le modèle avec un seul mode de relaxation n'est alors pas rejeté. Néanmoins, notons que nous n'avons pas obtenu d'approximation intérieure. Etant donné la difficulté du problème, des méthodes d'optimisation ponctuelle doivent être utilisées afin d'obtenir cette approximation intérieure.

Supposons maintenant qu'on est en présence de deux modes de relaxations. En ajoutant une contrainte imposant que les deux temps de relaxations sont différents, SIVIAP génère alors en quelques minutes un ensemble vide ; ceci signifie que les données utilisées ne sont pas compatibles avec le modèle à deux modes de relaxations, ce dernier est alors rejeté.

Deuxième cas d'étude : deux modes de relaxation

Dans ce deuxième cas, nous avons simulé le modèle d'Havriliak-Negami avec deux modes de relaxation avec :

$\alpha_1 = 0.6,\ \beta_1 = 1,\ \tau_1 = 15.915 \cdot 10^{-6}$ s, $\Delta\varepsilon_1 = 1,\ \alpha_2 = 0.8,\ \beta_2 = 0.7,\ \tau_1 = 0.15915$ s, $\Delta\varepsilon_2 = 6$ et $\varepsilon_\infty = 3.$

Les domaines *a priori* des pseudo-mesures sont données par :

$$\left[\varepsilon_j^{'}\right] = \left[0.99 \cdot \hat{\varepsilon}_j^{'}, 1.01 \cdot \hat{\varepsilon}_j^{'}\right] \text{ et } \left[\varepsilon_j^{''}\right] = \left[0.99 \cdot \hat{\varepsilon}_j^{''}, 1.01 \cdot \hat{\varepsilon}_j^{''}\right]$$

D'abord, on suppose qu'on est en présence d'un seul mode de relaxation, SIVIAP trouve un ensemble vide ; ceci montre que les données ne sont pas compatibles avec le modèle à un seul mode de relaxation.

Paramètres	Valeurs exactes	Estimées
α_1	0.6	[0.593 ; 0.633]
β_1	1	[0.9375 ; 1]
$\Delta\varepsilon_1$	1	[0.971 ; 1.015]
τ_1 ($\times 10^{-6}$)	15.915	[15.04 ; 17.39]
α_2	0.8	[0.796 ; 0.803]
β_2	0.7	[0.683 ; 0.713]
$\Delta\varepsilon_2$	6	[5.984 ; 6.029]
τ_2 (s)	0.15915	[0.154 ; 0.166]
ε_∞	3	[2.999 ; 3.001]

Tableau 3 : Projection par rapport aux paramètres à estimer de l'ensemble des pavés générés par SIVIAP

Avec deux modes de relaxation, SIVIAP génère, en 4h 8 min, un ensemble de pavés dont la projection par rapport à chacun des paramètres estimés donne une approximation extérieure de la valeur réelle de ce dernier (voir Tableau 3).

Conclusion

Dans ces deux exemples, nous avons constaté que l'estimation de paramètres dans un contexte à erreurs bornées permet de rejeter un modèle lorsque des données expérimentales ne sont pas compatibles avec ce modèle. Ces méthodes nous permettent de sélectionner le nombre optimal des modes de relaxation inconnu *a priori*.

2.4. Influence de la stratégie de bissection

Dans cette partie, nous étudions l'influence de la stratégie de bissection sur les performances de l'algorithme SIVIAP. On considère alors le modèle d'Havriliak-Negami avec un seul mode de relaxation. Dans cette section, nous allons considérer deux cas différents (deux valeurs du vecteur de paramètres) pour lesquels, les pseudo-mesures sont obtenues par simulation du modèle et ajout d'un bruit numérique uniforme borné d'amplitude égale à 1% de la valeur simulée. Dans les deux cas, nous avons retenu 40 pseudo-mesures réparties sur tout le spectre (fréquences entre 10^{-2}Hz et 10^{6} Hz).

L'estimation des paramètres τ_1, α_1, β_1, $\Delta\varepsilon_1$, ε_∞ est effectuée en utilisant SIVIAP avec les stratégies de bissection **A** et **C** présentées dans le chapitre 1. On doit noter que la stratégie **C** requiert la différentiation des fonctions (22) et (23) ; étant donné que la forme analytique de la dérivée est assez complexe et les paramètres multi-occurrents, nous avons opté pour l'utilisation de la différentiation automatique [Ral81] [RC96], introduite dans le chapitre 1.

Dans les deux cas, la borne d'erreur *a priori* est prise égale à 1% de la valeur mesurée.

Premier cas : Les vraies valeurs des paramètres à estimer sont :

$$\tau_1 = 0.001591\,\text{s},\ \alpha_1 = 1,\ \beta_1 = 1,\ \Delta\varepsilon_1 = 6,\ \varepsilon_\infty = 3$$

Les résultats de l'estimation des paramètres du modèle d'Havriliak-Negami avec un seul mode de relaxation en utilisant les stratégies **A** et **C** sont donnés dans le tableau 4 [RRIC04].

	Stratégie A	**Stratégie C**
$\tau_1\times10^{-3}$	[1.572,1.616]	[1.543, 1.658]
α_1	[0.9968, 1]	[0.995, 1]
β_1	[0.996, 1]	[0.994, 1]
$\Delta\varepsilon_1$	[5.925, 6.06]	[5.88, 6.12]
ε_∞	[2.97, 3.03]	[2.97, 3.03]
tc (s)	11.7	0.23

Tableau 4 : Paramètres estimés en utilisant les stratégies de bissection **A** et **C**

Deuxième cas : Les vraies valeurs des paramètres à estimer sont :

$$\tau_1 = 0.001591\,\text{s},\ \alpha_1 = 0.8,\ \beta_1 = 0.5,\ \Delta\varepsilon_1 = 6,\ \varepsilon_\infty = 3$$

Les résultats de l'estimation des paramètres du modèle d'Havriliak-Negami avec un seul mode de relaxation en utilisant les stratégies **A** et **C** sont donnés dans le tableau 5.

	Stratégie A	**Stratégie C**
τ_1(ms)	[1.483,1.717]	[1.346,1.872]
α_1	[0.794,0.806]	[0.786,0.812]
β_1	[0.485,0.515]	[0.47,0.532]
$\Delta\varepsilon_1$	[5.88,6.12]	[5.857,6.15]
ε_∞	[2.95,3.04]	[2.93,3.06]
tc (s)	50.1	11.65

Tableau 5 : Paramètres estimés en utilisant les stratégies de bissection **A** et **C**

Conclusion :

On remarque que les tailles des intervalles générés par SIVIAP sont plus grandes dans le deuxième cas que dans le premier (indépendamment de la stratégie). Ceci peut être expliqué par le fait que peu de pavés sont éliminés lorsque la sortie du modèle tend à être « plate ». Dans notre cas, les parties réelle et imaginaire de la permittivité diélectrique sont assez raides lorsque le paramètre α_1 tend vers zéro. Dans ce dernier cas, une faible incertitude sur les mesures induit une grande incertitude sur les paramètres estimés. Nous verrons que ce problème sera rencontré par la suite dans tous les exemples où le paramètre α_1 s'approchera de zéro.

Le temps de calcul est nettement amélioré en utilisant la stratégie de bissection **C**, qui privilégie les directions pour lesquelles la taille du pavé représentant la sortie du modèle est la plus grande. Ainsi, il est parfois inutile de bissecter un paramètre non influent. Cet exemple montre qu'un choix judicieux de la stratégie de bissection peut, dans certains cas, réduire de manière significative le temps de calcul de l'algorithme SIVIAP.

Néanmoins, les tailles des pavés définissant l'approximation extérieure sont plus petites lorsqu'on utilise un processus de bissection uniforme (stratégie **A**). Cette dernière stratégie bissecte uniformément tous les paramètres et la convergence des tailles des intervalles générés par les bissections est équivalente pour tous les paramètres (on ne privilégie la bissection d'aucun paramètre). En particulier, les tailles des intervalles tendent vers zéro lorsque η tend vers zéro. D'un autre coté, il est probable qu'une direction de bissection reste souvent privilégiée en utilisant la stratégie **C** et dans ce cas les autres paramètres ne seront pas suffisamment bissectés.

2.5. Etude de cas

Dans cette partie, nous évaluons les performances expérimentales de SIVIAP en fonction du nombre de modes de relaxation, de la vraie valeur des paramètres, de la gamme de fréquences, de la taille de l'échantillon et de l'amplitude des erreurs additives.

2.5.1. Cas d'un seul mode de relaxation

On considère le modèle d'Havriliak-Negami avec un seul mode de relaxation dont les parties réelle et imaginaire de la permittivité diélectrique sont données en fonction des paramètres ε_∞, $\Delta\varepsilon_1$, τ_1, α_1, β_1. Le but est d'estimer ces paramètres en mesurant la permittivité diélectrique complexe. Dans cette section, les pseudo-mesures sont obtenues en simulant le modèle avec plusieurs valeurs du vecteur des paramètres. Les domaines *a priori* des pseudo-mesures sont données par :

$$\left[\varepsilon_j^{'}\right]=\left[0.99\cdot\hat{\varepsilon}_j^{'},1.01\cdot\hat{\varepsilon}_j^{'}\right] \text{ et } \left[\varepsilon_j^{"}\right]=\left[0.99\cdot\hat{\varepsilon}_j^{"},1.01\cdot\hat{\varepsilon}_j^{"}\right]$$

Dans la suite de cette partie nous allons utiliser SIVIAP dans certains cas où le spectre n'est pas complet. En particulier, nous allons étudier l'influence de la largeur du spectre et du nombre de mesures.

2.5.1.1. Effet de l'échantillonnage

Considérons le spectre tracé sur la figure 6 et supposons qu'on ne dispose pas de toutes les mesures. Néanmoins, le domaine de fréquences est le même que précédemment $f \in [10^{-2}, 10^{6}]$ Hz. Les résultats obtenus pour 54, 27 et 20 mesures sont donnés dans le tableau 6 avec les temps de calcul t_c correspondants.

	$\tau_1\times10^{-4}$ s	α_1	β_1	$\Delta\varepsilon_1$	ε_∞	t_c (s)
54 mesures	[15.72, 16.16]	[0.9969,1]	[0.99671,1]	[5.9251,6.0602]	[2.97,3.03]	15.4
27 mesures	[15.72, 1616]	[0.99681,1]	[0.99668,1]	[5.9252,6.0603]	[2.97,3.03]	9.7
20 mesures	[15.72, 16.17]	[0.99674,1]	[0.99649,1]	[5.918,6.0613]	[2.97,3.03]	5.7

Tableau 6. Paramètres diélectriques estimés pour différents nombres d'échantillons

Dans les trois cas, l'incertitude sur les paramètres estimés est comparable et petite ; l'algorithme est efficace même avec un nombre réduit de mesures. Ceci peut s'expliquer par la présence de nombreuses données redondantes d'où un temps de calcul plus important dans le premier cas.

2.5.1.2. Effet du domaine de fréquence

Dans certains cas, il n'est pas possible d'explorer expérimentalement un domaine de fréquence aussi large que celui défini précédemment. D'autre part, la valeur du temps de relaxation étant dépendante de la température, les modes de relaxations sont le plus souvent excentrés par rapport au domaine de fréquence utilisé et seule la partie basse fréquence ou haute fréquence est généralement observée. Dans de tels cas, les approches classiques basées sur la minimisation de critères échouent notamment à cause de la difficulté de l'initialisation.

Dans ce paragraphe nous étudierons plusieurs cas pour des mesures effectuées sur seulement une partie de la bande de fréquence. Ces domaines sont présentés sur la figure 6. Plusieurs séries d'estimation ont été réalisées en couvrant différentes gammes de fréquences.

Les résultats des différentes estimations, y compris les temps de calcul, réalisées en utilisant SIVIAP sont présentés dans le tableau 7.

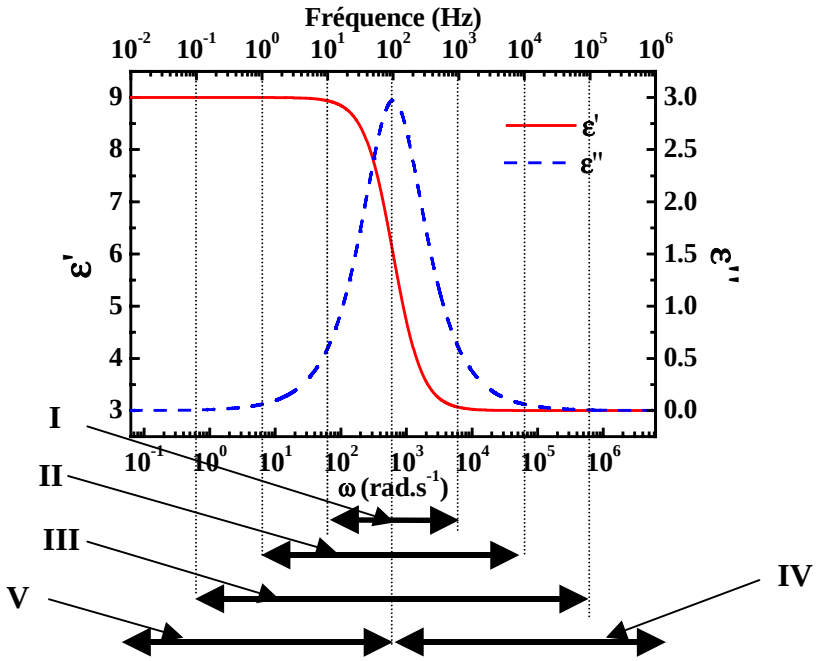

Figure 6. Différents domaines de fréquence étudiés pour le modèle d'Havriliak-Negami comportant une relaxation avec $\alpha_1 = \beta_1 = 1$.

Cas	$\tau_1 \times 10^{-4}$ s	α_1	β_1	$\Delta\varepsilon_1$	ε_∞	t_c (s)
I	[15.68, 16.51]	[0.98858,1]	[0.97579,1]	[5.9218,6.0768]	[2.9375,3.0348]	755
II	[15.71, 16.25]	[0.99357,1]	[0.99217,1]	[5.9241,6.0648]	[2.9686,3.03]	182.4
III	[15.73, 16.19]	[0.99579,1]	[0.99531,1]	[5.9223,6.0612]	[2.9699,3.03]	53.2
IV	[15.53, 16.88]	[0.99573,1]	[0.99573,1]	[5.9137,6.186]	[2.97,3.03]	494.1
V	[15.07, 21.57]	[0.99437,1]	[0.57636,1]	[5.9007,7.9653]	[1,3.1891]	749.3

Tableau 7. Paramètres diélectriques estimés pour plusieurs bandes de fréquence en utilisant toutes les mesures

L'incertitude sur les paramètres estimés est relativement petite pour les différents cas étudiés sauf pour le dernier cas d'étude pour lequel on ne dispose pas d'informations sur la partie haute fréquence du mode de relaxation. Dans ce dernier cas, l'incertitude sur β_1 est importante ; en effet β_1 n'influence la forme du spectre de relaxation qu'en haute fréquence. On remarque ainsi que cette incertitude se propage sur les autres paramètres à l'exception de α_1. Le temps de calcul est relativement faible lorsque les mesures couvrent tout le spectre.

2.5.1.3. Effet de la dissymétrie de la distribution des temps de relaxation

Pour étudier l'effet de la dissymétrie du spectre de relaxation, on considère plusieurs cas en faisant varier β_1 (voir figure 7), pour $\alpha_1 = 1$, ce qui correspond au modèle de Cole-Davidson [CA93]. Les différents spectres de relaxation simulés et analysés sont présentés sur la figure 7 ; les valeurs des autres paramètres utilisés sont les mêmes que précédemment ($\varepsilon_\infty = 3$, $\Delta\varepsilon_1 = 6$ et $\tau_1 = 0.001591$s).

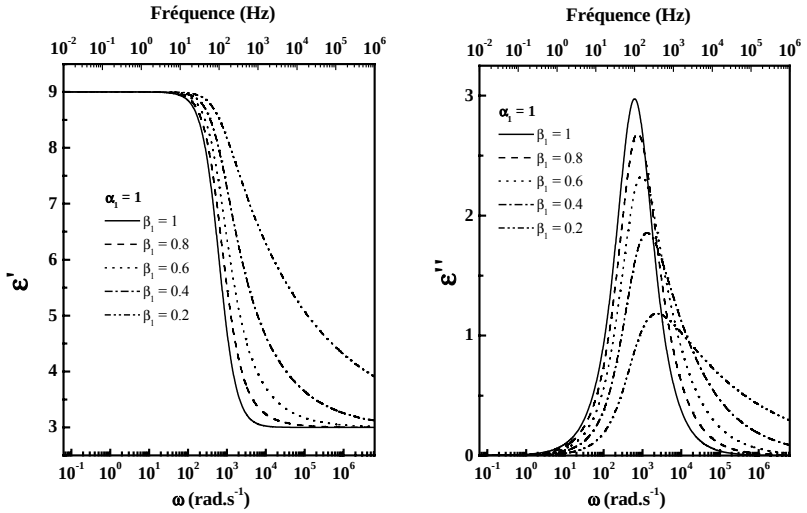

Figure 7. Evolution des spectres de relaxation pour $\alpha_1 = 1$ et différentes valeurs de β_1 (modèle de Cole-Davidson)

$\tau_1 \times 10^{-4}$ s	α_1	β_1	$\Delta\varepsilon_1$	ε_∞	t_c (s)
[15.41, 16.2]	[0.9956,1]	[0.79637,0.80803]	[5.91,6.0753]	[2.9699,3.0301]	33.7
[15.35, 16.26]	[0.99545,1]	[0.59607,0.60758]	[5.9101,6.0738]	[2.9691,3.0311]	35.4
[15.25, 16.33]	[0.99516,1]	[0.39611,0.40665]	[5.9181,6.0669]	[2.9637,3.0376]	37.2
[15.26, 16.39]	[0.99511,1]	[0.19633,0.20548]	[5.9184,6.0665]	[2.9242,3.0757]	54.8

Tableau 8. Paramètres diélectriques estimés pour $\alpha_1 = 1$ et plusieurs valeurs de β_1 ($\beta_1 = 0.8, 0.6, 0.4, 0.2$)

Les résultats de l'estimation des paramètres τ_1, α_1, β_1, $\Delta\varepsilon_1$, ε_∞ pour différentes valeurs de β_1 sont donnés par le tableau 8. Dans les différents cas étudiés, l'incertitude sur les paramètres estimés est petite. D'autre part, les temps de calcul obtenus restent faibles quelle que soit la valeur de β_1.

2.5.1.4. Effet de la largeur de la distribution des temps de relaxation

Pour étudier l'effet de la largeur de la distribution des temps de relaxation des spectres, on considère plusieurs cas en faisant varier α, pour $\beta_1 = 1$, ce qui correspond au modèle de Cole-Cole [CA93], les valeurs des autres paramètres étant les mêmes que précédemment. Les

spectres étudiés sont présentés sur la figure 8 et les résultats des estimations sont donnés dans le tableau 9.

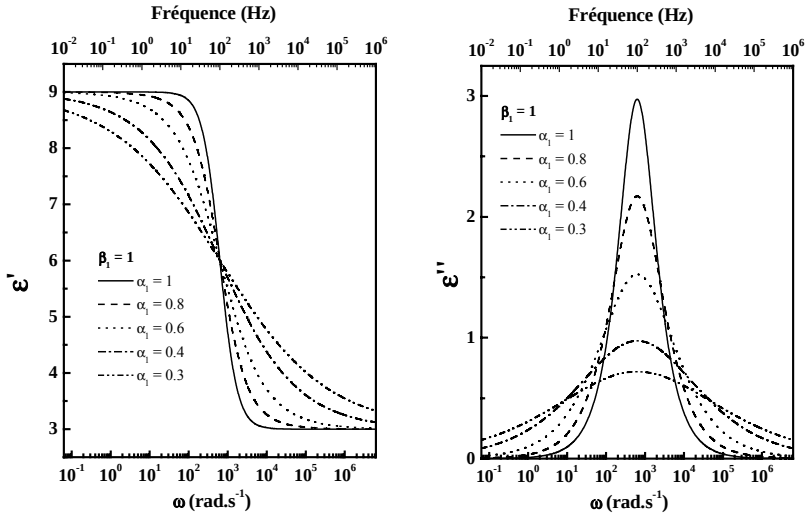

Figure 8. Paramètres diélectriques estimés pour $\beta_1 = 1$ et plusieurs valeurs de α_1.

$\tau_1 \times 10^{-4}$ s	α_1	β_1	$\Delta\varepsilon_1$	ε_∞	t_c (s)
[15.69, 16.59]	[0.79653,0.80496]	[0.98698,1]	[5.9104,6.0987]	[2.9698,3.0301]	53.3
[15.57, 17.33]	[0.59517,0.60596]	[0.9756,1]	[5.8953,6.124]	[2.9676,3.0316]	107.4
[14.97, 21.7]	[0.39373,0.40962]	[0.91988,1]	[5.8828,6.1357]	[2.9459,3.0437]	405.3
[14.26, 35.36]	[0.29148,0.31966]	[0.83244,1]	[5.87,6.146]	[2.8808,3.0656]	1659.1

Tableau 9. Paramètres diélectriques estimés pour $\beta_1 = 1$ et différentes valeurs de α_1 (α_1 = 0.8, 0.6, 0.4, 0.3)

D'après les résultats de l'estimation des paramètres τ_1, α_1, β_1, $\Delta\varepsilon_1$, ε_∞ présentés dans le tableau 9, on remarque que l'incertitude sur les paramètres estimés augmente avec la largeur de la distribution des temps de relaxation. Néanmoins, l'incertitude sur α_1 est relativement petite, ceci s'explique par le fait que α_1 est bien estimé lorsque les mesures couvrent tout le spectre de relaxation. Les temps de calcul obtenus, dépendent sensiblement de la valeur de α_1. Pour des valeurs de $\alpha_1 < 0.3$, le temps de calcul devient considérable ; ceci peut s'expliquer par le fait que pour des telles valeurs de α_1, les parties réelle et imaginaire de la permittivité complexe sont données par des fonctions qui tendent à être plates. Néanmoins, de telles valeurs de α_1 sont rarement rencontrées dans des cas réels.

2.5.1.5. Effet combiné de la dissymétrie et de la largeur de la distribution des temps de relaxation

Dans cette section nous étudions simultanément l'effet de la dissymétrie du spectre de relaxation ainsi que de sa largeur. Dans ce cas, quelques couples de valeurs de α_1 et β_1 ont été étudiés. Les spectres correspondants sont donnés sur la figure 9.

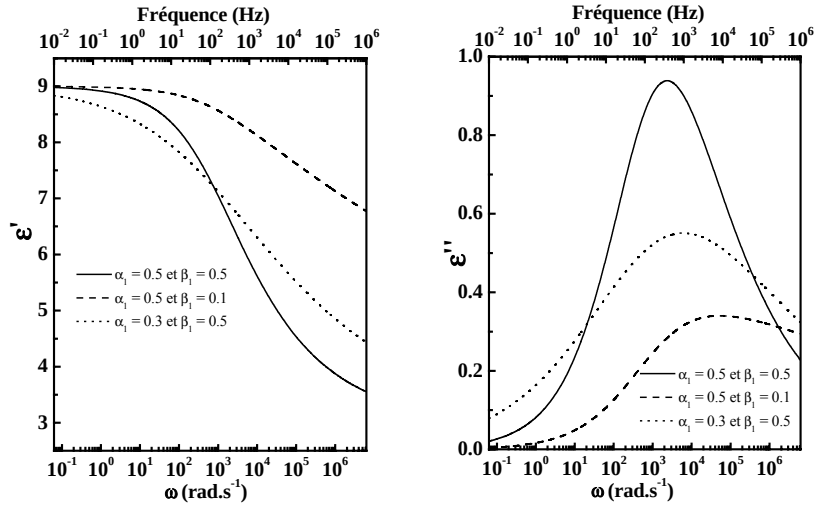

Figure 9. Spectres de relaxation pour différentes valeurs de β_1 et α_1.

D'après le tableau 10 on remarque que pour les trois cas étudiés l'incertitude sur le paramètre α_1 est relativement petite ; comme on l'a indiqué dans les paragraphes précédents, ce paramètre est toujours bien estimé lorsque des mesures couvrent la partie basse fréquence du mode de relaxation. Les incertitudes sur les paramètres β_1 et ε_∞ sont assez grandes pour les deux derniers cas, ceci s'explique par le fait que les mesures ne couvrent pas entièrement la partie haute fréquence du mode de relaxation. Ceci entraîne une incertitude très importante sur le temps de relaxation, notamment dans le dernier cas d'étude, ainsi qu'un accroissement du temps de calcul.

$\tau_1 \times 10^{-4}$ s	α_1	β_1	$\Delta\varepsilon_1$	ε_∞	t_c (s)
[12.43, 20.46]	[0.4907,0.51]	[0.45229,0.54588]	[5.8226,6.1787]	[2.863,3.118]	609
[10.74, 23.46]	[0.485,0.5215]	[0.06624,0.14494]	[4.7152,8.0619]	[1,4.321]	6480
[5.58, 107.66]	[0.2853,0.359]	[0.29723,0.66859]	[5.6476,6.8038]	[2.144,3.389]	7918

Tableau 10. Paramètres diélectriques estimés pour plusieurs valeurs de α_1 et β_1 ($\alpha_1 = \beta_1 = 0.5$, $\alpha_1 = 0.5$ et $\beta_1 = 0.1$, $\alpha_1 = 0.3$ et $\beta_1 = 0.5$)

2.5.1.6. Influence du bruit de mesure

Pour observer l'effet de l'amplitude de la borne de l'erreur de sortie, nous avons considéré le spectre de relaxation obtenu en simulant le modèle d'Havriliak-Negami comprenant un seul mode de relaxation avec les valeurs suivantes :

$$\tau_1 = 1/200\pi \simeq 0.001591\,\text{s},\ \alpha_1 = 0.8,\ \beta_1 = 0.5,\ \Delta\varepsilon_1 = 6,\ \varepsilon_\infty = 3$$

Nous donnons dans le tableau 11 les résultats de l'estimation des paramètres diélectriques en fonction de la largeur des domaines des mesures.

Bruit	1%	2%	5%
α_1	[0.793, 0.807]	[0.790, 0.810]	[0.781, 0.820]
β_1	[0.487, 0.513]	[0.480, 0.519]	[0.461, 0.538]
τ_1 (ms)	[1.490, 1.694]	[1.427, 1.783]	[1.277, 2.010]
$\Delta\varepsilon_1$	[5.890, 6.115]	[5.816, 6.189]	[5.628, 6.389]
ε_∞	[2.963, 3.035]	[2.931, 3.068]	[2.832, 3.164]
t_c (s)	108.43	453.84	4920

Tableau 11 : Paramètres diélectriques estimés avec amplitude de la borne de l'erreur de sortie : 1%, 2% et 5%

Nous remarquons d'après le tableau 11 que l'incertitude sur les paramètres estimés augmente en fonction de l'amplitude sur le bruit de mesure. Ce pessimisme est plus remarquable pour les paramètres β_1, τ_1 et $\Delta\varepsilon_1$. Néanmoins, les paramètres α_1 et ε_∞ sont relativement bien estimés dans les trois cas. L'accroissement du temps de calcul avec l'erreur peut s'expliquer par l'absence de solutions intérieures même dans le cas d'une erreur de 5% ; le nombre de bissections est considérable. On rappelle qu'on décide d'arrêter un pavé lorsqu'on prouve qu'il est *acceptable* ou le cas échéant inconsistant avec les mesures.

2.5.1.7. Conclusions

Dans cette partie, nous avons considéré le modèle d'Havriliak-Negami avec un seul mode de relaxation. Nous avons constaté que l'incertitude sur les paramètres estimés dépend surtout de la valeur du paramètre α_1. En effet, le mode de relaxation est mieux visible pour des valeurs de α_1 proches de 1. D'autre part, le temps du calcul ainsi que l'incertitude sur les paramètres estimées augmentent lorsque une partie du spectre n'est pas disponible. Ceci s'explique par le fait que d'une part, l'absence d'une partie du spectre rend la contraction non optimale et d'autre part, on ne dispose pas d'informations sur le comportement du modèle dans cette zone manquante.

D'un un autre coté, nous avons étudié l'effet de l'amplitude de la borne d'erreur des mesures sur les paramètres estimés. Nous avons alors constaté que la variation de cette amplitude ne se répercute pas sur tous les paramètres mais seulement sur β_1, τ_1 et $\Delta\varepsilon_1$. En effet, ε_∞ est toujours bien estimé lorsqu'on dispose des mesures hautes fréquences et α_1 des mesures basses fréquences.

2.5.2. Cas de deux modes de relaxation

On considère maintenant le modèle d'Havriliak-Negami avec deux modes de relaxation dont les parties réelle et imaginaire de la permittivité diélectrique sont calculées à partir des paramètres ε_∞ , $\Delta\varepsilon_1$, τ_1 , α_1 , β_1, $\Delta\varepsilon_2$, τ_2 , α_2 , β_2 donnés dans le tableau 12. Le spectre de relaxation correspondant est présenté sur la figure 10. L'amplitude de l'erreur de sortie est de 1%.

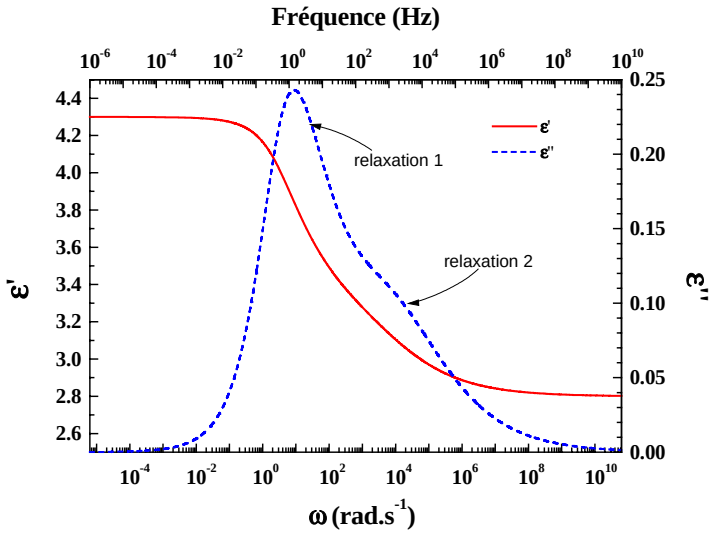

Figure 10. Spectre de relaxation obtenu pour les paramètres donnés dans le tableau 12

Paramètre	Valeur réelle	Encadrement extérieur
α_1	0.7	[0.6602; 0.74208]
β_1	0.5	[0.37108; 0.60837]
τ_1 (s)	0.3183	[0.2839; 0.3569]
$\Delta\varepsilon_1$	1	[0.9174; 1.25]
α_2	0.4	[0.33619; 0.43822]
β_2	0.8	[0.6875; 1]
τ_2 (s)	1.989×10^{-4}	$[8.481 \times 10^{-5} ; 3.045 \times 10^{-4}]$
$\Delta\varepsilon_2$	0.5	[0.25106 ; 0.57618]
ε_∞	2.8	[2.7931 ; 2.8018]

Tableau 12. Paramètres diélectriques estimés pour deux modes de relaxation

Le résultat de l'estimation des paramètres du modèle d'Havriliak-Negami avec deux modes de relaxations en utilisant SIVIAP est donnée dans le tableau 12. On remarque que l'incertitude sur certains paramètres estimés est importante. Pour avoir une solution moins pessimiste (petites largeurs des intervalles), il faudrait travailler avec une meilleure précision (il faudrait utiliser un paramètre η dans SIVIAP suffisamment petit) ce qui va engendrer un temps de calcul plus important.

Par ailleurs, le temps de calcul est de l'ordre de 15 heures, ceci s'explique par le fait que le contracteur utilisé ne permet pas de réduire considérablement les domaines des paramètres à estimer. Nous verrons dans la suite de ce chapitre que l'utilisation de la représentation polaire pour les intervalles complexes nous permettra de réduire de manière significative le temps de calcul.

2.6. Inconvénients de la décomposition en parties réelle et imaginaire

Dans les exemples traités dans les sections précédentes, nous avons utilisé des expressions analytiques pour les parties réelle et imaginaire de la permittivité diélectrique complexe. Ceci nous a permis d'utiliser des techniques d'analyse par intervalles réels. Néanmoins, l'utilisation de formes explicites pour ε' et ε'' limite les performances du contracteur associé à SIVIAP étant donné la présence de paramètres multioccurrents. En effet, le nombre d'occurrences de la majorité des paramètre est supérieur à 1. On donne dans le tableau 13 le nombre d'occurrences de chaque variable dans la partie réelle et la partie imaginaire.

	τ_i	α_i	β_i	$\Delta\varepsilon_i$	ε_∞	σ_0	s
ε'	4	7	2	1	1	1	2
ε''	4	7	2	1	0	1	2

Tableau 13. Nombre d'occurrences des paramètres du modèle d'Havriliak-Negami

On remarque que le paramètre α_i se répète 7 fois dans la partie réelle et autant de fois dans la partie imaginaire. L'évaluation des fonctions d'inclusion naturelles de ε' et ε'' est alors très pessimiste à cause du phénomène de dépendance.

En examinant l'expression complexe de la permittivité diélectrique donnée par l'équation (20), on observe que si on manipulait directement des intervalles complexes, tous les paramètres du modèle seraient mono-occurrents. Pour exploiter cette forme mono-occurrente, nous utiliserons ci-après des intervalles complexes.

2.7. Intervalles complexes

Dans cette section, nous allons utiliser le modèle donné par la fonction à variables complexes définie par (20). Ceci nécessite donc le bon choix de la représentation des intervalles complexes. Pour choisir une telle représentation, nous commençons d'abord par évaluer le dénominateur d'un des termes représentant une relaxation en utilisant les formes rectangulaire et polaire.

On considère la fonction complexe f définie par :

$$f(\omega, \tau, \alpha, \beta) = \left(1+(j\omega\tau)^{\alpha}\right)^{\beta} \quad \text{où} \qquad j^2 = -1 \tag{26}$$

et soient : $[\alpha]=[0.5, 1]$, $[\beta]=[0.4, 1]$, $[\tau]=[0.5, 1]$ s et $\omega = 2\pi$ rad.s^{-1}.

Evaluons d'abord la fonction d'inclusion naturelle de f en utilisant la représentation rectangulaire des intervalles complexes. On note alors par :

$$\begin{cases} Z_1 = (j\omega\tau)^{\alpha} = (\omega\tau)^{\alpha} \cdot \left(\cos\left(\frac{\alpha\pi}{2}\right) + j \sin\left(\frac{\alpha\pi}{2}\right)\right) \\ Z_2 = (1 + Z_1) \\ Z_3 = (Z_2)^{\beta} \end{cases}$$

Pour évaluer la fonction d'inclusion naturelle de *f*, il suffit d'évaluer celles de Z_1, Z_2 et Z_3 données par :

$$\begin{cases} [Z_1] = (j\omega[\tau])^{[\alpha]} = (\omega[\tau])^{[\alpha]} \cdot \left(\cos\left(\frac{[\alpha]\pi}{2}\right) + j \sin\left(\frac{[\alpha]\pi}{2}\right)\right) \\ [Z_2] = (1 + [Z_1]) \\ [Z_3] = ([Z_2])^{[\beta]} \end{cases}$$

On remarque que l'évaluation de Z_1 est pessimiste. En effet l'ensemble :

$$\left\{ z = (\omega\tau)^{\alpha} \cdot \left(\cos\left(\frac{\alpha\pi}{2}\right) + j \sin\left(\frac{\alpha\pi}{2}\right)\right) \mid \alpha \in [\alpha], \beta \in [\beta], \tau \in [\tau] \right\}$$

n'est pas un rectangle, étant donné que l'ensemble :

$$\left\{ z = \left(\cos\left(\frac{\alpha\pi}{2}\right) + j \sin\left(\frac{\alpha\pi}{2}\right)\right) \mid \alpha \in [\alpha] \right\}$$

décrit un arc de cercle. Les deux sources de pessimisme que sont le phénomène de dépendance et d'enveloppement influent dans notre cas.

De même, l'évaluation de la fonction d'inclusion naturelle de Z_3 est pessimiste ; ceci est dû au fait qu'elle est évaluée comme suit :

$$[Z_3] = ([Z_2])^{[\beta]} = \exp([\beta] \cdot \mathrm{Log}([Z_2]))$$

et l'évaluation de la fonction logarithme sur un rectangle n'est pas un rectangle.

Considérons maintenant la représentation polaire des intervalles complexes. On obtient alors :

$$\begin{cases} Z_1 = (j\omega\tau)^{\alpha} = \left\{ (\omega\tau)^{\alpha}, \dfrac{\alpha\pi}{2} \right\} = \{\rho_1, \theta_1\} \\ Z_2 = (1 + Z_1) = \{\rho_2, \theta_2\} \\ Z_3 = (Z_2)^{\beta} = \left\{ (\rho_2)^{\beta}, \beta \cdot \theta_2 \right\} = \{\rho_3, \theta_3\} \end{cases}$$

Les fonctions d'inclusion naturelles de Z_1, Z_2 et Z_3 sont ainsi données par :

$$\begin{cases} [Z_1] = (j\omega[\tau])^{[\alpha]} = \left\{ (\omega[\tau])^{[\alpha]}, \dfrac{[\alpha]\pi}{2} \right\} = \{[\rho_1], [\theta_1]\} \\ [Z_2] = (1 + [Z_1]) = \{[\rho_2], [\theta_2]\} \\ [Z_3] = ([Z_2])^{[\beta]} = \left\{ ([\rho_2])^{[\beta]}, [\beta] \cdot [\theta_2] \right\} = \{[\rho_3], [\theta_3]\} \end{cases}$$

On remarque alors que l'évaluation de Z_1 par le biais de sa fonction d'inclusion naturelle est exacte étant donné que les fonctions exp et Log sont minimales et que l'on n'a pas d'effet de dépendance. Le même raisonnement nous permet de conclure que l'évaluation de Z_3 est exacte (à condition que celle de Z_2 le soit). Néanmoins, Z_2 ne peut pas être représenté par un secteur, du fait de la présence de l'opération (+) ; un pessimisme est alors introduit.

Numériquement, nous obtenons les évaluations suivantes :

$$[f_r](\omega, [\tau], [\alpha], [\beta]) = [0.189847, 8.27881] + j[0.109196, 8.20953]$$

$$\left[f_p\right](\omega, [\tau], [\alpha], [\beta]) = \left[[1.32873, 7.02597] \ , \ [0.20304, 1.41297]\right]$$

$$\equiv [0.208844, 6.88164] + j[0.267933, 6.93864]$$

où $[f_r]$ est la fonction d'inclusion naturelle de f évaluée en utilisant la représentation rectangulaire et $[f_p]$ est celle obtenue par la forme polaire. On remarque que la seconde est moins pessimiste.

Dans la suite, nous allons alors utiliser l'arithmétique des intervalles complexes polaires développée dans le chapitre 2 afin d'étudier quelques cas expérimentaux.

2.7.1. Etude de cas expérimentaux

Dans cette partie, nous poursuivons l'étude de la performance expérimentale de SIVIAP en traitant des cas de complexité accrue, ce qui est rendu possible par le caractère mono-occurent

des paramètres du modèle et donc la bonne performance attendue du contracteur *propagation – rétropropagation.*

2.7.1.1. Cas d'un mode de relaxation en présence de la conductivité

Dans cette section, nous allons étudier divers cas expérimentaux où la contribution de la conductivité n'est pas négligée ; le nombre de paramètres à estimer est alors fixé à 7. Nous verrons que le temps de calcul n'augmente pas sensiblement par rapport au cas de l'estimation de seulement 5 paramètres. Ceci s'explique par le fait que les paramètres liés à la contribution de la conductivité sont mono-occurrents ; la contraction de leurs domaines par le contracteur *propagation – rétropropagation* est optimale.

Position du mode de relaxation :

On considère le cas des spectres de relaxation tracés sur la figure 11, ils sont obtenus en simulant le modèle d'Havriliak-Negami avec un seul mode de relaxation et en considérant que la contribution de la conductivité n'est pas négligée. Sur la figure 11, nous avons tracé 5 spectres obtenus pour 5 valeurs de τ_1 (10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5} (s)) et en prenant :

$$\alpha_1 = 0.8 \; ; \; \beta_1 = 0.6 \; ; \; \Delta\varepsilon_1 = 6 \; ; \; \varepsilon_\infty = 3 \; ; \; s = 1 \; ; \; \sigma_0 / \varepsilon_0 = 11.3 \; \text{s}^{-1}$$

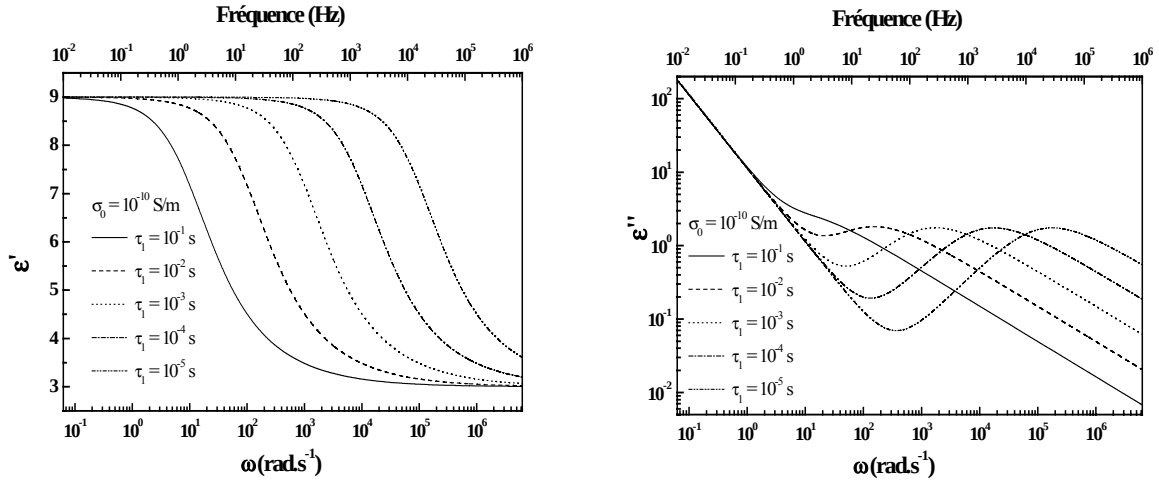

Figure 11. Spectres de relaxation pour différentes valeurs de τ_1.

Les résultats de l'estimation des paramètres du modèle d'Havriliak-Negami avec un seul mode de relaxation et tenant compte de la contribution de la conductivité sont donnés dans le tableau 14.

D'après les résultats donnés dans le tableau 14, nous constatons que le temps de calcul diminue lorsque le mode de relaxation et le phénomène de conduction sont séparés et lorsque le mode est centré sur le domaine de fréquence. D'autre part, nous remarquons que l'incertitude sur le paramètre β_1 augmente lorsque la partie haute fréquence du mode est tronquée. Enfin, les paramètres $\Delta\varepsilon_1$, ε_∞, σ_0 et s sont bien estimés dans tous les cas étudiés.

Paramètre	1er cas	2ème cas	3ème cas	4ème cas	5ème cas
α_1	[0.7795, 0.831]	[0.785, 0.8195]	[0.788, 0.81]	[0.79, 0.808]	[0.794, 0.806]
β_1	[0.569, 0.6244]	[0.577, 0.6027]	[0.5839, 0.6205]	[0.581, 0.6029]	[0.574, 0.625]
τ_1 (s)	[0.095, 0.106]	[$0.95.10^{-2}$, $1.04.10^{-2}$]	[$0.94.10^{-3}$, $1.04.10^{-3}$]	[$0.94.10^{-4}$, $1.05.10^{-4}$]	[$0.94.10^{-5}$, $1.07.10^{-5}$]
$\Delta\varepsilon_1$	[5.932, 6.038]	[5.97, 6.027]	[5.98, 6.02]	[5.97, 6.03]	[5.939, 6.06]
ε_∞	[2.995, 3.004]	[2.99, 3.007]	[2.988, 3.01]	[2.98, 3.02]	[2.948, 3.053]
σ_0/ε_0	[11.288, 11.32]	[11.288, 11.316]	[11.288, 11.315]	[11.288, 11.315]	[11.288, 11.3147]
s	[0.999, 1]	[0.999, 1]	[0.999, 1]	[0.999, 1]	[0.999, 1]
Tc (s)	19.6	9.54	6.76	7.8	20.25

Tableau 14. Paramètres diélectriques estimés pour plusieurs valeurs de τ_1 (10^{-1} s, 10^{-2} s, 10^{-3} s, 10^{-4} s, 10^{-5} s)

Influence de la contribution de la conductivité :

Considérons les spectres de relaxation tracés sur la figure 12 afin d'étudier l'influence de la contribution de la conductivité sur les paramètres estimés.

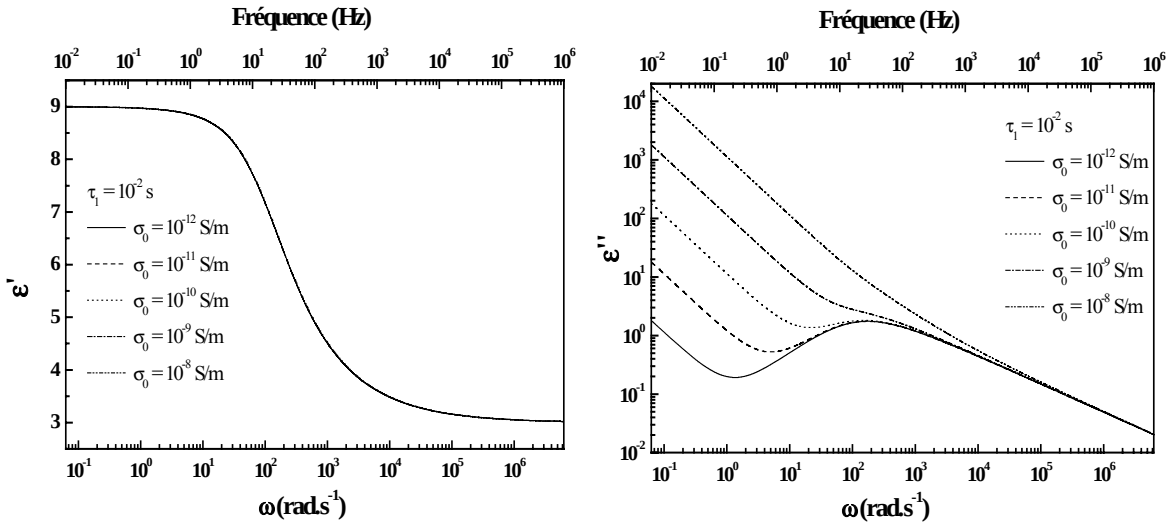

Figure 12. Spectres de relaxation pour différentes valeurs de σ_0.

Les spectres tracés sur la figure 12 sont obtenus en simulant le modèle d'Havriliak-Negami avec un mode de relaxation pour différentes valeurs de σ_0/ε_0 (s^{-1}) (σ_0/ε_0 = 0.113 , 1.13, 11.3, 113, 1130) et pour :

$$\tau_1 = 0.01 \text{ s} ; \alpha_1 = 0.8 ; \beta_1 = 0.6 ; \Delta\varepsilon_1 = 6 ; \varepsilon_\infty = 3 ; s = 1 ;$$

Les résultats de l'estimation des paramètres du modèle d'Havriliak-Negami avec un seul mode de relaxation et tenant compte de la contribution de la conductivité sont donnés dans le tableau 15.

Paramètre	1er cas	2ème cas	3ème cas	4ème cas	5ème cas
α_1	[0.791, 0.8097]	[0.786, 0.8125]	[0.785, 0.8195]	[0.7765, 0.8265]	[0.7619, 0.8399]
β_1	[0.5858, 0.613]	[0.582, 0.62]	[0.577, 0.6027]	[0.5708, 0.6282]	[0.5607, 0.6429]
τ_1 (s)	[0.00964, 0.0105]	[0.00963, 0.0106]	$[0.95.10^{-2}, 1.04.10^{-2}]$	[0.00938, 0.0105]	[0.00913, 0.0106]
$\Delta\varepsilon_1$	[5.98, 6.015]	[5.977, 6.019]	[5.97, 6.027]	[5.9596, 6.0349]	[5.957, 6.0358]
ε_∞	[2.994, 3.006]	[2.993, 3.0068]	[2.99, 3.007]	[2.9924, 3.0071]	[2.991, 3.007]
σ_0/ε_0	[0.1126, 0.1154]	[1.1283, 1.1362]	[11.288, 11.316]	[112.88, 113.11]	[1128.8, 1131.08]
s	[0.98977, 1]	[0.998, 1]	[0.999, 1]	[0.999, 1]	[0.99999, 1]
t_c (s)	1.78	2.2	9.54	110.5	1700

Tableau 15. Paramètres diélectriques estimés pour plusieurs valeurs de σ_0/ε_0

Nous remarquons d'après le tableau 15 que le temps de calcul augmente considérablement lorsque le mode de relaxation est entièrement masqué par le phénomène de conduction (5ème cas). Néanmoins, dans ce dernier cas les paramètres σ_0 et s sont très bien estimés ; en effet, on dispose de suffisamment d'informations relatives à ce phénomène. En revanche, les paramètres relatifs au mode de relaxation (α_1, β_1, τ_1, $\Delta\varepsilon_1$, ε_∞) sont dans ce cas moins bien estimés, l'incertitude est relativement grande (par rapport aux paramètres relatifs au phénomène de conduction).

2.7.1.2. Cas de deux modes de relaxation et comparaison des représentations

On considère le modèle d'Havriliak-Negami avec deux modes de relaxation dont les parties réelle et imaginaire de la permittivité diélectrique sont calculées à partir des paramètres ε_∞ , $\Delta\varepsilon_1$, τ_1 , α_1 , β_1, $\Delta\varepsilon_2$, τ_2 , α_2 , β_2 donnés dans le tableau 16. Dans le même tableau, nous avons donné le résultat de l'estimation de ces paramètres en utilisant SIVIAP avec $\eta = 0.01$. La deuxième colonne contient le résultat de l'estimation obtenue en utilisant la décomposition en partie réelle et partie imaginaire de la permittivité complexe ; la troisième colonne est relative aux intervalles polaires. La borne d'erreur *a priori* est prise égale à 0.1% de la pseudo-mesure.

	Vraies valeurs	Réelle + Imaginaire	Forme polaire
α_1	0.6	[0.593 ; 0.633]	[0.599, 0.609]
β_1	1	[0.9375 ; 1]	[0.983, 1]
$\Delta\varepsilon_1$	1	[0.971 ; 1.015]	[0.987, 1.008]
τ_1 ($\times 10^{-6}$)	15.915	[15.04 ; 17.39]	[15.62, 16.33]
α_2	0.8	[0.796 ; 0.803]	[0.797, 0.8016]
β_2	0.7	[0.683 ; 0.713]	[0.692, 0.7101]
$\Delta\varepsilon_2$	6	[5.984 ; 6.029]	[5.988, 6.013]
τ_2	0.15915	[0.154 ; 0.166]	[0.154, 0.162]
ε_∞	3	[2.999 ; 3.001]	[2.99999, 3]
t_c		**4h 8 min**	**20 min**

Tableau 16. Paramètres diélectriques estimés pour le modèle d'Havriliak-Negami avec deux modes de relaxation et en utilisant deux représentations des intervalles complexes.

On remarque d'après le tableau 16 que le temps de calcul est beaucoup moins important dans le cas des intervalles complexes polaires, ceci grâce au nombre d'occurrences des variables

assez réduit qui rend le contracteur *propagation-rétropropagation* plus efficace. On doit noter que ceci est valable pour un borne d'erreur *a priori* d'amplitude assez petite.

2.7.2. Avantages et inconvénients de la représentation complexe

Considérons le cas d'un spectre complet obtenu en simulant le modèle d'Havriliak-Negami avec un seul mode de relaxation avec les valeurs suivantes :

$$\alpha_1 = 0.8,\ \beta_1 = 0.5,\ \tau_1 = 0.001591\text{s},\ \Delta\varepsilon_1 = 6,\ \varepsilon_\infty = 3$$

Nous supposons que l'erreur de sortie est bornée et d'amplitude 1% de la pseudo-mesure. L'estimation des paramètres du modèle d'Havriliak-Negami avec un seul mode de relaxation génère en 363 secondes un ensemble de pavés contenant la solution exacte. La projection de cet ensemble sur les axes correspondant aux différents paramètres, donne :

$$\begin{pmatrix} \alpha_1 \\ \beta_1 \\ \tau_1 \\ \Delta\varepsilon_1 \\ \varepsilon_\infty \end{pmatrix} \in \begin{pmatrix} [0.79122,\ 0.809] \\ [0.46914,\ 0.531] \\ [0.001412,\ 0.001785] \\ [5.8689,\ 6.1352] \\ [2.9537,\ 3.0452] \end{pmatrix}$$

On remarque que le temps de calcul est très important par rapport à celui obtenu en utilisant les fonctions explicites des parties réelle et imaginaire (le temps de calcul en utilisant les formes explicites de la partie réelle et la partie imaginaire est $t_c = 108.4$ s) ; on donnera dans la suite une raison à cette constatation.

Considérons maintenant le même cas mais en supposant que la borne sur l'erreur *a priori* est de 0.1% de la pseudo-mesure. L'estimation des paramètres du modèle d'Havriliak-Negami avec un seul mode de relaxation génère en 2.2 secondes les résultats suivants :

$$\begin{pmatrix} \alpha_1 \\ \beta_1 \\ \tau_1 \\ \Delta\varepsilon_1 \\ \varepsilon_\infty \end{pmatrix} \in \begin{pmatrix} [0.79745,\ 0.80252] \\ [0.49307,\ 0.50753] \\ [0.001542,\ 0.001635] \\ [5.9822,\ 6.0175] \\ [2.9887,\ 3.0112] \end{pmatrix}$$

L'estimation de ces paramètres, avec les mêmes hypothèses, en utilisant la forme explicite des parties réelle et imaginaire génère en 14.9 secondes un ensemble dont la projection par rapport aux différents paramètres est donnée par :

$$\begin{pmatrix} \alpha_1 \\ \beta_1 \\ \tau_1 \\ \Delta\varepsilon_1 \\ \varepsilon_\infty \end{pmatrix} \in \begin{pmatrix} [0.797,\ 0.803] \\ [0.493,\ 0.506] \\ [0.001546,\ 0.001643] \\ [5.982,\ 6.015] \\ [2.991,\ 3.009] \end{pmatrix}$$

Ainsi, pour une borne de 1%, le temps de calcul le plus petit est obtenu pour les formes explicites des parties réelle et imaginaire. A l'inverse, pour une borne de 0.1%, le temps de calcul le plus petit est obtenu pour la représentation polaire.

Ces résultats peuvent s'expliquer par le phénomène de dépendance rencontré en utilisant la décomposition en parties réelle et imaginaire, à cause notamment du caractère multi-occurant des paramètres. Dans une telle situation, le contracteur *propagation – rétropropagation* utilisé est moins efficace. Pour une borne de 0.1%, l'algorithme fondé sur les formes polaires est le plus rapide.

Inversement, l'utilisation de la forme polaire contourne le problème de dépendance, mais introduit un pessimisme induit par enveloppement. Lorsque les données expérimentales sont décrites par des intervalles de taille importante, ce pessimisme handicape la forme polaire. Pour une borne de 1%, l'algorithme fondé sur la décomposition en partie réelle/imaginaire est le plus rapide.

2.8. Conclusion

Dans cette première partie du chapitre, nous avons utilisé des techniques d'inversion ensembliste par arithmétique par intervalles pour l'estimation de paramètres diélectriques. Ces méthodes permettent de trouver toutes les valeurs du vecteur de paramètres compatibles avec les mesures et avec les bornes d'erreurs supposées connues *a priori*. Les données utilisées dans cette première partie sont obtenues par simulation.

Dans cette partie, nous avons montré que l'estimation de paramètres dans un contexte à erreurs bornées permet de rejeter un modèle lorsque des données expérimentales ne sont pas compatibles avec ce modèle. Cette approche nous a permis de déterminer le nombre de modes de relaxation diélectrique, inconnu *a priori*.

D'un autre coté, nous avons montré sur un exemple qu'un bon choix de la stratégie de bissection permet de réduire le temps de calcul. En général, la stratégie C est la plus efficace, néanmoins, elle nécessite de calculer la dérivée d'une fonction ; ceci est fait à l'aide de la différentiation automatique.

D'autre part, étant donné que le modèle d'Havriliak-Negami utilisé est à variable complexe, nous avons considéré deux représentations pour la permittivité diélectrique. Dans un premier lieu, nous avons décomposé la sortie du modèle en une partie réelle et une partie imaginaire données par des fonctions explicites. L'avantage d'une telle approche est de ne travailler qu'avec des intervalles réels ; ceci nous a permis d'éviter l'effet d'enveloppement. Dans un

second lieu, nous avons représenté la sortie incertaine du modèle par un intervalle complexe polaire. L'avantage de cette approche est de ne pas chercher des fonctions explicites donnant les parties réelle et imaginaire de la permittivité diélectrique. L'utilisation du modèle à variable complexe est préférable pour la propagation de contraintes étant donné le nombre d'occurrences réduit pour chaque variable. Néanmoins, l'effet d'enveloppement est amplifié en utilisation des intervalles complexes, notamment en présence d'un bruit de mesure important.

3. Estimation de paramètres thermophysiques

Dans cette partie nous allons considérer une seconde application qui consiste à identifier la diffusivité thermique a et la conductivité thermique λ d'un matériau. Le dispositif expérimental utilisé a fait l'objet des deux thèses [TK98] ; [Bou03]. Il a aussi fait l'objet d'une caractérisation garantie par inversion et projection ensembliste [Bra02].

3.1. Le banc d'essai

Le principe de la mesure consiste à placer un échantillon de conductivité et de diffusivité inconnues entre deux plaques métalliques. La plaque dite, « plaque avant » est soumise à une excitation périodique de température. L'autre plaque dite, « plaque arrière » a une face en contact avec l'air à température ambiante (voir figures 13 et 14). Le contact thermique entre différents éléments est assuré par une graisse à haute conductivité thermique. L'ensemble du dispositif est placé dans une enceinte sous vide secondaire. Les températures sont mesurées au sein des deux plaques métalliques. En utilisant ces deux mesures de température, on se propose d'identifier la conductivité et la diffusivité de l'échantillon.

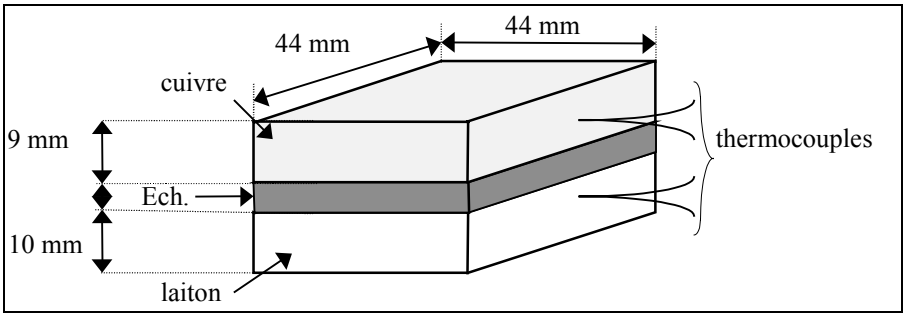

Figure 13 : Schéma du porte échantillon

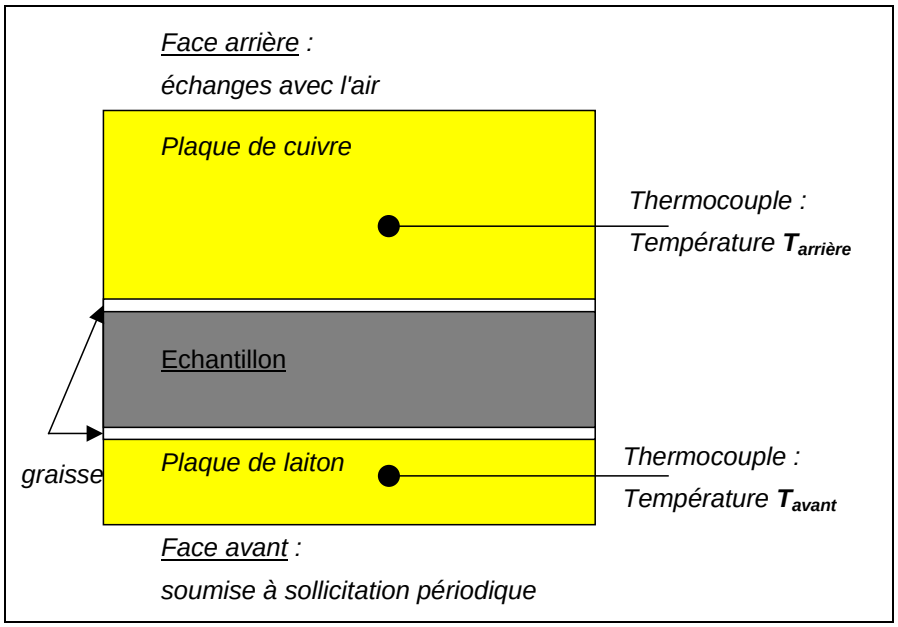

Figure 14 : Plan du banc d'essai

3.2. Le modèle

3.2.1. Les hypothèses de modélisation

Le système thermique étudié est constitué de cinq couches : deux couches métalliques servant de porte-échantillon, une couche homogène constituant l'échantillon et de couches de graisse thermique. Pour chaque couche de matériau homogène, on résout l'équation de la conservation de l'énergie. On utilise aussi les relations suivantes :

- à l'interface entre les couches, la continuité des températures et des flux ;
- sur la face avant, la température imposée est supposée connue ;
- sur la face arrière, le flux échangé avec l'atmosphère résiduelle dans l'enceinte est modélisé par une relation mettant en jeu la température de surface, la température ambiante, supposée connue et un coefficient d'échange surfacique.

3.2.2. Fonction de transfert

Le système est représenté par la fonction de transfert écrite comme étant le rapport entre les températures des portes-échantillons arrière et avant, soit :

$$H\ (j\omega) = \frac{T_{arrière}(j\omega)}{T_{avant}(j\omega)} \tag{27}$$

Nous utilisons une écriture explicite de cette fonction de transfert à l'aide de la représentation « quadripôle » des matériaux homogènes. Cette dernière permet de modéliser chaque matériau

par un quadripôle thermique $\mathbf{Z}(j\omega)$ ayant pour entrées les transformées de Fourier de la température et du flux thermique sur la face avant et pour sorties celles sur la face arrière [LAG 99] [WDM 02].

3.2.3. La modélisation "quadripôle" d'un matériau homogène

On suppose que le système est mono-dimensionnel, c'est-à-dire invariant par translation dans un plan (voir figure 15).

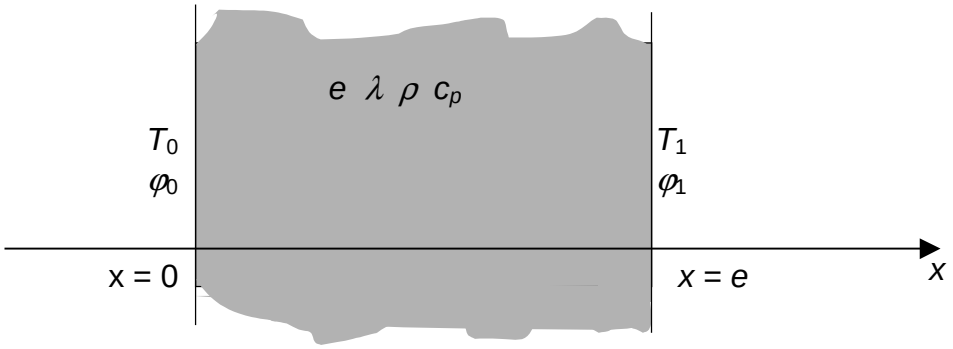

Figure 15 : Matériau homogène

Les variables du systèmes sont dans le tableau 17 :

Grandeur	***Symbole***	***Unité***
Température de la face arrière	T_1	*K*
Température de la face avant	T_0	*K*
Flux surfacique échangé à face arrière	φ_1	W/m^2
Flux surfacique échangé à face avant	φ_0	W/m^2
Masse volumique	ρ	kg/m^3
Capacité calorifique massique	c_p	*J/kgK*
Conductivité	λ	*W/mK*
Epaisseur	e	*m*
Diffusivité	a	m^2/s
Coefficient d'échange surfacique	h	W/m^2K

Tableau 17 : Variables du système

L'équation de la conservation de l'énergie dans le matériau s'écrit :

$$\frac{\partial^2 T(x,t)}{\partial x^2} = \frac{1}{a} \cdot \frac{\partial T(x,t)}{\partial t} \tag{28}$$

où la diffusivité a est définie par $a = \dfrac{\lambda}{\rho c_p}$

On pose les variables suivantes :

- $T(x=0,t) = T_0(t)$
- $T(x=e,t) = T_1(t)$
- $-\lambda \cdot \left.\dfrac{\partial T(x,t)}{\partial x}\right|_{x=0} = \varphi_0(t)$
- $-\lambda \cdot \left.\dfrac{\partial T(x,t)}{\partial x}\right|_{x=e} = \varphi_1(t)$

En procédant à une transformée de Laplace du système d'équations, on obtient :

$$\frac{\partial^2 T(x,s)}{\partial x^2} = \frac{s}{a} \cdot T(x,s) \tag{29}$$

où s désigne la variable de Laplace. On obtient :

- $T(x=0,s) = T_0(s)$
- $T(x=e,s) = T_1(s)$
- $-\lambda \cdot \left.\dfrac{\partial T(x,s)}{\partial x}\right|_{x=0} = \varphi_0(s)$
- $-\lambda \cdot \left.\dfrac{\partial T(x,s)}{\partial x}\right|_{x=e} = \varphi_1(s)$

La solution générale de l'équation est de la forme :

$$0 < x < e, \quad T(x,s) = A(s) \cdot \cosh\left(\sqrt{s/a} \cdot x\right) + B(s) \cdot \sinh\left(\sqrt{s/a} \cdot x\right) \tag{30}$$

On obtient pour A et B les expressions suivantes :

$$A(s) = T_0(s) \quad \text{et} \quad B(s) = \frac{T_1(s) - T_0(s) \cdot \cosh\left(\sqrt{s/a} \cdot e\right)}{\sinh\left(\sqrt{s/a} \cdot e\right)} \tag{31}$$

On obtient pour les flux surfaciques les expressions suivantes :

$$\begin{cases} \varphi_0(s) = -\lambda \cdot \left.\dfrac{\partial T(x,s)}{\partial x}\right|_{x=0} = \dfrac{\lambda \cdot \sqrt{s/a}}{\tanh\left(\sqrt{s/a} \cdot e\right)} \cdot T_0(s) - \dfrac{\lambda \cdot \sqrt{s/a}}{\sinh\left(\sqrt{s/a} \cdot e\right)} \cdot T_1(s) \\ \varphi_1(s) = -\lambda \cdot \left.\dfrac{\partial T(x,s)}{\partial x}\right|_{x=e} = \dfrac{\lambda \cdot \sqrt{s/a}}{\sinh\left(\sqrt{s/a} \cdot e\right)} \cdot T_0(s) - \dfrac{\lambda \cdot \sqrt{s/a}}{\tanh\left(\sqrt{s/a} \cdot e\right)} \cdot T_1(s) \end{cases} \tag{32}$$

Ainsi, il est possible de trouver une relation entre le couple (température – flux surfacique) de la face arrière et le couple (température – flux surfacique) de la face avant, sous une forme matricielle connue en thermique sous le nom de « quadripôle » d'une paroi :

$$\begin{bmatrix} T_0(s) \\ \varphi_0(s) \end{bmatrix} = \mathrm{Z}(s) \cdot \begin{bmatrix} T_1(s) \\ \varphi_1(s) \end{bmatrix} \tag{33}$$

avec

$$\mathrm{Z}(s) = \begin{bmatrix} \cosh\left(\sqrt{s/a} \cdot e\right) & \dfrac{1}{\lambda \cdot \sqrt{s/a}} \cdot \sinh\left(\sqrt{s/a} \cdot e\right) \\ \lambda \cdot \sqrt{s/a} \cdot \sinh\left(\sqrt{s/a} \cdot e\right) & \cosh\left(\sqrt{s/a} \cdot e\right) \end{bmatrix} \tag{34}$$

En utilisant le temps caractéristique de Fourier défini par la relation :

$$\tau = \frac{e^2}{a} \tag{35}$$

et la résistance surfacique du matériau homogène, définie par la relation :

$$R = \frac{e}{\lambda} \tag{36}$$

le quadripôle s'écrit sous la forme suivante :

$$\mathrm{Z}(s) = \begin{bmatrix} \cosh\left(\sqrt{\tau \cdot s}\right) & R \cdot \dfrac{\sinh\left(\sqrt{\tau \cdot s}\right)}{\sqrt{\tau \cdot s}} \\ \dfrac{\sqrt{\tau \cdot s} \cdot \sinh\left(\sqrt{\tau \cdot s}\right)}{R} & \cosh\left(\sqrt{\tau \cdot s}\right) \end{bmatrix} \tag{37}$$

3.2.4. Cas d'un matériau sans inertie

Lorsque le matériau étudié est un matériau qui présente une inertie négligeable, le « quadripôle » devient :

$$Z(s)\big|_{\rho c_p=0} = \begin{bmatrix} 1 & \frac{e}{\lambda} \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & R \\ 0 & 1 \end{bmatrix} \tag{38}$$

où R représente la résistance du matériau.

3.2.5. Echanges paroi-air

On peut aussi écrire le « quadripôle » pour décrire les relations entre le couple (température – flux surfacique) à la surface d'un matériau et le couple (température – flux surfacique) du fluide en contact avec cette surface. On obtient :

$$Z(s)\big|_h = \begin{bmatrix} 1 & \frac{1}{h} \\ 0 & 1 \end{bmatrix} \tag{39}$$

où h est le coefficient d'échange global tenant compte des transferts par convection et rayonnement.

3.2.6. Modèle du dispositif expérimental

<u>Température « avant »</u>

En utilisant les quadripôles, on peut écrire les relations suivantes pour le flux et la température de la plaque avant :

$$\begin{bmatrix} T_{avant}(s) \\ \varphi_{avant}(s) \end{bmatrix} = Z_{Laiton}(s) \cdot Z_{Graisse}(s) \cdot Z_{Ech}(s) \cdot Z_{Graisse}(s) \cdot Z_{Cuivre}(s) \cdot \begin{bmatrix} T_0(s) \\ hT_0(s) \end{bmatrix} \tag{40}$$

avec :

$$Z_{Laiton}(s) = \begin{bmatrix} \cosh\left(\sqrt{\tau_l \cdot s}\right) & R_l \cdot \dfrac{\sinh\left(\sqrt{\tau_l \cdot s}\right)}{\sqrt{\tau_l \cdot s}} \\ \dfrac{\sqrt{\tau_l \cdot s} \cdot \sinh\left(\sqrt{\tau_l \cdot s}\right)}{R_l} & \cosh\left(\sqrt{\tau_l \cdot s}\right) \end{bmatrix} \tag{41}$$

$$Z_{Graisse}(s) = \begin{bmatrix} 1 & R_{Graisse} \\ 0 & 1 \end{bmatrix} \tag{42}$$

$$Z_{Ech}(s)=\begin{bmatrix} \cosh\left(\sqrt{\tau_p \cdot s}\right) & R_p \cdot \dfrac{\sinh\left(\sqrt{\tau_p \cdot s}\right)}{\sqrt{\tau_p \cdot s}} \\ \dfrac{\sqrt{\tau_p \cdot s} \cdot \sinh\left(\sqrt{\tau_p \cdot s}\right)}{R_p} & \cosh\left(\sqrt{\tau_p \cdot s}\right) \end{bmatrix} \tag{43}$$

$$Z_{Cuivre}(s)=\begin{bmatrix} \cosh\left(\sqrt{\tau_c \cdot s}\right) & R_c \cdot \dfrac{\sinh\left(\sqrt{\tau_c \cdot s}\right)}{\sqrt{\tau_c \cdot s}} \\ \dfrac{\sqrt{\tau_c \cdot s} \cdot \sinh\left(\sqrt{\tau_c \cdot s}\right)}{R_c} & \cosh\left(\sqrt{\tau_c \cdot s}\right) \end{bmatrix} \tag{44}$$

Dans l'équation (40), T_0 représente la température de surface de la face arrière de la plaque de cuivre.

Température « arrière »

Pour la température « arrière », on écrit de même :

$$\begin{bmatrix} T_{arrière}(s) \\ \varphi_{arrière}(s) \end{bmatrix} = Z_{CUIVRE/thermocouple}(s) \cdot \begin{bmatrix} T_0(s) \\ hT_0(s) \end{bmatrix} \tag{45}$$

avec :

$$Z_{Cuivre/Thermocouple}(s)=\begin{bmatrix} \cosh\left(\sqrt{\xi^2\tau_c \cdot s}\right) & R_c \cdot \dfrac{\sinh\left(\sqrt{\xi^2\tau_c \cdot s}\right)}{\sqrt{\tau_c \cdot s}} \\ \dfrac{\sqrt{\tau_c \cdot s} \cdot \sinh\left(\sqrt{\xi^2\tau_c \cdot s}\right)}{R_c} & \cosh\left(\sqrt{\xi^2\tau_c \cdot s}\right) \end{bmatrix} \tag{46}$$

où ξ représente le rapport entre la distance thermocouple - face arrière et l'épaisseur de la plaque de cuivre. Ainsi ce dernier quadripôle modélise la partie de la plaque de cuivre située entre le thermocouple et la face arrière.

Réponse fréquentielle

La réponse fréquentielle est donnée par le rapport entre la transformée de Fourier de la température « arrière » et de la température « avant » :

$$H\ (j2\pi f)=\frac{T_{arrière}(j2\pi f)}{T_{avant}(j2\pi f)} \tag{47}$$

3.2.7. Les paramètres du modèle

Les paramètres thermophysiques du laiton et du cuivre sont tirés de la littérature et les paramètres géométriques sont mesurés. Les paramètres supposés connus sont en réalité incertains. Néanmoins, dans le cadre de notre étude, nous allons les considérer certains.

Nous rappelons que le premier but de ce chapitre est de valider la bibliothèque d'intervalles polaires développée dans le chapitre 2. Le cas incertain sera considéré dans une étude ultérieure.

Les paramètres de la plaque de laiton

La plaque de laiton est en contact avec le bloc Peltier.

- Diffusivité : $a_l = 0.33\times10^{-4}\ m^2.s^{-1}$
- Conductivité : $\lambda_l = 100\ W.m^{-1}.K^{-1}$
- Distance entre le thermocouple soudé dans la plaque de laiton et l'interface laiton – échantillon : $e_l = 5\times10^{-3}$ m.

Les paramètres de la plaque de cuivre

- Diffusivité : $a_c = 1.14\times10^{-4}\ m^2.s^{-1}$
- Conductivité : $\lambda_c = 389\ W.m^{-1}.K^{-1}$
- Distance entre le thermocouple soudé dans la plaque de cuivre et l'interface cuivre – échantillon : $\xi\times e_c = 4.5\times10^{-3}$ m.
- Distance entre l'interface cuivre – échantillon et la surface en contact avec l'air ambiant : $e_c = 9\times10^{-3}$ m

Donc $\xi = 0.5$.

Les paramètres des couches de la graisse

- La résistance thermique de la couche de graisse est estimée égale à : $R_c = 1.18\times10^{-4}\ m^2KW^{-1}$

Les échanges thermiques en face arrière

L'utilisation d'un coefficient d'échange h_c pour modéliser la convection à la surface d'une paroi à température ambiante est courante en thermique. Les coefficients utilisés sont soit constants soit variables et dépendent du sens des flux, de la vitesse de fluide, de la position (horizontal – vertical) de la surface etc… Cependant, à l'heure actuelle, on ne sait pas

quantifier ce coefficient pour une paroi dont la température est modulée. D'autre part, il est nécessaire de tenir compte des échanges par rayonnement entre la face arrière et l'environnement. Etant donné que l'ensemble du dispositif de mesure ainsi que l'enceinte sont à température ambiante et que les modulations de température imposées sont de quelques degrés, il est possible de prendre en compte ces échanges en utilisant un coefficient d'échange par rayonnement linéarisé h_r. L'ensemble des échanges par convection et rayonnement est modélisé en introduisant un coefficient d'échange global h :

$$h = h_c + h_r \tag{48}$$

On suppose que le coefficient d'échange à la surface est :

$$\mathrm{h} = 5\ \mathrm{W.m^{-2}.K^{-1}}$$

3.3. Estimation

L'objectif de cette partie est d'étudier la faisabilité de l'inversion ensembliste en utilisant la représentation polaire des intervalles complexes présentée dans le chapitre 2. La caractérisation complète du banc d'essai n'est pas considérée dans ce chapitre.

3.3.1. Représentation complexe

Nous avons vu dans les sections précédentes que le modèle utilisé est à variable complexe. En plus, il est très difficile de décomposer analytiquement la fonction de transfert liant la température de sortie à la température d'entrée en une partie réelle et une partie imaginaire. Il est donc nécessaire d'utiliser une représentation des intervalles complexes. Dans la thèse [Bra02], la représentation rectangulaire a été préférée étant donné la simplicité de l'implémentation des opérations arithmétiques. Dans la suite de ce chapitre, nous allons utiliser la représentation polaire développée dans le chapitre 2 ; ce choix est motivé d'une part par la présence dans le modèle du terme $\sqrt{j2\pi f}$ et d'autre part par la présence de la fonction exponentielle. A noter que ces deux dernières fonctions sont minimales lorsqu'on utilise la forme polaire.

Exemple : Afin d'illustrer les performances de la représentation polaire par rapport à la représentation rectangulaire, on se propose d'évaluer le terme suivant figurant dans le modèle présenté ci-dessus :

$$g(\tau,\ f) = \frac{\sinh\left(\sqrt{\tau \cdot s}\right)}{\sqrt{\tau \cdot s}} = \frac{\sinh\left(\sqrt{\tau \cdot j2\pi f}\right)}{\sqrt{\tau \cdot j2\pi f}} \tag{49}$$

On considère le cas où f = 0.005 Hz et $\tau \in [4,\ 6]$. L'évaluation de la fonction g en utilisant respectivement la représentation rectangulaire g_r et la représentation polaire g_p donne :

$$\left[g_r(\tau,\ f)\right] = [0.646304,\ 1.54104] + j[-0.2521, 0.319481]$$

$$\left[g_p(\tau,\ f)\right] = \left\{[0.816569,1.22499]\ ,\ [6.20811,6.40957]\right\}$$

$$\equiv [0.810056,1.22499]\ + j[-0.0918891,0.154406]$$

On remarque que, pour cet exemple, l'évaluation utilisant la représentation polaire est moins pessimiste que celle basée sur l'arithmétique des intervalles rectangulaires. Cela a donc motivé notre choix pour les intervalles polaires étant donné que le modèle comporte plusieurs fois le terme (49).

3.3.2. Fonctions d'inclusion

Dans la suite de ce chapitre nous allons utiliser uniquement la fonction d'inclusion naturelle de la fonction de transfert liant la température de sortie à celle d'entrée.

3.3.3. Estimation des paramètres thermophysiques d'un échantillon de PVDF

On se propose dans cette section d'identifier les paramètres thermophysiques du PVDF (Polyfluorure de Vinylidène) en utilisant les mesures réelles représentées sur les figures 16 et 17. Les températures des faces avant (plaque de laiton) et arrière (plaque de cuivre) sont mesurées à l'aide de thermocouples. Les signaux fournis par les thermocouples sont amplifiés, filtrés (filtre passe-bas à 4 Hz) et linéarisés à l'aide de modules de conditionnement [Bou03]. Les données sont échantillonnées à une fréquence de 1kHz. Une moyenne de 50 mesures est effectuée à chaque seconde. La durée totale d'une expérience est fixée à 67 minutes.

Une série de 20 expériences a été réalisée en utilisant les mêmes conditions expérimentales. Lors de chaque expérience, on utilise un signal d'excitation composé d'une somme de 5 sinusoïdes de fréquences f_0, $2f_0$, $4f_0$, $8f_0$ et $16f_0$, avec f_0 = 2.5mHz. Une fonction de transfert H($j\omega$) est calculée après chaque expérience. On obtient ainsi pour chaque fréquence d'excitation, 20 valeurs de la fonction de transfert thermique expérimentale, à partir desquelles sont obtenues les bornes d'erreur de $H\left(j2\pi f\right)$.

Sur les figures 16 et 17, nous avons tracé respectivement les bornes inférieure et supérieure du module (en échelle logarithmique) et de la phase (en échelle semi-logarithmique) de la fonction de transfert pour différentes fréquences.

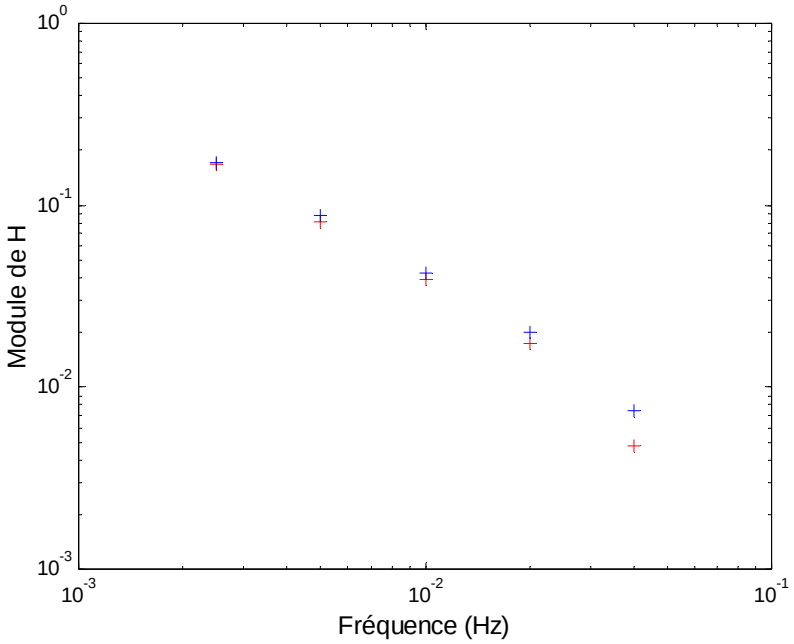

Figure 16 : Bornes du module des mesures

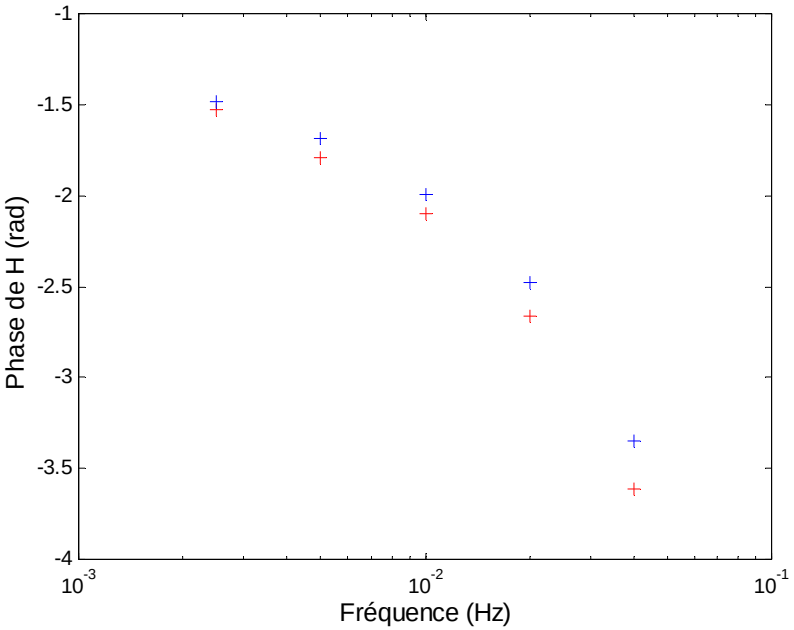

Figure 17 : Bornes de la phase des mesures

Les domaines de recherche initiaux des paramètres à estimer sont :

$$\sqrt{\tau_p} \in [1, 30]\,\mathrm{s}^{1/2} \text{ et } R_p \in \left[10^{-4}, 5\right]\mathrm{m}^2.\mathrm{K}.\mathrm{W}^{-1}$$

En utilisant SIVIAP avec $\eta = 0.001$, nous avons obtenu en 20 secondes l'ensemble des pavés tracés sur la figure 18. L'approximation intérieure est tracée en clair et l'ensemble des pavés indéterminés en foncé.

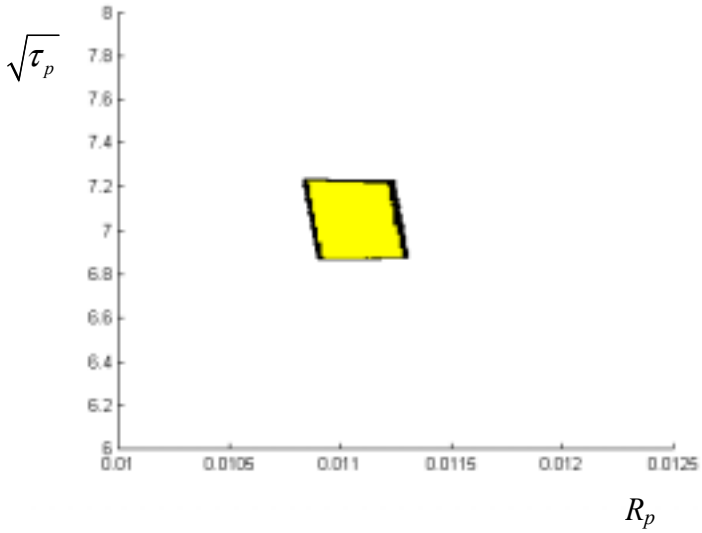

Figure 18 : Ensemble de pavés obtenus par SIVIAP pour le PVDF

La projection de ces ensembles par rapport aux deux axes fournit un encadrement extérieur de l'incertitude associée à chacun des paramètres :

$$\sqrt{\tau_p} \in [6.8694,\ 7.2375]\ s^{1/2}$$

$$R_p \in [0.010819,\ 0.011315]\ m^2.K.W^{-1}$$

On obtient alors pour un échantillon d'épaisseur $e_p = 2mm$:

$$\lambda_p \in [0.1767,\ 0.1848]\ Wm^{-1}K^{-1}$$

$$a_p \in [7.636,\ 8.476]\cdot 10^{-8}\ m^2s^{-1}$$

soit :

$$\lambda_p = 0.181\ \pm\ 2.2\%\ Wm^{-1}K^{-1}$$

$$a_p = 8.056\cdot 10^{-8} \pm 5.2\%\ m^2s^{-1}$$

Par ailleurs, les valeurs obtenues pour ce même matériau dans des conditions expérimentales similaires en utilisant une minimisation quadratique sont [Bou03] :

$$\lambda_p = 0.180\ \pm\ 0.6\%\ Wm^{-1}K^{-1}$$

$$a_p = 9.120\cdot 10^{-8}\ \pm\ 4.2\%\ m^2s^{-1}$$

Enfin, les valeurs fournies par le fabriquant sont :

$$0.1 \leq \lambda_p \leq 0.25$$

$$4.2 \cdot 10^{-8} \leq a_p \leq 1.1 \cdot 10^{-7}\, m^2 s^{-1}$$

On constate donc que les valeurs estimées sont compatibles avec celles proposées dans la littérature et celles obtenues par minimisation quadratique.

3.4. Conclusion

Dans cette partie, nous avons identifié les paramètres thermophysiques du PVDF à l'aide de l'algorithme d'inversion ensembliste SIVIAP, où l'évaluation de la sortie du modèle a été réalisée en utilisant la bibliothèque d'arithmétique des intervalles complexes développée dans le chapitre 2. Nous avons obtenu une approximation intérieure pour l'ensemble solution prouvant ainsi l'existence d'une solution. Les valeurs identifiées par SIVIAP sont en accord avec les valeurs issues de la littérature.

La caractérisation complète du dispositif expérimental avec la nouvelles bibliothèque, incluant la prise en compte des incertitudes de modèle par projection ensembliste, fera l'objet d'une étude ultérieure.

4. Conclusion générale

Ce chapitre a été consacré à l'application des méthodes ensemblistes à l'estimation de paramètres physiques pour deux applications différentes. Dans le cadre de l'analyse diélectrique, le modèle utilisé, dit de Havriliak-Negami, est donné par une fonction à variable complexe.

Pour l'évaluer, deux approches ont été considérées : la première est basée sur la décomposition explicite de cette fonction en deux parties réelle et imaginaire ; l'arithmétique des intervalles réels est alors utilisée. Néanmoins, ces fonctions explicites contiennent des variables multi-occurrentes, le contracteur *propagation – rétropropagation* n'est alors pas optimal. La seconde approche consiste à utiliser l'arithmétique des intervalles complexes polaires introduite dans le chapitre 2. L'avantage est d'éviter le problème de dépendance ; ainsi, le contracteur utilisé avec SIVIA est optimal. Néanmoins, l'effet d'enveloppement rend ces fonctions d'inclusion pessimistes. Dans une étude ultérieure, nous allons combiner la décomposition en parties réelle et imaginaire pour l'évaluation des fonctions d'inclusion et la représentation polaire pour la contraction. Par ailleurs, nous notons qu'aucune approximation intérieure de l'ensemble solution n'a été atteinte ; ceci explique la difficulté numérique du problème étudié. Dans une étude ultérieure, nous proposons de combiner les méthodes garanties avec des techniques de recherche ponctuelles afin de pouvoir caractériser les approximations intérieures.

Dans un second temps, nous avons montré que l'estimation de paramètres dans un contexte à erreurs bornées permet de rejeter un modèle lorsque des données expérimentales ne sont pas compatibles avec ce modèle. Cette approche nous a permis de déterminer le nombre de modes de relaxation, inconnu *a priori*.

Enfin, nous avons montré sur un exemple numérique que le choix de la stratégie de bissection permet dans certain cas, de réduire de manière significative le temps de calcul de l'algorithme SIVIAP. En effet, le temps de calcul est nettement amélioré en utilisant la stratégie de bissection **C** : ainsi, il est parfois inutile de bissecter un paramètre non influent.

Dans la deuxième partie du chapitre, nous nous sommes intéressés à une deuxième application concernant l'estimation des propriétés thermo-physiques (conductivité et diffusivité thermique) de matériaux par inversion ensembliste. Le modèle utilisé est non linéaire et à variable complexe. De plus, nous ne disposons pas de formule explicite pour la décomposition en parties réelle et imaginaire. Dans ce chapitre, nous avons montré qu'il est possible d'estimer ces paramètres thermophysiques en utilisant l'arithmétique des intervalles complexes polaires. La continuation de ce travail consistera à associer la représentation polaire aux algorithmes de projection développés dans [Bra02] et utilisés pour la caractérisation garantie du banc d'essai.

Chapitre 4

Intégration numérique garantie des équations différentielles

1. Introduction

Les modèles obtenus à l'aide des principes fondamentaux de la physique sont souvent décrits par des équations différentielles. La résolution symbolique de ces équations est dans la plupart des cas très difficile, voire impossible à réaliser étant donné le caractère non linéaire de ces équations. Plusieurs schémas numériques permettant de calculer une approximation de ces équations ont été proposés afin de résoudre ce problème. Dans certaines applications, ces équations différentielles contiennent des paramètres incertains appartenant à un ensemble connu *a priori*. Dans d'autres cas, la condition initiale n'est pas non plus connue d'une manière exacte. Pour résoudre ce problème à l'aide de méthodes ponctuelles, on effectue des tirages aléatoires dans l'ensemble initial. Ces méthodes sont généralement robustes et fiables pour la plupart des applications. Mais il existe des cas où elles retournent des solutions imprécises. D'autre part, il existe des situations où les bornes de l'erreur sont désirées ou nécessaires et peuvent être critiques pour la fiabilité du système. Il est alors préférable de s'assurer que la solution appartient à un domaine connu.

L'utilisation des méthodes garanties permet de remédier à ce problème ; ces dernières permettent de calculer deux bornes inférieure et supérieure d'un pavé dont on garantit qu'il contient la solution de l'équation différentielle pour toute valeur des paramètres incertains ou de la condition initiale. Ces méthodes permettent aussi de s'assurer que le problème étudié contient ou non une solution. En effet, si le pavé retourné est vide, alors le système d'équations différentielles ne possède pas de solution ; dans le cas contraire son existence est assurée. A noter que la propriété de garantie n'est pas assurée en utilisant les méthodes classiques d'intégrations d'équations différentielles ordinaires (EDOs) étant donné que seule une approximation de la solution est calculée.

Parmi les raisons qui ont empêché l'utilisation des techniques d'analyse par intervalles pour la résolution des équations différentielles, on peut noter le temps de calcul très important par rapport aux méthodes standards. Cependant ce problème a été résolu par l'utilisation de calculateurs à capacité importante. De plus, dans certaines situations, l'utilisation de l'analyse par intervalles peut ne pas exiger un temps de calcul et un espace mémoire plus importants

que dans le cas des méthodes standards, par exemple dans le cas de la résolution d'équations avec des paramètres non connus mais situés dans des domaines connus.

Plusieurs méthodes ont été développées afin de résoudre le système (1) ci-dessous. Dans ce chapitre, le principe général de résolution ainsi que les méthodes les plus utilisées seront présentés. Pour plus de détails, on peut consulter [Moo66] [Eij81] [Loh88] [Rih94] [BM98], [Ned99]. Dans la section 2, nous allons rappeler le principe général des méthodes garanties fondé sur deux étapes : la première, détaillée dans la section 3, consiste à prouver l'existence et l'unicité de la solution et déterminer un encadrement *a priori* de la solution. L'encadrement *a priori* est contracté en une deuxième phase afin de donner l'ensemble solution, cette étape est détaillée dans la section 4.

2. Principe général

Considérons le système d'équations différentielles défini par :

$$\begin{cases} \dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}(t)) \\ \mathbf{x}(t_0) \in [\mathbf{x}_0] \end{cases} \tag{1}$$

Avec $t \in [t_0, T]$ et $\mathbf{f} \in C^{k-1}(D)$, $D \subseteq \mathbb{R}^n$ un ensemble ouvert. On peut bien sûr considérer le cas où l'EDO (1) contient des paramètres incertains représentés par des intervalles réels.

Le but de ce chapitre est de calculer, à chaque instant $t_j \in \{t_1, t_2 \ldots t_N\}$ dans l'intervalle $[t_0, t_N]$, un pavé $[\mathbf{x}_j]$, le plus petit possible, contenant d'une manière garantie la solution de (1) à t_j et pour tout $\mathbf{x}(t_0) \in [\mathbf{x}_0]$.

La majorité des méthodes de résolution garantie d'équations différentielles comportent, à chaque pas d'intégration, deux étapes :

- La première consiste à vérifier l'existence et l'unicité de la solution. Dans le cas où l'existence est prouvée, un pas d'intégration h_j ainsi qu'un encadrement *a priori* $[\tilde{\mathbf{x}}_j]$ vérifiant :

  $$\forall t \in [t_j, t_{j+1}],\ \mathbf{x}(t) \in [\tilde{\mathbf{x}}_j]$$

  sont calculés, avec $t_{j+1} = t_j + h_j$.

- La deuxième consiste à calculer un pavé $[\mathbf{x}_{j+1}] \subseteq [\tilde{\mathbf{x}}_j]$ contenant la solution de (1) à t_{j+1} pour tout $\mathbf{x}(t_j) \in [\mathbf{x}_j]$. Ceci consiste donc à contracter, ou réduire, le pavé $[\tilde{\mathbf{x}}_j]$. Cette deuxième étape est généralement réalisée à l'aide d'un développement de Taylor d'ordre élevé et en utilisant des fonctions d'inclusion centrée. Il est aussi possible d'utiliser des contracteurs à point fixe.

3. Existence, unicité et solution *a priori*

Cette étape est réalisée à l'aide de l'opérateur de Picard-Lindelöf et du théorème du point fixe [Ned99]. C'est une méthode du premier ordre permettant de calculer un pas $h_j = t_{j+1} - t_j$ et un ensemble *a priori* $[\tilde{\mathbf{x}}_j]$ contenant la trajectoire de la solution de (1) pour $t \in [t_j, t_{j+1}]$ et pour $\mathbf{x}(t_j) \in [\mathbf{x}_j]$. Considérons alors l'équation suivante :

$$\begin{cases} \dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}(t)), \quad t \in [t_j, t_{j+1}] \\ \mathbf{x}(t_j) = \mathbf{x}_j \end{cases} \tag{2}$$

où la fonction **f** vérifie les mêmes hypothèses données dans la section 1.

Définition 1 : On appelle opérateur de Picard-Lindelöf l'opérateur **T** défini par :

$$\mathbf{T}\big(\mathbf{x}(t)\big) = \mathbf{x}_j + \int_{t_j}^{t} \mathbf{f}(\mathbf{x}(s))ds \tag{3}$$

où **x** et **f** sont deux fonctions et t_j est une variable indépendante qui peut représenter la variable temps. ♦

Propriété 1 : Lorsque **x** et **f** représentent les fonctions définies dans (2), alors $(T\mathbf{x})(t_{j+1}) = \mathbf{x}_{j+1}$.

La démonstration de cette propriété est immédiate, il suffit d'intégrer l'équation $\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}(t))$ sur $t \in [t_j, t_{j+1}]$. ♦

Théorème 1 - Théorème du point fixe ou de Banach :

Soit φ : Y $\rightarrow$ Y définie dans un espace métrique complet avec $d(.,.)$ comme distance métrique, et soit $0 \leq \gamma < 1$, alors si :

$$d(\varphi(\mathbf{x}) , \varphi(\mathbf{y})) \leq \gamma d(\mathbf{x} , \mathbf{y}) \tag{4}$$

pour tout **x** et **y** $\in$ Y. Alors φ possède un seul point fixe $\mathbf{x}^* \in$ Y.

Proposition 1 : On considère l'équation différentielle (2), soient $h_j = t_{j+1} - t_j$ et $[\tilde{\mathbf{x}}_j] \subseteq D$ tels que :

$$\mathbf{x}_j + [0, h_j] \cdot \mathbf{f}([\tilde{\mathbf{x}}_j]) \subseteq [\tilde{\mathbf{x}}_j] \tag{5}$$

alors le pavé $[\tilde{\mathbf{x}}_j]$ contient toute la trajectoire de $\mathbf{x}(t)$, solution de l'équation différentielle (2), entre t_j et t_{j+1}.

Preuve : Supposons que l'inclusion (5) est vérifiée ; soit U l'ensemble des fonctions continues sur $[t_j, t_{j+1}]$ et à valeurs dans le pavé $[\tilde{\mathbf{x}}_j]$:

$$U = \left\{ \mathbf{u} \mid \mathbf{u} \in C^0\left([t_j, t_{j+1}]\right) \text{ et } \forall t \in [t_j, t_{j+1}],\ \mathbf{u}(t) \in [\tilde{\mathbf{x}}_j] \right\} \tag{6}$$

En appliquant l'opérateur de Picard-Lindelöf à une fonction $\mathbf{u} \in U$, on obtient :

$$\mathbf{T}\left(\mathbf{u}(t)\right) = \mathbf{x}_j + \int_{t_j}^{t} \mathbf{f}(\mathbf{u}(s))ds = \mathbf{x}(t)$$

D'autre part on a $\mathbf{u}(s) \in [\tilde{\mathbf{x}}_j]$, donc :

$$\mathbf{T}\left(\mathbf{u}(t)\right) \in \mathbf{x}_j + \int_{t_j}^{t} \mathbf{f}([\tilde{\mathbf{x}}_j])ds$$

Sachant que $t \in [t_j, t_{j+1}]$ et $h_j = t_{j+1} - t_j$, alors :

$$\mathbf{T}\left(\mathbf{u}(t)\right) \in \mathbf{x}_j + [0, h_j] \cdot \mathbf{f}([\tilde{\mathbf{x}}_j])$$

En tenant compte de (5), on obtient :

$$\mathbf{T}\left(\mathbf{u}(t)\right) \in [\tilde{\mathbf{x}}_j] \tag{7}$$

On constate alors d'après (7) que :

$$\forall \mathbf{u} \in U,\ \mathbf{T}\left(\mathbf{u}(t)\right) \in [\tilde{\mathbf{x}}_j]$$

Et comme $\mathbf{T}\left(\mathbf{u}(t)\right) \in C^0([t_j, t_{j+1}])$, alors :

$$\mathbf{T}\left(U\right) \subseteq U \tag{8}$$

L'opérateur de Picard-Lindelöf est alors contractant [Eijgenraam, 81] et possède un seul point fixe dans U. D'autre part, on sait que le point fixe de l'opérateur de Picard-Lindelöf est solution de l'équation différentielle (2) pour $t \in [t_j, t_{j+1}]$ avec $\mathbf{x}_j$ comme condition initiale. On peut donc déduire que le pavé $[\tilde{\mathbf{x}}]$ contient toute la trajectoire de la solution de (2) pour $t \in [t_j, t_{j+1}]$. ♦

Proposition 2 : Supposons maintenant que dans (2), l'état initial $\mathbf{x}(t_j)$ n'est pas connu d'une manière exacte mais $\mathbf{x}(t_j) \in [\mathbf{x}_j]$; dans ce cas un pavé $[\tilde{\mathbf{x}}]$ vérifiant l'inclusion suivante :

$$[\mathbf{x}_j] + [0, h_j] \cdot \mathbf{f}([\tilde{\mathbf{x}}]) \subseteq [\tilde{\mathbf{x}}] \tag{9}$$

contient d'une manière garantie la trajectoire de l'équation différentielle définie dans (2) et ce quelle que soit la condition initiale $\mathbf{x}(t_j) \in [\mathbf{x}_j]$. ♦

La preuve de la proposition 2 est immédiate.

Il est clair d'après la proposition 2 que si on réussit à calculer un pavé $[\tilde{\mathbf{x}}_j]$ vérifiant l'inclusion (9), alors l'équation différentielle :

$$\begin{cases} \dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}(t)), \quad t \in [t_j, t_{j+1}] \\ \mathbf{x}(t_j) \in [\mathbf{x}_j] \end{cases} \tag{10}$$

possède une solution unique à chaque instant *t* et pour une condition initiale $\mathbf{x}(t_j) \in [\mathbf{x}_j]$.

Calculer le plus petit pavé $[\tilde{\mathbf{x}}_j]$ vérifiant (9) est un problème numériquement difficile ; en général on se contente d'une approximation extérieure. Néanmoins, le fait d'avoir une solution *a priori* pessimiste peut être amorti dans la phase de correction (ou de contraction) consistant à calculer un pavé $[\mathbf{x}_{j+1}]$ contenant la solution de l'EDO à l'instant t_{j+1}.

On présente ici un algorithme assez simple à mettre en œuvre afin de calculer un pavé $[\tilde{\mathbf{x}}_j]$ et un pas d'intégration h_j vérifiant l'inclusion (9).

Algorithme 1 : **validation**

1. Entrées : $[\mathbf{x}_{j+1}]$, h_j, h_{min}, α, ε, **f**, *Valide*:= *faux* ;
2. Tant que ($h_j > h_{min}$ et $Valide \neq vrai$)
3. $\quad [\tilde{\mathbf{x}}_j] := [\mathbf{x}_j] + [0, h_j] \cdot \mathbf{f}([\mathbf{x}_j])$
4. $\quad [\tilde{\mathbf{x}}_j] := [\tilde{\mathbf{x}}_j] + [-\varepsilon, \varepsilon] \cdot \left|[\tilde{\mathbf{x}}_j]\right|$
5. $\quad$ if $\left([\mathbf{x}_j] + [0, h_j] \cdot \mathbf{f}([\tilde{\mathbf{x}}_j]) \subseteq [\tilde{\mathbf{x}}_j]\right)$
6. $\quad\quad Valide := vrai$;
7. $\quad$ sinon
8. $\quad\quad h_j := \alpha \cdot h_j$;
9. $\quad\quad$ Aller à l'étape 3.
10. Sorties : *Valide*, $[\tilde{\mathbf{x}}_j]$, h_j.

Le principe de l'algorithme est de prendre initialement pour $[\tilde{\mathbf{x}}_j]$ un pavé contenant $[\mathbf{x}_j]$ puis de le gonfler en espérant vérifier l'inclusion (9). Si tel n'est pas le cas, alors le pas d'intégration fixé au début est réduit. Si après quelques itérations, le pas devient plus petit que la valeur minimale fixée *a priori*, l'algorithme est arrêté. Le paramètre α est en général choisi tel que $0.5 \leq \alpha \leq 1$; en pratique, on prend $\alpha = 0.8$ [Ned99]. Le choix de ε influe sur le pessimisme de la solution *a priori* ; en effet, s'il est trop petit le temps d'exécution de l'algorithme devient assez important. En contre partie, une grande valeur de ε rend la solution *a priori* pessimiste.

Cet algorithme est une adaptation d'un algorithme proposé dans [Ned99] et qui utilise un développement de Taylor d'ordre élevé à la place de l'inclusion de premier ordre (9) afin de prouver l'existence et l'unicité de la solution de l'EDO.

Remarque 1 : On a vu que la vérification de l'existence et de l'unicité de la solution a été effectuée en utilisant l'opérateur de Picard-Lindelöf. L'encadrement a priori est donc calculé à l'aide d'un développement en série de Taylor de premier ordre ; il est général plus pessimiste que celui calculé dans [NJP01] en se basant sur un développement de Taylor d'ordre élevé.

4. Réduction de la solution

On considère de nouveau l'équation différentielle (10) ; dans ce paragraphe on suppose que l'existence et l'unicité de la solution de (10) pour chaque $t \in [t_j, t_{j+1}]$ ont été validées et que le pas h_j et la solution *a priori* $[\tilde{\mathbf{x}}_j]$ ont été calculés.

4.1. Intégration à l'aide du développement de Taylor

Le but de cette partie est de pouvoir calculer, à l'aide d'un développement de Taylor, un pavé $[\mathbf{x}_{j+1}]$ vérifiant :

$$\begin{cases} [\mathbf{x}_{j+1}] \subseteq [\tilde{\mathbf{x}}_j] \\ \mathbf{x}(t_{j+1}) \in [\mathbf{x}_{j+1}],\ \forall\, \mathbf{x}(t_j) \in [\mathbf{x}_j] \end{cases} \tag{11}$$

Théoreme 2 : Théorème de Taylor:

On considère l'équation différentielle (2), avec **f** une fonction de classe C^k sur $[t_j, t_{j+1}]$, alors:

$$\mathbf{x}(t_{j+1}) = \mathbf{x}(t_j) + \sum_{i=1}^{k-1} \frac{1}{i!} \left(\mathbf{x}^{(i)}(t_j)\right) \cdot \left(t_{j+1} - t_j\right)^i + \mathbf{R}_k\left(t_{j+1}, t_j, \theta\right) \tag{12}$$

et

$$\exists\, \theta \in [0,1] \mid \mathbf{R}_k\left(t_{j+1}, t_j, \theta\right) = (t_{j+1} - t_j)^k \int_0^1 \mathbf{x}^{(k)}(\theta t_{j+1} + (1-\theta) t_j) \frac{(1-\theta)^{k-1}}{(k-1)!} d\theta \tag{13}$$

Proposition 2 : La solution $\mathbf{x}_{j+1}$ de l' EDO (2) à l'instant t_{j+1} avec $\mathbf{x}(t_j)=\mathbf{x}_j$ vérifie l'inclusion suivante :

$$\mathbf{x}_{j+1} \in \mathbf{x}_j + \sum_{i=1}^{k-1} \mathbf{f}^{[i]}(\mathbf{x}_j) h_j^i + h_j^k \mathbf{f}^{[k]}([\tilde{\mathbf{x}}_j]) \tag{14}$$

avec $h_j = t_{j+1} - t_j$, $[\tilde{\mathbf{x}}_j]$ vérifiant (9) et :

$$\begin{aligned} \mathbf{f}^{[1]} &= \mathbf{x}^{(1)} = \mathbf{f} \\ \mathbf{f}^{[2]} &= \frac{1}{2}\mathbf{x}^{(2)} = \frac{1}{2}\frac{d\mathbf{f}}{d\mathbf{x}}\mathbf{f} \\ &\vdots \\ \mathbf{f}^{[i]} &= \frac{1}{i!}\mathbf{x}^{(i)} = \frac{1}{i}\frac{d\mathbf{f}^{[i-1]}}{dx}\mathbf{f} \end{aligned} \tag{15}$$

Les coefficients de Taylor $\mathbf{f}^{[i]}$ sont obtenus à l'aide de la différentiation automatique [Ral81] [RC96].

Preuve : On suppose que nous avons prouvé l'existence et l'unicité de la solution de l' EDO (2) pour tout $t \in [t_j, t_{j+1}]$ et qu'un pavé $[\tilde{\mathbf{x}}_j]$ vérifiant (9) est disponible, alors la solution à t_{j+1} vérifie (12) et (13), avec :

$$\mathbf{R}_k(t_{j+1}, t_j) = h^k \int_0^1 \mathbf{x}^{(k)}(\theta t_{j+1} + (1-\theta)t_j) \frac{(1-\theta)^{k-1}}{(k-1)!} d\theta$$

D'autre part on a :

$$\mathbf{x}^{(k)} = k \cdot \mathbf{f}^{[k]}$$

alors :

$$\mathbf{R}_k(t_{j+1}, t_j) = h^k \int_0^1 k\mathbf{f}^{[k]}(\mathbf{x}(\theta t_{j+1} + (1-\theta)t_j))(1-\theta)^{k-1} d\theta$$

Etant donné que $[\tilde{\mathbf{x}}_j]$ contient toute la trajectoire de la solution de l'EDO pour $t \in [t_j, t_{j+1}]$, alors on obtient l'appartenance :

$$\mathbf{R}_k(t_{j+1}, t_j) \in h^k \int_0^1 k\mathbf{f}^{[k]}([\tilde{\mathbf{x}}_j])(1-\theta)^{k-1} d\theta = h^k \mathbf{f}^{[k]}([\tilde{\mathbf{x}}_j]) \int_0^1 k(1-\theta)^{k-1} d\theta$$

enfin on obtient :

$$\mathbf{R}_k(t_{j+1}, t_j) \in h^k \mathbf{f}^{[k]}([\tilde{\mathbf{x}}_j]) \tag{16}$$

La relation (14) découle directement de (12) et (16). ♦

L'extension de (14) au cas où la condition initiale $\mathbf{x}(t_j) \in [\mathbf{x}_j]$ est donnée par la fonction d'inclusion naturelle de (14), s'obtient de la sorte :

$$\mathbf{x}_{j+1} \in [\mathbf{x}_{j+1}] = [\mathbf{x}_j] + \sum_{i=1}^{k-1} \mathbf{f}^{[i]}([\mathbf{x}_j]) h^i + h^k \mathbf{f}^{[k]}([\tilde{\mathbf{x}}_j]) \quad (17)$$

A l'aide de la formule (17) on peut donc intégrer l'équation différentielle (1) d'une manière garantie. Néanmoins, la taille des intervalles $[\mathbf{x}_1], [\mathbf{x}_2] \ldots [\mathbf{x}_N]$ est dans notre cas une suite strictement croissante, en effet :

$$w\left(\left[\mathbf{x}_{j+1}\right]\right) = w\left(\left[\mathbf{x}_j\right]\right) + \sum_{i=1}^{k-1} w\left(\left[\mathbf{x}_j\right]_i\right) h^i + w\left(\mathbf{f}^{[k]}\left(\left[\tilde{\mathbf{x}}_j\right]\right)\right) h^k$$

$$\geq w\left(\left[\mathbf{x}_j\right]\right)$$

Par conséquent, la méthode d'intégration de l'EDO en utilisant le schéma (17) diverge après quelques pas d'intégration. Dans les sections suivantes, la formule (17) est améliorée afin d'éviter ce problème.

4.2. Méthode directe

On considère de nouveau (10) ; pour améliorer (17), les coefficients de Taylor sont évalués en utilisant la forme centrée [Rih94].

A l'aide de la forme centrée donnée dans le chapitre 1, on obtient :

$$\mathbf{f}^{[i]}([\mathbf{x}_j]) = \mathbf{f}^{[i]}(\hat{\mathbf{x}}_j) + \mathbf{J}\,(\mathbf{f}^{[i]}; [\mathbf{x}_j])([\mathbf{x}_j] - \hat{\mathbf{x}}_j) \quad (18)$$

où :

- $\hat{\mathbf{x}}_j \in [\mathbf{x}_j]$: en général on prend le centre,
- $\mathbf{J}\,(\mathbf{f}^{[i]}; [\mathbf{x}_j])$: une fonction d'inclusion du Jacobien du $i^{ème}$ coefficient de Taylor évalué sur le pavé $[\mathbf{x}_j]$.

Cette forme centrée est préférée à la fonction d'inclusion naturelle du $i^{ème}$ coefficient de Taylor lorsque la taille du pavé $[\mathbf{x}_j]$ est assez petite, dans le cas contraire, cette forme centrée n'est pas moins pessimiste.

En introduisant la forme centrée dans (17), on obtient :

$$[\mathbf{x}_{j+1}] = \hat{\mathbf{x}}_j + \sum_{i=1}^{k-1} h_j^i \cdot \mathbf{f}^{[i]}\left(\hat{\mathbf{x}}_j\right) + h_j^k \cdot \mathbf{f}^{[k]}\left([\tilde{\mathbf{x}}_j]\right) + \left\{\mathbf{I} + \sum_{i=1}^{k-1} h_j^i \mathbf{J}\left(\mathbf{f}^{[i]}; [\mathbf{x}_j]\right)\right\}\left([\mathbf{x}_j] - \hat{\mathbf{x}}_j\right) \quad (19)$$

où $\mathbf{I}$ est la matrice identité de dimension appropriée.

La majorité des méthodes d'intégration garantie d'équations différentielles sont basées sur l'expression (19) [Moo66] [Eij81] [Rih94] [Loh88].

On pose alors :

$$\left[\mathbf{S}_j\right] = \left\{\mathbf{I} + \sum_{i=1}^{k-1} h_j^i \mathbf{J}\left(\mathbf{f}^{[i]};\ [\mathbf{x}_j]\right)\right\} \tag{20}$$

$$\left[\mathbf{v}_{j+1}\right] = \hat{\mathbf{x}}_j + \sum_{i=1}^{k-1} h_j^i \cdot \mathbf{f}^{[i]}\left(\hat{\mathbf{x}}_j\right) + h_j^k \cdot \mathbf{f}^{[k]}\left([\tilde{\mathbf{x}}_j]\right) \tag{21}$$

$$\hat{\mathbf{x}}_{j+1} = m\left(\left[\mathbf{v}_{j+1}\right]\right) = m\left(\hat{\mathbf{x}}_j + \sum_{i=1}^{k-1} h_j^i \cdot \mathbf{f}^{[i]}\left(\hat{\mathbf{x}}_j\right) + h_j^k \cdot \mathbf{f}^{[k]}\left([\tilde{\mathbf{x}}_j]\right)\right) \tag{22}$$

où $m([\mathbf{x}])$ représente le centre du pavé $[\mathbf{x}]$. Le pavé $[\mathbf{v}_{j+1}]$ est une approximation ponctuelle de la solution à l'instant t_{j+1} en prenant comme condition initiale $\mathbf{x}_j = \hat{\mathbf{x}}_j$.

En tenant compte des notations (20) à (22), l'expression (19) peut être réécrite sous la forme :

$$[\mathbf{x}_{j+1}] = \left[\mathbf{v}_{j+1}\right] + \left[\mathbf{S}_j\right]\left([\mathbf{x}_j] - \hat{\mathbf{x}}_j\right) \tag{23}$$

La méthode d'intégration, appelée méthode directe, est basée sur le schéma numérique (23), elle est résumée dans l'algorithme suivant :

Algorithme 2 : Méthode directe

1. Entrées : $[\tilde{\mathbf{x}}_j]$, h_j, $[\mathbf{x}_j]$, $\hat{\mathbf{x}}_j$;
2. Calculer
3. $\left[\mathbf{v}_{j+1}\right] = \hat{\mathbf{x}}_j + \sum_{i=1}^{k-1} h_j^i \cdot \mathbf{f}^{[i]}\left(\hat{\mathbf{x}}_j\right) + h_j^k \cdot \mathbf{f}^{[k]}\left([\tilde{\mathbf{x}}_j]\right)$;
4. $\left[\mathbf{S}_j\right] = \left\{\mathbf{I} + \sum_{i=1}^{k-1} \mathbf{J}\left(\mathbf{f}^{[i]};\ [\mathbf{x}_j]\right)\right\}$;
5. $[\mathbf{x}_{j+1}] = \left[\mathbf{v}_{j+1}\right] + \left[\mathbf{S}_j\right]\left([\mathbf{x}_j] - \hat{\mathbf{x}}_j\right)$
6. $\hat{\mathbf{x}}_{j+1} = m\left(\left[\mathbf{v}_{j+1}\right]\right) = m\left([\mathbf{x}_{j+1}]\right)$;
7. Sorties : $[\mathbf{x}_{j+1}]$, $\hat{\mathbf{x}}_{j+1}$

Comme on l'a noté dessus, le pavé $[\mathbf{v}_{j+1}]$ est une approximation ponctuelle de la solution à l'instant t_{j+1} en prenant comme condition initiale $\mathbf{x}_j = \hat{\mathbf{x}}_j$. L'incertitude sur ce terme est

introduite par le dernier élément de la série de Taylor. Néanmoins, cette incertitude peut être réduite en utilisant un ordre élevé dans le développement de Taylor donné par les lignes 3 et 4 ; ceci se fait alors au détriment du temps de calcul.

L'intégration de l'équation différentielle (10) à l'aide de l'algorithme de la méthode directe n'est pas en général possible ; en effet, l'algorithme diverge au bout de quelques pas à cause de la propagation du pessimisme dû au terme $\left[\mathbf{S}_j\right]\left([\mathbf{x}_j] - \hat{\mathbf{x}}_j\right)$. L'analyse suivante nous permet de bien comprendre ce problème (voir, [Rih94] [Ned99] pour plus de détails).

Etant donné un pavé initial $\left[\mathbf{x}(t_0)\right] = \left[\mathbf{x}_0\right]$, on obtient d'après (23) :

$$[\mathbf{x}_1] = \left[\mathbf{v}_1\right] + \left[\mathbf{S}_0\right]\left([\mathbf{x}_0] - \hat{\mathbf{x}}_0\right)$$

$$[\mathbf{x}_2] = \left[\mathbf{v}_2\right] + \left[\mathbf{S}_1\right]\left([\mathbf{x}_1] - \hat{\mathbf{x}}_1\right)$$

$$= \left[\mathbf{v}_2\right] + \left[\mathbf{S}_1\right]\left(\left([\mathbf{v}_1] - \hat{\mathbf{x}}_1\right) + \left[\mathbf{S}_0\right]\left([\mathbf{x}_0] - \hat{\mathbf{x}}_0\right)\right)$$

$$= \left[\mathbf{v}_2\right] + \left[\mathbf{S}_1\right]\left([\mathbf{v}_1] - \hat{\mathbf{x}}_1\right) + \left[\mathbf{S}_1\right]\left(\left[\mathbf{S}_0\right]\left([\mathbf{x}_0] - \hat{\mathbf{x}}_0\right)\right)$$

En procédant de la même manière pour le terme suivant :

$$\begin{aligned}[\mathbf{x}_{j+1}] &= \left[\mathbf{v}_{j+1}\right] + \left[\mathbf{S}_j\right]\left([\mathbf{x}_j] - \hat{\mathbf{x}}_j\right) \\ &= \left[\mathbf{v}_{j+1}\right] + \left[\mathbf{S}_j\right]\left([\mathbf{v}_j] - \hat{\mathbf{x}}_j\right) \\ &+ \left[\mathbf{S}_j\right]\left(\left[\mathbf{S}_{j-1}\right]\left([\mathbf{v}_{j-1}] - \hat{\mathbf{x}}_{j-1}\right)\right) \\ &+ \cdots \\ &+ \left[\mathbf{S}_j\right]\left(\left[\mathbf{S}_{j-1}\right]\cdots\left(\left[\mathbf{S}_0\right]\left([\mathbf{v}_0] - \hat{\mathbf{x}}_0\right)\right)\right)\end{aligned} \tag{24}$$

Analysons par exemple le dernier terme $\left[\mathbf{S}_j\right]\left(\left[\mathbf{S}_{j-1}\right]\cdots\left(\left[\mathbf{S}_0\right]\left([\mathbf{v}_0] - \hat{\mathbf{x}}_0\right)\right)\right)$:

- pour $j = 0$, on calcule $\left[\mathbf{S}_0\right]\left([\mathbf{x}_0] - \hat{\mathbf{x}}_0\right)$
- pour $j = 1$, le dernier terme est donné par $\left[\mathbf{S}_1\right]\left(\left[\mathbf{S}_0\right]\left([\mathbf{x}_0] - \hat{\mathbf{x}}_0\right)\right)$; pour l'évaluer, on calcule d'abord le produit $\left[\mathbf{S}_0\right]\left([\mathbf{x}_0] - \hat{\mathbf{x}}_0\right)$ puis on multiplie le résultat par la matrice $\left[\mathbf{S}_1\right]$. Ceci introduit alors un pessimisme compte tenu du fait que :

$$w\left(\left[\mathbf{S}_1\right]\left(\left[\mathbf{S}_0\right]\left([\mathbf{x}_0] - \hat{\mathbf{x}}_0\right)\right)\right) = \left|\left[\mathbf{S}_1\right]\right|\left|\left[\mathbf{S}_0\right]\right| w\left([\mathbf{x}_0] - \hat{\mathbf{x}}_0\right)$$

et comme

$$\left|\left[\mathbf{S}_1\right]\right| \cdot \left|\left[\mathbf{S}_0\right]\right| \geq \left|\left[\mathbf{S}_1\right]\left[\mathbf{S}_0\right]\right|$$

alors

$$w\left(\left[\mathbf{S}_1\right]\left(\left[\mathbf{S}_0\right]\left(\left[\mathbf{x}_0\right] - \hat{\mathbf{x}}_0\right)\right)\right) \geq \left\|\left[\mathbf{S}_1\right]\left[\mathbf{S}_0\right]\right\| w\left(\left[\mathbf{x}_0\right] - \hat{\mathbf{x}}_0\right)$$

$$= w\left(\left(\left[\mathbf{S}_1\right]\left[\mathbf{S}_0\right]\right)\left(\left[\mathbf{x}_0\right] - \hat{\mathbf{x}}_0\right)\right)$$

- en procédant de la même manière, on obtient

$$w\left(\left[\mathbf{S}_j\right]\left(\left[\mathbf{S}_{j-1}\right]\cdots\left(\left[\mathbf{S}_0\right]\left(\left[\mathbf{v}_0\right] - \hat{\mathbf{x}}_0\right)\right)\right)\right) \geq w\left(\left[\mathbf{S}_j\right]\left[\mathbf{S}_{j-1}\right]\left[\mathbf{S}_0\right]\left(\left[\mathbf{v}_0\right] - \hat{\mathbf{x}}_0\right)\right)$$

On constate alors qu'un pessimisme dû au phénomène d'enveloppement est introduit à chaque pas d'intégration. L'algorithme de la méthode directe diverge au bout de quelques itérations.

Dans la plupart des autres méthodes d'intégration basées sur (23), le pessimisme est contrôlé en prenant des précautions avant de calculer le terme $\left[\mathbf{S}_j\right]\left(\left[\mathbf{x}_j\right] - \hat{\mathbf{x}}_j\right)$.

4.3. Réduction du pessimisme

Supposons qu'on arrive à choisir un pas d'intégration h_j permettant d'avoir :

$$\left\|\left[\mathbf{S}_j\right]\right\| < 1$$

on obtient alors:

$$w\left(\left[\mathbf{S}_j\right]\left(\left[\mathbf{x}_j\right] - \hat{\mathbf{x}}_j\right)\right) = \left\|\left[\mathbf{S}_j\right]\right\| w\left(\left[\mathbf{x}_j\right] - \hat{\mathbf{x}}_j\right)$$

$$\leq w\left(\left[\mathbf{x}_j\right] - \hat{\mathbf{x}}_j\right)$$

La solution est alors contractée à chaque pas d'intégration. Néanmoins, il est rare de rencontrer une matrice $\left[\mathbf{S}_j\right]$ vérifiant $\left\|\left[\mathbf{S}_j\right]\right\|<1$. On verra dans les paragraphes suivant que d'autres méthodes permettent, dans plusieurs cas, de contrôler le pessimisme.

4.4. Méthode de Lohner

4.4.1. Principe de la méthode de Lohner

Lohner [Loh88] propose d'écrire la solution à l'instant t_{j+1} sous la forme suivante :

$$\mathbf{x}_{j+1} = \left\{ \hat{\mathbf{x}}_{j+1} + \mathbf{A}_{j+1}\mathbf{r}_{j+1} \mid \mathbf{r}_{j+1} \in \left[\mathbf{r}_{j+1}\right] \right\} \tag{25}$$

où $\hat{\mathbf{x}}_{j+1}$ est une approximation donnée par (22), $\mathbf{A}_{j+1}$ est une matrice ponctuelle et $[\mathbf{r}_{j+1}]$ est un pavé. Le bon choix de la matrice $\mathbf{A}_{j+1}$ assure le bon fonctionnement de la méthode de Lohner.

Considérons les mêmes notations données dans la méthode directe. Et soient :

$$\begin{cases} \left[\mathbf{z}_{j+1}\right] = h_j^k \mathbf{f}^{[k]}\left(\left[\tilde{\mathbf{x}}_j\right]\right) \\ \hat{\mathbf{s}}_{j+1} = m\left(\left[\mathbf{z}_{j+1}\right]\right) \end{cases} \tag{26}$$

on obtient alors :

$$\hat{\mathbf{x}}_{j+1} = \hat{\mathbf{x}}_j \; + \sum_{i=1}^{k-1} h_j^i \cdot \mathbf{f}^{[i]}\left(\hat{\mathbf{x}}_j\right) + \hat{\mathbf{s}}_{j+1} \tag{27}$$

A partir de (19), (20), (26) et (27) on obtient :

$$[\mathbf{x}_{j+1}] = \left(\hat{\mathbf{x}}_{j+1} - \hat{\mathbf{s}}_{j+1}\right) + \left[\mathbf{z}_{j+1}\right] + \left[\mathbf{S}_j\right]\left([\mathbf{x}_j] \; - \; \hat{\mathbf{x}}_j\right) \tag{28}$$

D'autre part, d'après (25) on pose :

$$[\mathbf{x}_j] - \hat{\mathbf{x}}_j = \mathbf{A}_j[\mathbf{r}_j] \tag{29}$$

En utilisant (29) et en réorganisant (28), on obtient :

$$[\mathbf{x}_{j+1}] = \hat{\mathbf{x}}_{j+1} + \left(\left[\mathbf{S}_j\right]\mathbf{A}_j\right)[\mathbf{r}_j] + \left[\mathbf{z}_{j+1}\right] - \mathbf{s}_{j+1} \tag{30}$$

On fait maintenant apparaître une matrice inversible $\mathbf{A}_{j+1}$, on a donc :

$$[\mathbf{x}_{j+1}] = \hat{\mathbf{x}}_{j+1} + \mathbf{A}_{j+1}\left(\mathbf{A}_{j+1}^{-1}\left(\left[\mathbf{S}_j\right]\mathbf{A}_j\right)[\mathbf{r}_j] + \mathbf{A}_{j+1}^{-1}\left(\left[\mathbf{z}_{j+1}\right] - \mathbf{s}_{j+1}\right)\right) \tag{31}$$

Notons $\left[\mathbf{r}_{j+1}\right]$ par :

$$\left[\mathbf{r}_{j+1}\right] = \left(\mathbf{A}_{j+1}^{-1}\left(\left[\mathbf{S}_j\right]\mathbf{A}_j\right)[\mathbf{r}_j] + \mathbf{A}_{j+1}^{-1}\left(\left[\mathbf{z}_{j+1}\right] - \mathbf{s}_{j+1}\right)\right) \tag{32}$$

On obtient alors :

$$\left[\mathbf{x}_{j+1}\right] = \hat{\mathbf{x}}_{j+1} + \mathbf{A}_{j+1}\left[\mathbf{r}_{j+1}\right] \tag{33}$$

Le bon choix de la matrice $\mathbf{A}_{j+1}$ nous permettra alors de bien contrôler le pessimisme. Cette méthode d'intégration suppose que la matrice $\mathbf{A}_{j+1}$ est inversible (dans le cas contraire, la méthode d'intégration échoue). Le choix de la matrice $\mathbf{A}_{j+1}$ sera discuté plus loin dans ce chapitre.

Le pavé $[\mathbf{x}_{j+1}]$ calculé à l'aide de (33) est une approximation extérieure de l'ensemble suivant :

$$\left\{\hat{\mathbf{x}}_{j+1} + \mathbf{A}_{j+1}\mathbf{r}_{j+1} \mid \mathbf{r}_{j+1} \in \left[\mathbf{r}_{j+1}\right]\right\} \tag{34}$$

Comme cet ensemble n'est pas un pavé, un pessimisme dû à l'effet d'enveloppement est alors introduit. Pour réduire cet effet, deux méthodes sont proposées dans la littérature [Loh88] [Ned99].

4.4.2. Méthode parallélépipédique

Considérons la solution de l'EDO (1) donnée par la formule (31). Etant donné que le pessimisme dû au terme $\mathbf{A}_{j+1}^{-1}\left(\left[\mathbf{z}_{j+1}\right]-\mathbf{s}_{j+1}\right)$ peut être facilement contrôlé en utilisant un développement de Taylor d'ordre élevé et en choisissant un pas d'intégration assez petit (sachant que $\left[\mathbf{z}_{j+1}\right]$ est proportionnel à $(1/k!)h_j^k$), on examine uniquement le pessimisme introduit par le terme $\mathbf{A}_{j+1}^{-1}\left(\left[\mathbf{S}_j\right]\mathbf{A}_j\right)[\mathbf{r}_j]$.

On pose :

$$\left[\mathbf{A}_{j+1}\right]=m\left(\left[\mathbf{S}_j\right]\mathbf{A}_j\right) \tag{35}$$

D'autre part, on peut décomposer la matrice $[\mathbf{S}_j]$ en deux termes :

$$\left[\mathbf{S}_j\right]=\hat{\mathbf{S}}_j+\left[\mathbf{E}_j\right] \tag{36}$$

on obtient alors :

$$\mathbf{A}_{j+1}=\hat{\mathbf{S}}_j\mathbf{A}_j \tag{37}$$

Par la suite on obtient :

$$\begin{aligned}\mathbf{A}_{j+1}^{-1}\left(\left[\mathbf{S}_j\right]\mathbf{A}_j\right)&=\mathbf{A}_j^{-1}\hat{\mathbf{S}}_j^{-1}\left(\hat{\mathbf{S}}_j\mathbf{A}_j+\left[\mathbf{E}_j\right]\mathbf{A}_j\right)\\&=\mathbf{I}+\mathbf{A}_j^{-1}\hat{\mathbf{S}}_j^{-1}\left[\mathbf{E}_j\right]\mathbf{A}_j\end{aligned} \tag{38}$$

Le terme considéré est presque égale à la matrice identité lorsque le dernier terme est proche de la matrice nulle. Ceci peut être le cas lorsque $\left\|\hat{\mathbf{S}}_j^{-1}[\mathbf{E}_j]\right\|\approx 0$ et lorsque la matrice $\mathbf{A}_j$ est bien conditionnée. Lorsque ces dernières conditions sont satisfaites, le pessimisme sur le terme $\mathbf{A}_{j+1}^{-1}\left(\left[\mathbf{S}_j\right]\mathbf{A}_j\right)[\mathbf{r}_j]$, qui représente la principale source de pessimisme dans (32), est assez réduit. Néanmoins, ces conditions ne sont pas assurées par le choix (35), ce qui peut alors conduire à des résultats très pessimistes.

La méthode parallélépipédique est très efficace lorsque la taille des éléments de la matrice $[\mathbf{S}_j]$ n'est pas grande.

4.4.3. Factorisation QR

Cette méthode a été proposée par Lohner [Loh88] qui l'a utilisé pour l'implémentation du solveur AWA [Loh94] ; elle est généralement très robuste. Le principe de cette méthode

consiste à représenter un ensemble quelconque dans une nouvelle base permettant de réduire l'effet d'enveloppement.

Soit $\tilde{\mathbf{A}}_{j+1}$ une matrice ponctuelle vérifiant :

$$\tilde{\mathbf{A}}_{j+1} \in \left[\mathbf{S}_j\right]\mathbf{A}_j$$

On note alors par

$$\hat{\mathbf{A}}_{j+1} = \tilde{\mathbf{A}}_{j+1}\mathbf{P}_{j+1}$$

où $\mathbf{P}_{j+1}$ est une matrice de permutation permettant de minimiser le pessimisme ; on peut consulter [Loh88] [Ned99] pour plus de détails sur le choix de cette matrice.

La matrice $\hat{\mathbf{A}}_{j+1}$ est alors factorisée à l'aide de la factorisation QR, on obtient :

$$\hat{\mathbf{A}}_{j+1} = \mathbf{Q}_{j+1}\mathbf{R}_{j+1}$$

où $\mathbf{Q}_{j+1}$ est une matrice orthogonale et $\mathbf{R}_{j+1}$ est une matrice triangulaire supérieure.

Dans la méthode de Lohner [Loh88], on prend alors :

$$\mathbf{A}_{j+1} = \mathbf{Q}_{j+1} \tag{39}$$

Cette méthode consiste donc à représenter l'ensemble $([\mathbf{S}_j]\mathbf{A}_j)[\mathbf{r}_j]$ dans une nouvelle base permettant de réduire l'effet d'enveloppement. Elle est particulièrement efficace, comme dans le cas de la méthode parallélépipédique, lorsque la taille des éléments de la matrice $[\mathbf{S}_j]$ n'est pas grande.

4.4.4. Algorithme de Lohner

L'algorithme de la méthode de Lohner est basé sur l'utilisation des expressions (32) et (33) ainsi que la factorisation QR.

Algorithme 3 : Méthode de Lohner

1. Entrées : $[\tilde{\mathbf{x}}_j]$, h_j, $[\mathbf{x}_j]$, $[\mathbf{r}_j]$, $\hat{\mathbf{x}}_j$, $\mathbf{A}_j$;
2. Calculer
3. $\left[\mathbf{z}_{j+1}\right] := h_j^k \mathbf{f}^{[k]}\left(\left[\tilde{\mathbf{x}}_j\right]\right)$
4. $\hat{\mathbf{s}}_{j+1} := m\left(\left[\mathbf{z}_{j+1}\right]\right)$

5. $\hat{\mathbf{x}}_{j+1} := \hat{\mathbf{x}}_j + \sum_{i=1}^{k-1} h_j^i \cdot \mathbf{f}^{[i]}\left(\hat{\mathbf{x}}_j\right) + \hat{\mathbf{s}}_{j+1}$

6. $\left[\mathbf{S}_j\right] := \left\{\mathbf{I} + \sum_{i=1}^{k-1} h_j^i \mathbf{J}\left(\mathbf{f}^{[i]};\ [\mathbf{x}_j]\right)\right\}$

7. Calculer $\mathbf{A}_{j+1}$ à l'aide de la factorisation QR

8. $[\mathbf{x}_{j+1}] := \hat{\mathbf{x}}_{j+1} + \left(\left[\mathbf{S}_j\right]\mathbf{A}_j\right)[\mathbf{r}_j] + \left[\mathbf{z}_{j+1}\right] - \mathbf{s}_{j+1}$

9. $\left[\mathbf{r}_{j+1}\right] := \left(\mathbf{A}_{j+1}^{-1}\left(\left[\mathbf{S}_j\right]\mathbf{A}_j\right)[\mathbf{r}_j] + \mathbf{A}_{j+1}^{-1}\left(\left[\mathbf{z}_{j+1}\right] - \mathbf{s}_{j+1}\right)\right)$

10. Sorties : $[\mathbf{x}_{j+1}]$, $\hat{\mathbf{x}}_{j+1}$, $\mathbf{A}_{j+1}$, $[\mathbf{r}_{j+1}]$.

4.5. Méthode de la valeur moyenne étendue

Cette méthode a été proposée par Rihm [Rih94], elle consiste à réduire le pessimisme de la méthode directe avec un pré-conditionnement matriciel. Pour simplifier la présentation de cette méthode, nous commençons d'abord par les premiers pas d'intégration.

En utilisant l'expression (23), on obtient :

$$\begin{aligned}[\mathbf{x}_1] &= \left[\mathbf{v}_1\right] + \left[\mathbf{S}_0\right]\left([\mathbf{x}_0] - \hat{\mathbf{x}}_0\right) \\ &= \left[\mathbf{v}_1\right] + \left[\mathbf{q}_1\right]\end{aligned}$$

avec :

$$\begin{aligned}\left[\mathbf{q}_1\right] &= \left[\mathbf{S}_0\right]\left([\mathbf{x}_0] - \hat{\mathbf{x}}_0\right) \\ &= \mathbf{A}_1\left(\left(\mathbf{A}_1^{-1}\left[\mathbf{S}_0\right]\right)\left([\mathbf{x}_0] - \hat{\mathbf{x}}_0\right)\right) \\ &= \mathbf{A}_1\left[\mathbf{p}_1\right]\end{aligned}$$

où $\mathbf{A}_1$ est une matrice non singulière et

$$\left[\mathbf{p}_1\right] = \left(\left(\mathbf{A}_1^{-1}\left[\mathbf{S}_0\right]\right)\left([\mathbf{x}_0] - \hat{\mathbf{x}}_0\right)\right)$$

Comme pour la méthode de Lohner, l'introduction de la matrice $\mathbf{A}_1^{-1}$ permet de réduire l'effet d'enveloppement dans le produit $\left[\mathbf{S}_0\right]\left([\mathbf{x}_0] - \hat{\mathbf{x}}_0\right)$. Néanmoins, ceci dépend de plusieurs facteurs i.e. la matrice **A** doit être inversible et bien conditionnée et la taille des éléments de la matrice $\left[\mathbf{S}_0\right]$ assez petite.

De même, une approximation de la solution à l'instant t_2 est donnée par :

$$[\mathbf{x}_2] = [\mathbf{v}_2] + [\mathbf{S}_1]([\mathbf{x}_1] - \hat{\mathbf{x}}_1)$$

Etant donné que :

$$[\mathbf{x}_1] = [\mathbf{v}_1] + [\mathbf{q}_1]$$

alors :

$$[\mathbf{x}_2] = [\mathbf{v}_2] + ([\mathbf{S}_1]\mathbf{A}_1)[\mathbf{p}_1] + [\mathbf{S}_1]([\mathbf{v}_1] - \hat{\mathbf{x}}_1)$$

Introduisons maintenant une matrice $\mathbf{A}_2$ supposée non singulière et bien conditionnée, alors :

$$[\mathbf{x}_2] = [\mathbf{v}_2] + \mathbf{A}_2\left(\left(\mathbf{A}_2^{-1}([\mathbf{S}_1]\mathbf{A}_1)\right)[\mathbf{p}_1] + \mathbf{A}_2^{-1}[\mathbf{S}_1]([\mathbf{v}_1] - \hat{\mathbf{x}}_1)\right)$$

Notons par :

$$[\mathbf{p}_2] = \left(\mathbf{A}_2^{-1}([\mathbf{S}_1]\mathbf{A}_1)\right)[\mathbf{p}_1] + \mathbf{A}_2^{-1}[\mathbf{S}_1]([\mathbf{v}_1] - \hat{\mathbf{x}}_1)$$

on obtient finalement :

$$[\mathbf{x}_2] = [\mathbf{v}_2] + \mathbf{A}_2[\mathbf{p}_2]$$

$$= [\mathbf{v}_2] + [\mathbf{q}_2]$$

En procédant de la même manière, on obtient à l'instant t_{j+1} :

$$[\mathbf{q}_{j+1}] = ([\mathbf{S}_j]\mathbf{A}_j)[\mathbf{p}_j] + [\mathbf{S}_j]([\mathbf{v}_j] - \hat{\mathbf{x}}_j)$$

et

$$[\mathbf{p}_{j+1}] = \left(\left(\mathbf{A}_{j+1}^{-1}([\mathbf{S}_j]\mathbf{A}_j)\right)[\mathbf{p}_j] + \mathbf{A}_{j+1}^{-1}[\mathbf{S}_j]([\mathbf{v}_j] - \hat{\mathbf{x}}_j)\right)$$

La suite des matrices $\mathbf{A}_j$ peut être calculée en utilisant la méthode QR présentée ci-dessus, néanmoins, on ne peut pas garantir que ces matrices soient toutes inversibles et bien conditionnées.

L'algorithme de la méthode de la valeur moyenne étendue est résumé dans l'algorithme 4.

L'algorithme est initialisé avec $[\mathbf{p}_j] = \mathbf{0}$, $\mathbf{A}_1 = \mathbf{I}$ et la solution *a priori* $[\tilde{\mathbf{x}}_j]$ est calculée à l'aide de l'algorithme 1.

Rihm [Rih1994] affirme que la méthode de la valeur moyenne étendue produit des résultats moins pessimistes par rapport à la méthode de Lohner [Loh88].

Algorithme 4 : Méthode de la valeur moyenne étendue

1. Entrées: $[\mathbf{x}_j]$, $[\tilde{\mathbf{x}}_j]$, $[\mathbf{p}_j]$, h_j, $\hat{\mathbf{x}}_j$, $\mathbf{A}_j$
2. Calculer :
3. $$[\mathbf{v}_{j+1}] = \hat{\mathbf{x}}_j + \sum_{i=1}^{k-1} \mathbf{f}^{[i]}(\hat{\mathbf{x}}_j)h^i + \mathbf{f}^{[k]}([\tilde{\mathbf{x}}_j])h^k$$
4. $$[\mathbf{S}_j] = \mathbf{I} + \sum_{i=1}^{k-1} \mathbf{J}(\mathbf{f}^{[i]};[\mathbf{x}_j])h^i$$
5. $$[\mathbf{q}_{j+1}] = ([\mathbf{S}_j]\mathbf{A}_j)[\mathbf{p}_j] + [\mathbf{S}_j]([\mathbf{v}_j] - \hat{\mathbf{x}}_j)$$
6. $$[\mathbf{x}_{j+1}] = [\mathbf{v}_{j+1}] + [\mathbf{q}_{j+1}]$$
7. $$[\mathbf{p}_{j+1}] = \mathbf{A}_{j+1}^{-1}([\mathbf{S}_j]\mathbf{A}_j)[\mathbf{p}_j] + (\mathbf{A}_{j+1}^{-1}[\mathbf{S}_j])([\mathbf{v}_j] - \hat{\mathbf{x}}_j)$$
8. $\mathbf{A}_{j+1}$ (à l'aide de la factorisation QR)
9. Sorties: $[\mathbf{x}_{j+1}]$, $[\mathbf{p}_{j+1}]$, $\mathbf{A}_{j+1}$

4.6. Méthode d'Hermite-Obreschkoff

Cette méthode a été étendue du cas ponctuel aux intervalles par Nedialkov [NJ99]. Elle est généralement plus efficace que les méthodes basées sur un développement de Taylor explicite. Pour résoudre une équation différentielle ordinaire avec la méthode d'Hermite-Obreschkoff, deux étapes sont nécessaires : la première, appelée prédiction, utilise un développement de Taylor et la deuxième, dite correction, consiste à réduire la solution obtenue dans la phase de prédiction ; un contracteur est alors utilisé.

4.6.1. Cas ponctuel

Dans cette section nous donnons un petit succinct du schéma classique de la méthode d'Hermite-Obreschkoff (cas ponctuel et sans garantie).

On considère alors l'équation différentielle (2) et notons par :

$$\begin{cases} P_{p,q}(s) = \dfrac{s^q (s-1)^p}{(p+q)!} \\ c_i^{p,q} = \dfrac{q!(p+q-i)!}{(p+q)!(q-i)!} \\ \mathbf{g}_i(s) = \dfrac{\mathbf{g}^{(i)}(s)}{i!} \end{cases} \tag{40}$$

où $s \in \mathbb{R}$, $p, q, i \in \mathbb{N}$, $0 \le i \le q$ et g est une fonction au moins $(p+q+1)$ fois différentiable.

Les intégrations par parties successives de l'intégrale $\int_0^1 P_{p,q}(s)\mathbf{g}^{(p+q+1)}(s)ds$, obtenues par [Dar1876] et [Her1878], donnent :

$$(-1)^{(p+q)}\int_0^1 P_{p,q}(s)\mathbf{g}^{(p+q+1)}(s)ds = \sum_{i=0}^{q}(-1)^i c_i^{q,p}\mathbf{g}_i(1) - \sum_{i=0}^{p} c_i^{p,q}\mathbf{g}_i(0) \tag{41}$$

Notons par **x**(*t*) la solution de l'EDO (2) et par :

$$\mathbf{g}(s) = \mathbf{x}\left(t_j + sh_j\right) \tag{42}$$

g(*s*) est alors la solution de l'EDO (2) à l'instant ($t_j + sh_j$) étant donné $\mathbf{x}(t_j)$ solution à t_j. On obtient alors :

$$\begin{cases} \mathbf{g}_i(0) = \dfrac{\mathbf{g}^{(i)}(0)}{i!} = h_j^i \dfrac{\mathbf{x}^{(i)}(t_j)}{i!} = h_j^i \mathbf{f}^{[i]}(\mathbf{x}_j) \\ \mathbf{g}_i(1) = \dfrac{\mathbf{g}^{(i)}(1)}{i!} = h_j^i \dfrac{\mathbf{x}^{(i)}(t_j + h_j)}{i!} = h_j^i \mathbf{f}^{[i]}(\mathbf{x}_{j+1}) \\ \mathbf{g}^{(p+q+1)}(s) = h_j^{p+q+1}\mathbf{x}^{(p+q+1)}(t_j + sh_j) \end{cases} \tag{43}$$

où $\mathbf{f}^{[i]}$ est le $i^{ème}$ coefficient de Taylor de la fonction $\mathbf{x}(t)$ solution de (2).

En utilisant la notation (43), on a :

$$\begin{aligned} (-1)^{p+q}\int_0^1 P_{p,q}(s)\mathbf{g}^{(p+q+1)}(s)ds &= (-1)^{p+q}h_j^{p+q+1}\int_0^1 P_{p,q}(s)\mathbf{x}^{(p+q+1)}(t_j + sh_j)ds \\ &= (-1)^q \frac{q!\,p!}{(p+q)!}h_j^{p+q+1}\frac{\mathbf{x}^{(p+q+1)}(\xi_l)}{(p+q+1)!} \end{aligned} \tag{44}$$

où le $l^{ème}$ élément de $\mathbf{x}^{(p+q+1)}(\xi_l)$ est évalué à un instant $\xi_l \in [t_j, t_{j+1}]$.

Les expressions (41) et (44) donnent :

$$\sum_{i=0}^{q}(-1)^i c_i^{q,p}h_j^i\mathbf{f}^{[i]}(\mathbf{x}_{j+1}) = \sum_{i=0}^{p} c_i^{p,q}h_j^i\mathbf{f}^{[i]}(\mathbf{x}_j) \;+\; (-1)^q\frac{q!\,p!}{(p+q)!}h_j^{p+q+1}\frac{\mathbf{x}^{(p+q+1)}(\xi_l)}{(p+q+1)!} \tag{45}$$

Si on suppose que la solution de l'EDO (2) à l'instant t_j est connue, alors la résolution du système d'équations (45) permet de trouver la solution de (2) à t_{j+1}. La méthode d'intégration ponctuelle d'Hermite-Obreschkoff [GCHK97] [Wan77] est basée sur la résolution de (45) où le dernier terme correspond à l'erreur de troncature. Ce schéma a été étendu dans [Ned99] au cas où la solution à t_j appartient à un intervalle.

4.6.2. Extension aux intervalles

L'extension de la méthode d'Hermite-Obreschkoff aux intervalles consiste d'abord à prédire la solution à l'instant t_{j+1} en utilisant une des méthodes basées sur un développement de Taylor. L'approximation extérieure obtenue est ensuite contractée en résolvant le système d'équations (45). Dans la suite de cette section nous allons décrire cette procédure.

Soit $[\mathbf{x}_j]$ la solution de l'EDO (10) à l'instant t_j, et notons par $[\mathbf{x}^0_{j+1}]$ une approximation extérieure de la solution de l' EDO à l'instant t_{j+1} calculée par exemple à l'aide de la méthode Lohner. La phase de correction consiste alors à calculer un pavé $[\mathbf{x}_{j+1}] \subseteq [\mathbf{x}^0_{j+1}]$ contenant la solution de l' EDO (10) étant donné une condition initiale $\mathbf{x}_j \in [\mathbf{x}_j]$. Cette étape est réalisée à l'aide d'un contracteur de Newton.

En évaluant les coefficients de Taylor dans l'expression (45) à l'aide de la forme centrée, on obtient :

$$\begin{cases} \mathbf{f}^{[i]}(\mathbf{x}_j) = \mathbf{f}^{[i]}(\hat{\mathbf{x}}_j) + \mathbf{J}\left(\mathbf{f}^{[i]}, \mathbf{x}_j, \hat{\mathbf{x}}_j\right)\left(\mathbf{x}_j - \hat{\mathbf{x}}_j\right) \\ \mathbf{f}^{[i]}(\mathbf{x}_{j+1}) = \mathbf{f}^{[i]}(\hat{\mathbf{x}}^0_{j+1}) + \mathbf{J}\left(\mathbf{f}^{[i]}, \mathbf{x}_{j+1}, \hat{\mathbf{x}}^0_{j+1}\right)\left(\mathbf{x}_{j+1} - \hat{\mathbf{x}}^0_{j+1}\right) \end{cases} \tag{46}$$

avec $\hat{\mathbf{x}}_j \in [\mathbf{x}_j]$, $\hat{\mathbf{x}}^0_{j+1} \in [\mathbf{x}^0_{j+1}]$, $\mathbf{J}\left(\mathbf{f}^{[i]}, \mathbf{x}_j, \hat{\mathbf{x}}_j\right)$ est le Jacobien du coefficient de Taylor $\mathbf{f}^{[i]}$ dont la ligne *i* est évaluée à partir de $\mathbf{x}_j + \theta_i(\hat{\mathbf{x}}_j - \mathbf{x}_j)$ pour $\theta_i \in [0, 1]$.

$$\sum_{i=0}^{q}(-1)^i c_i^{q,p} h_j^i \mathbf{f}^{[i]}(\hat{\mathbf{x}}^0_{j+1}) + \left(\sum_{i=0}^{q}(-1)^i c_i^{q,p} h_j^i \mathbf{J}\left(\mathbf{f}^{[i]}, \mathbf{x}_{j+1}, \hat{\mathbf{x}}^0_{j+1}\right)\right)\left(\mathbf{x}_{j+1} - \hat{\mathbf{x}}^0_{j+1}\right)$$

$$= \sum_{i=0}^{q} c_i^{p,q} h_j^i \mathbf{f}^{[i]}(\hat{\mathbf{x}}_j) + \left(\sum_{i=0}^{p} c_i^{p,q} h_j^i \mathbf{J}\left(\mathbf{f}^{[i]}, \mathbf{x}_j, \hat{\mathbf{x}}_j\right)\right)\left(\mathbf{x}_j - \hat{\mathbf{x}}_j\right) \tag{47}$$

$$+ (-1)^q \frac{q!\,p!}{(p+q)!} h_j^{p+q+1} \frac{\mathbf{x}^{(p+q+1)}(\xi_l)}{(p+q+1)!}$$

Supposons maintenant que $\mathbf{x}_j \in [\mathbf{x}_j]$ et qu'une approximation extérieure $[\mathbf{x}^0_{j+1}]$de $[\mathbf{x}_{j+1}]$ ait été calculée à l'aide d'une méthode d'intégration de Taylor. L'expression (47) peut être réécrite sous la forme suivante :

$$\left[\mathbf{S}_{j+1}^{-}\right]\left(\left[\mathbf{x}_{j+1}\right]-\hat{\mathbf{x}}_{j+1}^{0}\right)=\left[\mathbf{S}_{j}^{+}\right]\left(\left[\mathbf{x}_{j}\right]-\hat{\mathbf{x}}_{j}\right)+\left[\boldsymbol{\delta}_{j+1}\right] \tag{48}$$

où :

$$\begin{cases}\left[\mathbf{S}_{j+1}^{-}\right] = \left(\sum_{i=0}^{q}(-1)^{i} c_i^{q,p} h_j^i \mathbf{J}\left(\mathbf{f}^{[i]}, \left[\mathbf{x}_{j+1}^{0}\right]\right)\right) \\ \hat{\mathbf{S}}_{j+1}^{-} = m\left(\left[\mathbf{S}_{j+1}^{-}\right]\right) \\ \left[\mathbf{S}_{j}^{+}\right] = \left(\sum_{i=0}^{p} c_i^{p,q} h_j^i \mathbf{J}\left(\mathbf{f}^{[i]}, \left[\mathbf{x}_{j}\right]\right)\right) \\ \left[\boldsymbol{\varepsilon}_{j+1}\right] = (-1)^{q}\frac{q!\,p!}{(p+q)!} h_j^{p+q+1}\mathbf{f}^{[p+q+1]}\left(\left[\tilde{\mathbf{x}}_{j}\right]\right) \\ \mathbf{g}_{j+1} = \sum_{i=0}^{p} c_i^{p,q} h_j^i \mathbf{f}^{[i]}(\hat{\mathbf{x}}_{j}) - \sum_{i=0}^{q}(-1)^{i} c_i^{q,p} h_j^i \mathbf{f}^{[i]}(\hat{\mathbf{x}}_{j+1}^{0}) \\ \left[\boldsymbol{\delta}_{j+1}\right] = \mathbf{g}_{j+1} + \left[\boldsymbol{\varepsilon}_{j+1}\right]\end{cases} \tag{49}$$

Pour déduire l'expression de $[\mathbf{x}_{j+1}]$ à partir de (48), on doit alors inverser la matrice $[\mathbf{S}_{j+1}^{-}]$; ceci est numériquement coûteux. On montrera dans la suite qu'il est possible d'éviter cette inversion [Ned99].

En introduisant dans l'expression (48) la matrice $\hat{\mathbf{S}}_{j+1}^{-}$ définie dans (49), on obtient :

$$\hat{\mathbf{S}}_{j+1}^{-}\left(\left[\mathbf{x}_{j+1}\right]-\hat{\mathbf{x}}_{j+1}^{0}\right)=\left[\mathbf{S}_{j}^{+}\right]\left(\left[\mathbf{x}_{j}\right]-\hat{\mathbf{x}}_{j}\right)-\left(\left[\mathbf{S}_{j+1}^{-}\right]-\hat{\mathbf{S}}_{j+1}^{-}\right)\left(\left[\mathbf{x}_{j+1}\right]-\hat{\mathbf{x}}_{j+1}^{0}\right)+\left[\boldsymbol{\delta}_{j+1}\right] \tag{50}$$

On suppose que la matrice $\hat{\mathbf{S}}_{j+1}^{-}$ est inversible alors on obtient :

$$\begin{aligned}\left[\mathbf{x}_{j+1}\right]-\hat{\mathbf{x}}_{j+1}^{0} = &\left(\left(\hat{\mathbf{S}}_{j+1}^{-}\right)^{-1}\left[\mathbf{S}_{j}^{+}\right]\right)\left(\left[\mathbf{x}_{j}\right]-\hat{\mathbf{x}}_{j}\right)+\left(\mathbf{I}-\left(\hat{\mathbf{S}}_{j+1}^{-}\right)^{-1}\left[\mathbf{S}_{j+1}^{-}\right]\right)\left(\left[\mathbf{x}_{j+1}\right]-\hat{\mathbf{x}}_{j+1}^{0}\right) \\ &+\left(\hat{\mathbf{S}}_{j+1}^{-}\right)^{-1}\left[\boldsymbol{\delta}_{j+1}\right]\end{aligned} \tag{51}$$

On remarque que $[\mathbf{x}_{j+1}]$ apparaît à la fois dans les termes de gauche et de droite dans (51). D'autre part, une approximation extérieure $[\mathbf{x}_{j+1}^{0}]$ de $[\mathbf{x}_{j+1}]$ a été préalablement calculée à l'aide d'une des méthodes de Taylor. On remplace dans les termes de droite de (51) $[\mathbf{x}_{j+1}]$ par l'encadrement a priori $[\mathbf{x}_{j+1}^{0}]$ déjà calculé, on obtient alors une expression explicite de $[\mathbf{x}_{j+1}]$:

$$\left[\mathbf{x}_{j+1}\right] - \hat{\mathbf{x}}^0_{j+1} \equiv \left(\left(\hat{\mathbf{S}}^-_{j+1}\right)^{-1}\left[\mathbf{S}^+_j\right]\right)\left(\left[\mathbf{x}_j\right] - \hat{\mathbf{x}}_j\right) + \left(\mathbf{I} - \left(\hat{\mathbf{S}}^-_{j+1}\right)^{-1}\left[\mathbf{S}^-_{j+1}\right]\right)\left(\left[\mathbf{x}^0_{j+1}\right] - \hat{\mathbf{x}}^0_{j+1}\right) + \left(\hat{\mathbf{S}}^-_{j+1}\right)^{-1}\left[\boldsymbol{\delta}_{j+1}\right] \tag{52}$$

Rappelons que dans la méthode de Lohner, le pavé solution à l'instant t_{j+1} est décrit par la forme suivante:

$$\left[\mathbf{x}_{j+1}\right] = \left[\hat{\mathbf{x}}_{j+1}\right] + \mathbf{A}_{j+1}\left[\mathbf{r}_{j+1}\right]$$

En utilisant cette description dans (52), on obtient :

$$\left[\mathbf{x}_{j+1}\right] - \hat{\mathbf{x}}^0_{j+1} = \left(\left(\hat{\mathbf{S}}^-_{j+1}\right)^{-1}\left[\mathbf{S}^+_j\right]\mathbf{A}_j\right)\left[\mathbf{r}_j\right] + \left(\mathbf{I} - \left(\hat{\mathbf{S}}^-_{j+1}\right)^{-1}\left[\mathbf{S}^-_{j+1}\right]\right)\left(\left[\mathbf{x}^0_{j+1}\right] - \hat{\mathbf{x}}^0_{j+1}\right) + \left(\hat{\mathbf{S}}^-_{j+1}\right)^{-1}\left[\boldsymbol{\delta}_{j+1}\right] \tag{53}$$

Utilisons maintenant les notations suivantes afin de simplifier l'expression (53) :

$$\begin{cases} \left[\mathbf{B}_j\right] = \left(\hat{\mathbf{S}}^-_{j+1}\right)^{-1}\left[\mathbf{S}^+_j\right]\mathbf{A}_j \\ \left[\mathbf{C}_j\right] = \mathbf{I} - \left(\hat{\mathbf{S}}^-_{j+1}\right)^{-1}\left[\mathbf{S}^-_{j+1}\right] \\ \left[\mathbf{v}_j\right] = \left[\mathbf{x}^0_{j+1}\right] - \hat{\mathbf{x}}^0_{j+1} \end{cases} \tag{54}$$

On obtient alors :

$$\left[\mathbf{x}_{j+1}\right] = \hat{\mathbf{x}}^0_{j+1} + \left[\mathbf{B}_j\right]\left[\mathbf{r}_j\right] + \left[\mathbf{C}_j\right]\left[\mathbf{v}_j\right] + \left(\hat{\mathbf{S}}^-_{j+1}\right)^{-1}\left[\boldsymbol{\delta}_{j+1}\right] \tag{55}$$

L'expression (55) permet de calculer une approximation extérieure de la solution de l'EDO (10) à t_{j+1} lorsque la solution à t_j vérifie $\mathbf{x}_j \in [\mathbf{x}_j]$. L'approximation extérieure de l'ensemble solution est alors donnée par :

$$\left[\mathbf{x}_{j+1}\right] = \left(\hat{\mathbf{x}}^0_{j+1} + \left[\mathbf{B}_j\right]\left[\mathbf{r}_j\right] + \left[\mathbf{C}_j\right]\left[\mathbf{v}_j\right] + \left(\hat{\mathbf{S}}^-_{j+1}\right)^{-1}\left[\boldsymbol{\delta}_{j+1}\right]\right) \cap \left[\mathbf{x}^0_{j+1}\right] \tag{56}$$

L'expression (56) définit alors un contracteur semblable à celui de Newton. On doit noter que ce contracteur ne produit pas une contraction optimale étant donné que les coefficients de Taylor contiennent des variables multioccurrentes. Pour améliorer la contraction on peut envisager d'appliquer ce contracteur tant qu'une amélioration est possible, ceci a bien sûr des conséquences négatives au niveau du temps de calcul.

Chapitre 4

Algorithme 5 : Méthode d'Hermite-Obreschkoff

Entrées : $[\mathbf{x}_j]$, $[\tilde{\mathbf{x}}_j]$, $[\mathbf{r}_j]$, h_j, $\hat{\mathbf{x}}_j$, $\mathbf{A}_j$

Calculer :

1. $\left[\mathbf{x}_{j+1}^0\right]$ à l'aide de la méthode de Lohner par exemple

2. $\hat{\mathbf{x}}_{j+1}^0 := m\left(\left[\mathbf{x}_{j+1}^0\right]\right)$

3. $\mathbf{g}_{j+1} := \sum_{i=0}^{p} c_i^{p,q} h_j^i \mathbf{f}^{[i]}(\hat{\mathbf{x}}_j) - \sum_{i=0}^{q} (-1)^i c_i^{q,p} h_j^i \mathbf{f}^{[i]}(\hat{\mathbf{x}}_{j+1}^0)$

4. $\left[\mathbf{S}_j^+\right] := \left(\sum_{i=0}^{p} c_i^{p,q} h_j^i \mathbf{J}\left(\mathbf{f}^{[i]}, \left[\mathbf{x}_j\right]\right)\right)$

5. $\left[\mathbf{S}_{j+1}^-\right] := \left(\sum_{i=0}^{q} (-1)^i c_i^{q,p} h_j^i \mathbf{J}\left(\mathbf{f}^{[i]}, \left[\mathbf{x}_{j+1}^0\right]\right)\right)$

6. $\hat{\mathbf{S}}_{j+1}^- := m\left(\left[\mathbf{S}_{j+1}^-\right]\right)$

7. $\left[\mathbf{B}_j\right] := \left(\hat{\mathbf{S}}_{j+1}^-\right)^{-1}\left[\mathbf{S}_j^+\right]\mathbf{A}_j$

8. $\left[\mathbf{C}_j\right] := \mathbf{I} - \left(\hat{\mathbf{S}}_{j+1}^-\right)^{-1}\left[\mathbf{S}_{j+1}^-\right]$

9. $\left[\mathbf{v}_j\right] := \left[\mathbf{x}_{j+1}^0\right] - \hat{\mathbf{x}}_{j+1}^0$

10. $\left[\mathbf{x}_{j+1}\right] := \left(\hat{\mathbf{x}}_{j+1}^0 + \left[\mathbf{B}_j\right]\left[\mathbf{r}_j\right] + \left[\mathbf{C}_j\right]\left[\mathbf{v}_j\right] + \left(\hat{\mathbf{S}}_{j+1}^-\right)^{-1}\left[\boldsymbol{\delta}_{j+1}\right]\right) \cap \left[\mathbf{x}_{j+1}^0\right]$

11. $\hat{\mathbf{A}}_{j+1} := m\left(\left[\mathbf{B}_j\right]\right)$

12. $\mathbf{A}_{j+1}$: A l'aide d'une factorisation QR

13. $\hat{\mathbf{x}}_{j+1} := m\left(\left[\mathbf{x}_{j+1}\right]\right)$

14. $\left[\mathbf{r}_{j+1}\right] := \left(\mathbf{A}_{j+1}^{-1}\left[\mathbf{B}_j\right]\right)\left[\mathbf{r}_j\right] + \left(\mathbf{A}_{j+1}^{-1}\left[\mathbf{C}_j\right]\right)\left[\mathbf{v}_j\right] + \left(\mathbf{A}_{j+1}^{-1}\left(\hat{\mathbf{S}}_{j+1}^-\right)^{-1}\right)\left[\boldsymbol{\delta}_{j+1}\right] + \mathbf{A}_{j+1}^{-1}\left(\hat{\mathbf{x}}_{j+1}^0 - \hat{\mathbf{x}}_{j+1}\right)$

Sorties : $[\mathbf{x}_{j+1}]$, $[\mathbf{r}_{j+1}]$, $\mathbf{A}_{j+1}$, $\hat{\mathbf{x}}_{j+1}$

Dans cet algorithme, la phase de prédiction n'a pas été détaillée étant donné qu'elle est basée sur une des méthodes de Taylor présentées dans les sections précédentes. L'utilisation de la méthode d'Hermite-Obreschkoff n'apporte pas réellement de complexité supplémentaire par rapport aux schémas de Taylor. En effet, les coefficients de Taylor ainsi que l'inverse de la matrice $\mathbf{A}_j$ sont calculés pour toutes les méthodes.

5. Conclusion

Ce chapitre a été consacré aux méthodes d'intégration numérique garantie des équations différentielles. Ces méthodes permettent de vérifier l'existence et l'unicité de la solution de l'EDO considérée et fournissent, à chaque instant, un pavé contenant d'une manière garantie la solution de l'EDO.

La majorité des techniques d'intégration garantie comportent deux étapes ; la première consiste à prouver, à l'aide du théorème du point fixe, l'existence et l'unicité de la solution à chaque pas d'intégration. Un pavé, appelé aussi solution *a priori*, contenant toute la trajectoire de la solution entre deux pas successifs ainsi que le pas d'intégration sont calculés. La deuxième phase consiste à contracter ce pavé. Cette contraction est effectuée à l'aide d'un développement de Taylor d'ordre élevé dont les coefficients sont évalués en utilisant la forme centrée.

La dernière section de ce chapitre a été consacrée à la méthode d'Hermite-Obreschkoff qui génère, en général, une approximation moins pessimiste que celle calculée à l'aide des méthodes de Taylor ; ceci est obtenu grâce à l'utilisation d'un contracteur.

Ces outils d'intégration numérique garantie d'équations différentielles seront utilisés dans le chapitre suivant afin de développer des observateurs ensemblistes pour des systèmes non linéaires à temps continu.

Chapitre 5

Estimation d'état et de paramètres de systèmes non linéaires à temps continu

1. Introduction

Pour élaborer une loi de commande ou pour prendre une décision, il est souvent nécessaire de connaître l'évolution des variables d'état du système considéré. Néanmoins, dans la plupart des applications réelles, ces variables ne sont pas directement accessibles. Dans ce cas, on a recours à l'estimation de l'évolution de l'état à partir des mesures de l'entrée et de la sortie du système. Ce processus d'estimation nécessite la construction d'un modèle permettant de reproduire au mieux le comportement du système réel.

Lorsque le modèle est linéaire et représente exactement le comportement du système et lorsque les données mesurées ne sont pas entachées de bruit, il est alors possible de reconstruire l'état du processus en utilisant l'observateur de Luenberger [Lue71]. Cet estimateur est souvent utilisé lorsqu'on ne possède pas d'informations satisfaisantes sur les perturbations ou lorsque l'effet de ces dernières est négligeable. Néanmoins, étant donné qu'un modèle n'est qu'une approximation du processus réel, l'état estimé par l'observateur de Luenberger est souvent biaisé. Le filtrage de Kalman [Kal60] représente une alternative à l'observateur de Luenberger, et produit des estimations plus précises lorsque le modèle est linéaire. L'importance du filtrage de Kalman vient du fait que sous certaines hypothèses statistiques, l'estimée calculée est optimale. Cet observateur a été étendu au cas non linéaire en utilisant des linéarisations. Dans ce dernier cas et lorsque le modèle est fortement non linéaire, l'estimée obtenue peut être éloignée de la vraie solution.

Comme nous l'avons noté dans le paragraphe précédent, le filtre de Kalman produit des résultats précis lorsque des distributions statistiques des perturbations sont disponibles. Dans certaines applications, le bruit ne peut pas être décrit par une loi de probabilité (on ne dispose pas d'assez de données par exemple), dans ce cas il est plus judicieux de considérer que ces perturbations sont bornées et de bornes connues. L'estimation dans un contexte à erreurs bornées constitue alors une alternative intéressante aux méthodes statistiques. Le problème est donc de caractériser à chaque instant, d'une manière garantie, toutes les valeurs du vecteur d'état compatibles avec les mesures et avec les bornes d'erreurs supposées connues.

Le problème d'estimation dans un contexte à erreurs bornées a été souvent traité pour des modèles linéaires (voir par exemple [Sch68] [DPW96] [PNDW04] [MN02] [DWP01]). Lorsque le modèle est linéaire, l'ensemble des solutions compatibles avec les mesures et avec les bornes d'erreurs est un polyèdre convexe qui peut être exactement décrit lorsque la dimension du vecteur d'état est réduite. En pratique, cette caractérisation exacte est un problème complexe et ainsi d'autres approximations extérieures utilisant des formes géométriques simples, i.e. des ellipsoïdes, des parallèlotopes ou autres, sont calculées [FH82] [BBC90] [CGZ96] [DPW96] [CGVZ98] [BH99] [DWP01] [LB02] [MN02] [Bec03].

Lorsque le modèle est non linéaire, ces dernières méthodes ne sont plus utilisables ; d'autres estimateurs basés sur l'analyse par intervalles et la propagation de contraintes ont été développés pour des systèmes à temps discret [JKDW01] [KJW02] [FA03]. On trouve aussi dans la littérature une extension du filtre de Kalman aux intervalles [CWS97] [BRG01].

Les techniques présentées ci-dessus sont applicables lorsque le modèle est à temps discret. Néanmoins, les systèmes réels sont souvent décrits par des équations différentielles non linéaires (EDOs). Ce problème a été résolu pour une certaine classe de systèmes à temps continu [GRH00] [HG01]. Mais dans le cas général, ce problème n'a jamais été abordé avant les travaux de Jaulin [Jau02], où un estimateur basé sur l'approche prédiction/correction a été proposé. Dans ce chapitre nous proposons une amélioration de cet estimateur afin de l'adapter à des cas où la dimension du vecteur d'état est élevée. Cet observateur, développé dans la section 3 de ce chapitre, est basé sur les méthodes d'intégration numériques garanties d'EDOs présentées dans le chapitre 4.

Dans la dernière partie de la section 3, nous proposons un observateur à horizon glissant basé sur l'estimateur prédicteur/correcteur et des techniques de propagation de contraintes.

Enfin, dans la dernière partie de ce chapitre, l'observateur présenté sera étendu au cas où l'on cherche à estimer des paramètres inconnus dans le modèle.

2. Estimation d'état pour des systèmes non linéaires à temps discret

On considère le système suivant :

$$\begin{cases} \mathbf{x}_{k+1} = \mathbf{f}(\mathbf{x}_k, \mathbf{u}_k, \mathbf{w}_k) \\ \mathbf{y}_k = \mathbf{g}(\mathbf{x}_k, \mathbf{u}_k, \mathbf{v}_k) \\ \mathbf{x}_0 \in [\mathbf{x}_0] \end{cases} \quad (1)$$

où $\mathbf{x}_k \in \mathbb{R}^n$, $\mathbf{u}_k \in \mathbb{R}^m$ et $\mathbf{y}_k \in \mathbb{R}^p$ sont respectivement les vecteurs d'état, d'entrée et de sortie. Les vecteurs **w** et **v** représentent les bruits d'état et de mesure et appartiennent respectivement aux pavés $[\mathbf{w}_k]$ et $[\mathbf{v}_k]$. Enfin, les fonctions **f** et **g** sont supposées non linéaires.

Dans la suite de ce chapitre, nous allons supposer que les bruits d'état et de mesure sont bornés et de bornes connues :

$$\begin{cases} \mathbf{w}_k \in [\mathbf{w}_k] \\ \mathbf{v}_k \in [\mathbf{v}_k] \end{cases} \tag{2}$$

Notre objectif est de déterminer à chaque instant *k*, le plus petit ensemble $\mathcal{X}_k$ contenant tous les vecteurs d'état $\mathbf{x}_k$ compatibles avec les mesures et avec les bornes d'erreur fixées *a priori*.

Avant de présenter une approche ensembliste pour résoudre le problème (1), telle que suggérée dans [Kie98] [KJW02], nous allons tout d'abord rappeler quelques définitions [FA03].

2.1. Trajectoires et tubes de trajectoires

Définition 1 - Trajectoires

On appelle trajectoire d'état $\mathbf{z}_{\mathbf{x}}\left(\mathbf{x}_i(k_0), k_0, k_N, \{\mathbf{u}\}\right)$ une suite discrète de vecteurs $\mathbf{x}_k$ obtenus à partir du modèle (1) pour un état initial $\mathbf{x}_i(k_0)$ donné, et une séquence d'entrée $\{\mathbf{u}\} = \{\mathbf{u}(k_0), \ldots, \mathbf{u}(k_N - 1)\}$. Les deux entiers k_0 et k_n vérifient : $k_0 < k_N$. De la même manière, on définit une trajectoire de sortie.

Définition 2 – Tube de trajectoire

On appelle un tube de trajectoire d'état $\mathbf{T}_{\mathbf{x}}\left(\mathbf{x}_i(k_0), k_0, k_N, \{\mathbf{u}\}, \delta\right)$ de taille δ une suite de sous-ensembles $\mathbf{X}_k$ de l'espace d'état, définis par :

$$\mathbf{X}_k = \left\{\mathbf{x} \in \mathbb{R}^n \mid \left\|\mathbf{x} - \mathbf{x}_z(k)\right\|_\infty < \delta\right\} \tag{3}$$

où $\mathbf{x}_z(k)$ est le $(k - k_0 + 1)^{\text{ème}}$ point de la trajectoire $\mathbf{z}_{\mathbf{x}}\left(\mathbf{x}_i(k_0), k_0, k_N, \{\mathbf{u}\}\right)$. Afin de simplifier les notations, un tube de trajectoire sera noté par $\mathbf{T}_{\mathbf{x}}\left(\mathbf{x}_i(k_0), \delta\right)$ [FA03].

De la même manière, on peut définir un tube de trajectoire de sortie.

2.2. Distingabilité et observabilité numériques

Avant de chercher un observateur d'état pour un système, il est indispensable de vérifier la condition d'observabilité. Sans cette dernière propriété, il n'est pas possible d'affirmer s'il est possible d'observer l'état du système. Dans cette section, nous allons donner une définition de l'observabilité numérique [FA03] qui donne une indication sur l'observabilité (ou *N-Observabilité*) du système.

Définition 3 – Indistinguabilité pour une précision *p*

Considérons un système dynamique observé avec une précision *p* ($p>0$), i.e. la taille du tube de sortie à un instant donné est inférieure ou égale à *p*. Deux points $\mathbf{x}_1$ et $\mathbf{x}_2$ sont dits indistingables si le tube de la trajectoire de sortie engendré par le premier point contient la trajectoire du second.

Définition 4 – Observabilité numérique

Un système est dit *numériquement observable* (ou *N-observable*), si, pour une précision donnée sur l'état, notée $\varepsilon_{\mathbf{x}}$, il est possible de trouver une précision *p* telle que tous les points séparés d'une distance supérieure à $\varepsilon_{\mathbf{x}}$ sont distingables pour la précision *p* :

$$\forall\ \mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^n\ \ \left\|\mathbf{x}_1 - \mathbf{x}_2\right\|_\infty > \varepsilon_{\mathbf{x}} \Rightarrow \mathbf{z}_{\mathbf{y}}\left(\mathbf{x}_2\right) \notin \mathbf{T}\left(\mathbf{x}_1,\ p\right) \tag{4}$$

Cette définition implique que deux points séparés de plus de $\varepsilon_{\mathbf{x}}$ ne peuvent pas donner lieu à des trajectoires contenues dans un tube de taille *p*. On doit noter que l'observabilité numérique n'implique pas forcément l'observabilité au sens classique mais donne uniquement une indication qualitative.

2.3. Estimation d'état

On rappelle dans ce paragraphe l'algorithme récursif présenté dans [Kie98] [KJW02]. On note alors par $\mathcal{X}_k$ un ensemble contenant d'une manière garantie tous les vecteurs d'état compatibles avec les données expérimentales et avec les bornes d'erreurs de mesure jusqu'à l'instant *k*. L'algorithme proposé dans [Kie98] [KJW02] permet de calculer un ensemble $\mathcal{X}_{k+1}$ contenant les vecteurs d'état compatibles avec $\mathcal{X}_k$, la mesure disponible à l'instant *k* et la borne d'erreur correspondante. Il est basé sur une approche prédiction/correction.

Remarque 1 : La caractérisation exacte de l'ensemble $\mathcal{X}_k$ est un problème très difficile étant donné que le modèle est supposé non linéaire ; une approximation extérieure est alors recherchée. Pour décrire $\mathcal{X}_k$, on peut faire appel à des formes géométriques simples à manipuler comme les pavés. Néanmoins, cette méthode conduit en général à une surestimation de l'ensemble solution. Pour réduire ce pessimisme, il est préférable d'utiliser une union de pavés de tailles plus petites (on parle alors de *sous-pavage*). Ceci s'explique par le fait que, d'une part les sous-pavages décrivent mieux la forme de l'ensemble solution et d'autre part l'évaluation des fonctions est plus précise sur des intervalles de petite taille.

Prédiction : Cette étape consiste à trouver l'ensemble $\mathcal{X}_{k+1}^{+}$ des vecteurs d'état à l'instant *k*+1 accessibles à partir de $\mathcal{X}_k$. Cet ensemble est donné par :

$$\begin{aligned}\mathcal{X}_{k+1}^{+} &= \left\{\mathbf{f}(\mathbf{x}_k, \mathbf{u}_k, \mathbf{w}_k) \mid \mathbf{x}_k \in \mathcal{X}_k,\ \mathbf{w}_k \in \left[\mathbf{w}_k\right]\right\} \\ &= \mathbf{f}\left(\mathcal{X}_k, \mathbf{u}_k, \left[\mathbf{w}_k\right]\right)\end{aligned} \tag{5}$$

L'étape de prédiction revient donc à l'évaluation d'une fonction sur un ensemble.

Correction : L'étape de correction consiste à trouver l'ensemble des vecteurs d'état $\mathbf{x}_{k+1} \in \mathcal{X}^{+}_{k+1}$ compatibles avec les données expérimentales disponibles à l'instant k+1.

L'ensemble des vecteurs d'état $\mathbf{x}_{k+1}$ compatibles à l'instant k avec les mesures et les bornes d'erreurs est donné par :

$$\begin{aligned} \mathcal{X}^{\circ}_{k+1} &= \left\{ \mathbf{x} \in \mathbb{R}^n \mid \mathbf{g}(\mathbf{x}) \in [\mathbf{y}_{k+1}] \right\} \\ &= \mathbf{g}^{-1}\left([\mathbf{y}_{k+1}]\right) \end{aligned} \tag{6}$$

L'ensemble $\mathcal{X}_{k+1}$ contenant d'une manière garantie tous les vecteurs d'état à l'instant k+1 compatibles avec l'état précédent, les mesures et les bornes d'erreurs est donné par :

$$\mathcal{X}_{k+1} = \mathcal{X}^{+}_{k+1} \cap \mathcal{X}^{\circ}_{k+1} \tag{7}$$

La figure 1, tirée de [Kie98], illustre le principe de cette méthode récursive d'estimation d'état.

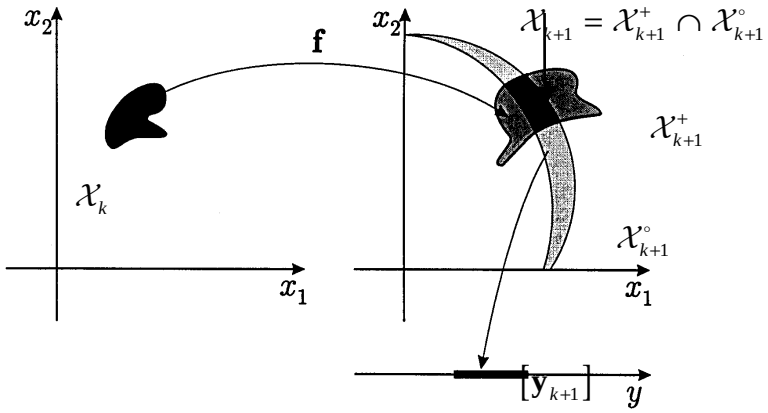

Figure 1 : Principe de l'algorithme d'estimation d'état basé sur (7)

Algorithme 1 : Estimateur d'état pour des systèmes à temps discret

Entrées : $\mathcal{X}_0$, $\{\mathbf{u}\}$, $\{[\mathbf{w}]\}$, $\mathbf{f}$, $\mathbf{g}$

1. Pour $k = 0$ à N calculer
2. $\mathcal{X}^{+}_{k+1} = \mathbf{f}\left(\mathcal{X}_k, \mathbf{u}_k, [\mathbf{w}_k]\right)$
3. $\mathcal{X}^{\circ}_{k+1} = \mathbf{g}^{-1}\left([\mathbf{y}_{k+1}]\right)$
4. $\mathcal{X}_{k+1} = \mathcal{X}^{+}_{k+1} \cap \mathcal{X}^{\circ}_{k+1}$

Sorties : $\{\mathcal{X}\}$

On démontre dans [Kie98] que l'approximation extérieure de l'ensemble solution est minimale, i.e. on ne peut pas trouver un sous-pavage $\mathcal{X}_{k+1}^{-} \subset \mathcal{X}_{k+1}$ qui contient l'ensemble solution. Dans [KJW02], on trouve une étude de convergence de l'algorithme 1.

3. Estimation d'état pour des systèmes non linéaires à temps continu

On considère maintenant un système décrit par le modèle suivant :

$$\begin{cases} \dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, \mathbf{u}) \\ \mathbf{y} = \mathbf{g}(\mathbf{x}, \mathbf{u}, \mathbf{v}) \\ \mathbf{x}_0 \in [\mathbf{x}_0] \end{cases} \tag{8}$$

où $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{u} \in \mathbb{R}^m$ et $\mathbf{y} \in \mathbb{R}^p$ sont respectivement les vecteurs d'état, d'entrée et de sortie. Le vecteur **v** représente le bruit de mesure supposé borné et de borne connue $\overline{e}$. Les fonctions **f** et **g** sont supposées non linéaires. On suppose dans la suite de ce chapitre qu'il n'y a pas de bruit d'état ; cette restriction est due au fait que la résolution des équations différentielles, contenant un bruit d'état, à l'aide des techniques d'analyse par intervalles fournit des résultats très pessimistes.

3.1. Estimateur causal

Dans cette section, nous allons étendre l'estimateur donné par l'algorithme 1 au cas des systèmes décrits par des équations différentielles non linéaires. Afin de simplifier la présentation, nous n'utiliserons pas de sous-pavages mais l'approximation extérieure sera donnée par un seul pavé. Néanmoins, il est aisé d'étendre cette approche en utilisant des sous-pavages afin de réduire le pessimisme.

On a vu dans l'algorithme 1 que l'étape de prédiction consiste à évaluer une fonction **f** sur un pavé (ou ensemble de pavés). Dans le cas des systèmes à temps continu et lorsqu'une solution explicite de l'EDO n'est pas disponible, il est nécessaire d'évaluer numériquement la solution de cette EDO. L'étape de prédiction consiste donc à calculer à chaque instant k+1 une approximation extérieure $[\mathbf{x}_{k+1}^{+}]$ de toutes les solutions $\mathbf{x}_{k+1}$ de l'EDO ($\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, \mathbf{u})$) partant d'un état initial appartenant à l'instant k à $[\mathbf{x}_k]$. Cette approximation est calculée à l'aide de l'une des méthodes d'intégration numérique garantie des équations différentielles présentées dans le chapitre 4.

$$[\mathbf{x}_{k+1}^{+}] = \mathrm{IODE}\left(\mathbf{f}, [\mathbf{x}_k]\right) \tag{9}$$

où IODE désigne l'une des méthodes d'intégration de l'EDO définie par ($\dot{\mathbf{x}}=\mathbf{f}(\mathbf{x},\mathbf{u})$) et $[\mathbf{x}_k]$ est un pavé contenant toutes les valeurs possibles du vecteur d'état à l'instant *k*.

Remarque 2 : Dans le chapitre 4, on a traité uniquement le cas d'EDOs autonomes ; dans le contexte de l'estimation d'état, les EDOs sont non autonomes étant donné que les systèmes sont souvent commandés. Néanmoins, ceci ne pose pas de problème lorsque la commande est donnée par une fonction continue ; dans ce cas, il suffit de considérer *t* (temps) comme une variable d'état pour obtenir une EDO autonome. On doit noter que le cas des systèmes commandés avec une commande discrète n'est pas abordé dans ce chapitre.

Remarque 3 : Les méthodes d'intégration d'EDOs présentées dans le chapitre 4 utilisent des pas d'intégration variables, cependant, les mesures sont en général prélevées avec une période d'échantillonnage fixe. En premier lieu, on utilisera donc un pas fixe égal à la période d'échantillonnage. ♦

Comme dans le cas des systèmes à temps discret, l'étape de correction consiste à trouver tous les vecteurs d'état compatibles avec l'état à l'instant *k*, les mesures disponibles à l'instant *k*+1 et les bornes d'erreur de mesure :

$$[\mathbf{x}_{k+1}^{+}] = [\mathbf{x}_{k+1}^{+}] \cap \mathbf{g}^{-1}\left([\mathbf{y}_{k+1}]\right) \tag{10}$$

L'estimateur d'état est alors donné par l'algorithme suivant [RRC03a] [RRC03b] :

Algorithme 2 : Estimateur d'état pour des systèmes à temps continu

Entrées : $[\mathbf{x}_0]$, **f**, **g**, $[\mathbf{y}_1],\ldots,[\mathbf{y}_N]$

1. Pour $k = 0$ à N-1, calculer
2. $\left[\mathbf{x}_{k+1}^{+}\right] = \text{IODE}\left(\mathbf{f},[\mathbf{x}_k]\right)$
3. $\left[\mathbf{x}_{k+1}^{\circ}\right] = \mathbf{g}^{-1}\left([\mathbf{y}_{k+1}]\right)$
4. $[\mathbf{x}_{k+1}] = \left[\mathbf{x}_{k+1}^{+}\right] \cap \left[\mathbf{x}_{k+1}^{\circ}\right]$

Sorties : $[\mathbf{x}_1],\ldots,[\mathbf{x}_N]$

Par souci de clarté, on n'a pas mentionné dans les entrées de l'algorithme les paramètres nécessaires à la résolution de l'EDO (pas d'intégration, ordre du développement de Taylor…).

Exemple 1 [RRC04] : On considère le modèle de Lotka-Volterra comportant deux variables d'état x_1 et x_2 qui représentent respectivement la taille de l'espèce proie et celle des prédateurs. On suppose qu'il n'y a aucune compétition entre les individus d'une même espèce. L'évolution de la taille des deux espèces est donnée par le modèle suivant :

$$\begin{cases} \dot{x}_1 = (a - b\,x_2)x_1 \\ \dot{x}_2 = (-c + d\,x_1)x_2 \end{cases} \tag{11}$$

où *a*, *b*, *c* et *d* sont des constantes positives. Les paramètres *a* et *c* sont respectivement les taux de naissance des proies et des prédateurs. Le terme $-bx_2x_1$ représente la décroissance de la population des proies due aux prédateurs et le terme dx_1x_2 est le taux de croissance grâce à cette même rencontre avec les proies. On suppose dans cet exemple que les quatre paramètres *a*, *b*, *c* et *d* sont parfaitement connus (a = 1, b = 0.01, c = 1 et d = 0.02).

Notre but est d'estimer l'évolution de la taille des deux populations lorsque la taille des populations initiales n'est pas connue de manière exacte, mais appartient aux pavés $\mathbf{x}_0 = [49;51]\times[49;51]$. On suppose qu'on mesure x_1 (l'équation d'observation est alors $y = g(x_1, x_2) = x_1$). Les pseudo-mesures sont obtenues en simulant le modèle avec $\mathbf{x}_0 = [50,50]\times[50,50]$ et en rajoutant un bruit numérique uniforme dans l'intervalle $[-1.5, 1.5]$, les domaines $[y_j]$ sont alors donnés par $[y_j] = [\hat{y}_j - 1.5, \hat{y}_j + 1.5]$. L'estimateur d'état donné par l'algorithme 2 génère en 0.3 s l'ensemble des pavés tracés sur la figure 2. Sur la figure 3, nous avons tracé l'ensemble des pavés obtenus sans correction. Dans les deux cas l'étape de prédiction a été réalisée en utilisant la méthode de la valeur moyenne étendue (voir chapitre 4) avec un pas d'intégration constant (égal à la période d'échantillonnage) h = 0.005 et un développement de Taylor d'ordre k = 4 ; le nombre de points est N = 900.

On remarque d'après les figures 2 et 3 que la taille des pavés n'augmente pas systématiquement. En particulier, la méthode d'intégration de l'EDO ($\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, \mathbf{u})$) avec un état initial incertain n'a pas divergé comme le montre la figure 3 contenant le résultat de l'estimation sans correction.

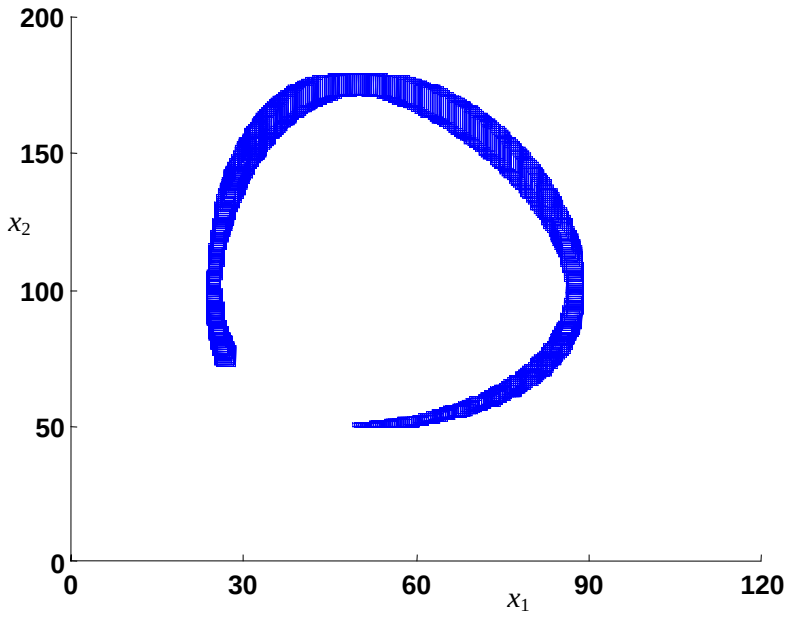

Figure 2 : Ensemble des pavés générés par l'algorithme 2, pour N = 1000 et $[\mathbf{x}_0] = [49,51]\times[49,51]$.

Les approximations de l'ensemble solution tracées sur les figures 2 et 3 sont moins pessimistes que celles obtenues dans [Jau02] ; dans l'estimateur présenté dans cette dernière publication, l'étape de prédiction était basée sur une méthode d'intégration numérique de premier ordre et sur des méthodes de consistance. De plus, contrairement à [Jau02], l'estimateur proposé dans ce chapitre ne nécessite aucune bissection du vecteur d'état lors de la phase de prédiction ; ceci le rend applicable à des systèmes de grande dimension.

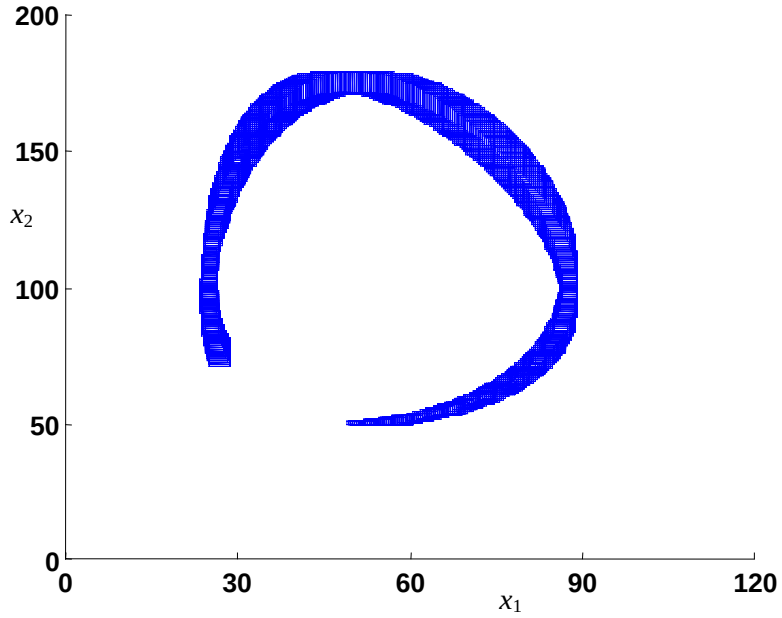

Figure 3 : Ensemble des pavés générés par l'algorithme 2 sans l'étape de correction pour N = 1000 et $[\mathbf{x}_0]$ = [49,51]×[49,51].

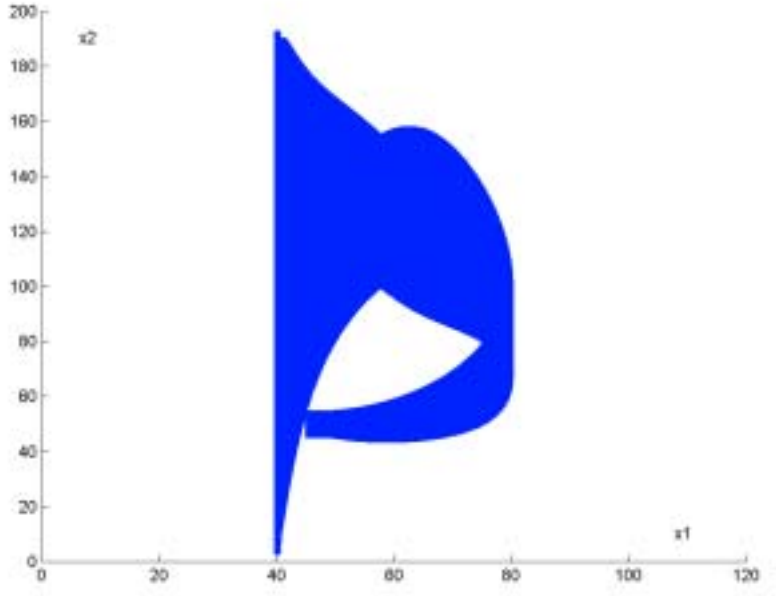

Figure 4 : Ensemble des pavés générés par l'algorithme 2 pour $[\mathbf{x}_0] = [45,55]\times[45,55]$

Considérons maintenant l'exemple 1 avec la condition initiale $[\mathbf{x}_0]=[45,55]\times[45,55]$; l'estimateur d'état donné par l'algorithme 2 génère une approximation extérieure de la trajectoire d'état tracée sur la figure 4. On remarque alors que la taille des pavés diverge après seulement quelques pas d'échantillonnage. Ceci montre bien que cet algorithme n'est utilisable que lorsque la taille de l'état initial est suffisamment petite.

3.2. Estimateur non causal

Dans la suite de ce paragraphe, nous allons étendre l'algorithme 2 dans un contexte hors ligne afin d'obtenir un estimateur utilisable lorsque le domaine d'incertitude de l'état initial est important. Dans un tel contexte, on suppose que toutes les mesures $(\hat{\mathbf{y}}_1, \hat{\mathbf{y}}_2,\ldots,\hat{\mathbf{y}}_N)$ sont disponibles et on estime le vecteur d'état aux différents instants simultanément.

Notre but est de trouver des approximations extérieures pour les variables $\mathbf{x}(1), \mathbf{x}(2),\ldots,\mathbf{x}(N)$ sachant que les domaines *a priori* $[\mathbf{y}_1],[\mathbf{y}_2],\ldots,[\mathbf{y}_N]$ des sorties sont tous disponibles. L'algorithme proposé est une extension d'un estimateur développé dans [JKDW01] pour le cas des systèmes non linéaires à temps discret. Nous proposons dans ce paragraphe de traiter le problème d'estimation d'état comme étant un problème de satisfaction de contraintes (CSP) dont les contraintes sont données par :

$$\text{CSP}\ \begin{cases} \mathbf{x}(1) = \text{IODE}(\mathbf{x}(0)) \\ \mathbf{x}(1) = \mathbf{g}^{-1}(\mathbf{y}(1)) \\ \quad\vdots \\ \mathbf{x}(N) = \text{IODE}(\mathbf{x}(N\text{-}1)) \\ \mathbf{x}(N) = \mathbf{g}^{-1}(\mathbf{y}(N)) \end{cases} \tag{12}$$

où IODE est l'une des méthodes d'intégration numérique garantie de l'EDO ($\dot{\mathbf{x}}=\mathbf{f}(\mathbf{x})$). On doit noter que l'entrée ne figure pas dans les équations compte tenu de la remarque 2. On suppose que les domaines initiaux *a priori* des variables $\mathbf{x}(1), \mathbf{x}(2),\ldots,\mathbf{x}(N)$ sont $[\mathbf{x}(0)]$, $[\mathbf{x}(1)],\ldots,[\mathbf{x}(N)]$; dans le cas où aucune information sur une variable $\mathbf{x}(i)$ n'est disponible, son domaine est $]-\infty,+\infty[$. Le graphe du CSP est présenté sur la figure 5.

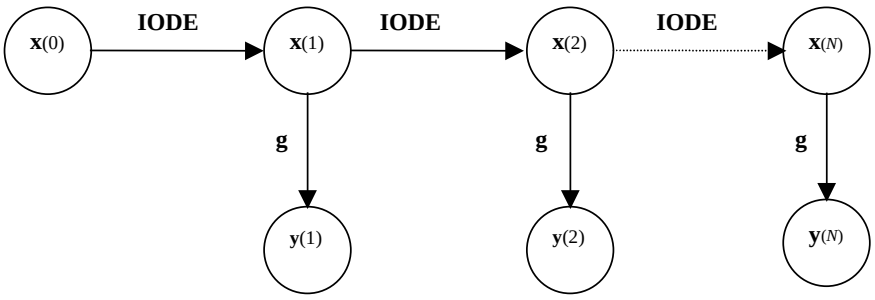

Figure 5: Graphe correspondant au CSP (12)

Nous allons alors utiliser le contracteur $\mathcal{C}_{\uparrow\downarrow}$ (*propagation – rétropropagation*) présenté au chapitre 1.

Propagation : La phase de propagation consiste à parcourir le graphe donné par la figure 5 en partant de l'état initial **x**(0). Cette étape peut parfois réussir à contracter le domaine des variables $\{\mathbf{x}_k\}$ à condition que le pessimisme de l'intégration numérique soit contourné et la taille de l'état initial ainsi que celle des mesures soient petites.

Algorithme 3 : Propagation

Entrées : $[\mathbf{x}_0]$, **f**, **g**, $[\mathbf{y}_1],\ldots,[\mathbf{y}_N]$

1. Pour $k = 0$ à N-1, calculer

2. $[\mathbf{x}_{k+1}] = \left(\text{IODE}\left(\mathbf{f},[\mathbf{x}_k]\right)\right) \cap [\mathbf{x}_{k+1}]$

3. $[\mathbf{x}_{k+1}] = \left(\mathbf{g}^{-1}\left([\mathbf{y}_{k+1}]\right)\right) \cap [\mathbf{x}_{k+1}]$

Sorties : $[\mathbf{x}_0],\ldots,[\mathbf{x}_N]$, $[\mathbf{y}_1],\ldots,[\mathbf{y}_N]$

Rétropropagation : La rétropropagation consiste à parcourir le graphe donné par la figure 5 en partant de $[\mathbf{y}_N]$. Ceci suppose que toutes les mesures sont disponibles ; cette étape nous permet alors de trouver les domaines des vecteurs d'état consistants avec toutes les mesures, i.e. chaque domaine est consistant avec les informations du passé et du futur. La rétropropagation est donnée par l'algorithme 4.

Algorithme 4 : Rétropropagation

Entrées : $[\mathbf{x}_0],\ldots,[\mathbf{x}_N]$, **f**, **g**, $[\mathbf{y}_1],\ldots,[\mathbf{y}_N]$

1. Pour $k = N$ à 1, calculer

2. $[\mathbf{y}_k] = \left(\mathbf{g}\left([\mathbf{x}_k]\right)\right) \cap [\mathbf{y}_k]$

3. $[\mathbf{x}_{k-1}] = \left(\text{IODE}^{-1}\left(\mathbf{f},[\mathbf{x}_k]\right)\right) \cap [\mathbf{x}_{k-1}]$

Sorties : $[\mathbf{x}_0],\ldots,[\mathbf{x}_N]$, $[\mathbf{y}_1],\ldots,[\mathbf{y}_N]$

Dans l'algorithme 4, on note par IODE^{-1} une méthode d'intégration numérique de l'EDO ($\dot{\mathbf{x}}=\mathbf{f}(\mathbf{x})$) dans le sens chronologique inverse, i.e. avec un pas d'intégration $-h$ (où h est le pas

d'intégration égal au pas d'échantillonnage). Ceci nous permet de calculer la solution de l'EDO à (*k*-1) connaissant la solution à l'instant *k*.

L'estimateur d'état proposé consiste alors à appliquer successivement la propagation et la rétropropagation. En général, en parcourant une seul fois (dans les deux sens) le graphe de la figure 5, on n'obtient pas d'approximations extérieures minimales ; ceci s'explique par le pessimisme dû aux effets d'enveloppement et de dépendance présents dans l'analyse par intervalles. Pour réduire ce pessimisme, on effectue des propagations – rétropropagations tant que l'on note des améliorations (réduction de pessimisme) de la taille des domaines de l'état.

Exemple 2 : On considère de nouveau l'exemple 1 avec les mêmes conditions ; on suppose alors que l'état initial est $[\mathbf{x}_0]=[45,55]\times[45,55]$. L'estimateur non causal basé sur le contracteur propagation–rétropropagation génère l'ensemble des pavés tracés sur la figure 6. On remarque alors que la taille des pavés ne diverge pas comme dans le cas de l'estimateur donné par l'algorithme 2 (voir figure 4). On remarque aussi que même le domaine de l'état initial a été contracté.

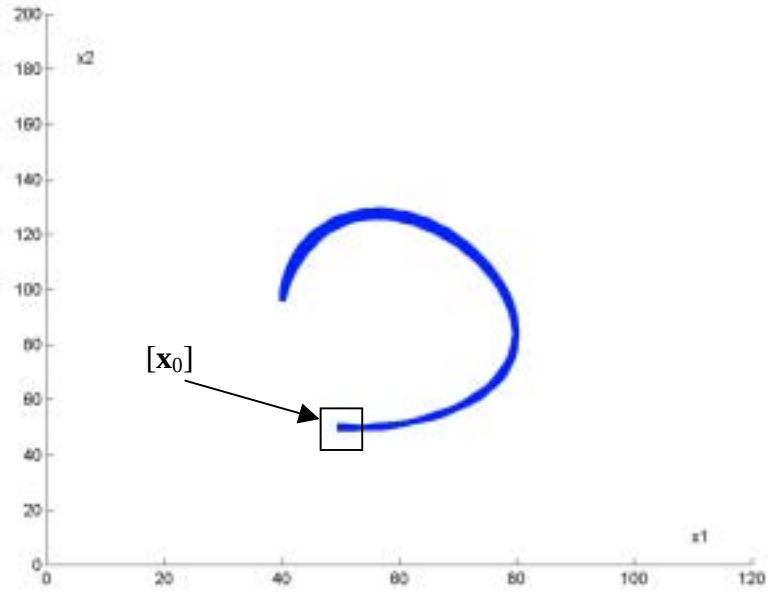

Figure 6 : Ensemble des pavés générés par l'estimateur non causal pour $[\mathbf{x}_0]=[45,55]\times[45,55]$

3.3. Estimateur d'état à horizon glissant

3.3.1. Estimateur

Comme on l'a indiqué au début du chapitre, l'utilisation de l'observateur de Luenberger et du filtre de Kalman pour des systèmes linéaires permet de trouver des résultats satisfaisants sous certaines conditions. Ces estimateurs sont aussi utilisés pour des modèles non linéaires en se basant sur des linéarisations locales. Comme nous l'avons mentionné, dans ce dernier cas,

l'estimée obtenue peut être éloignée de la vraie solution. Une alternative intéressante consiste à utiliser des techniques à horizon glissant [MM95] [RRL01] [GFA03]. L'idée consiste à transformer le problème d'estimation en un problème de résolution d'un système de *N* équations (où *N* est le nombre de mesures) ou sous la forme d'un problème d'optimisation. Il est généralement résolu en utilisant des méthodes d'optimisation itératives.

Dans ce paragraphe nous allons étendre l'algorithme présenté dans la section 3.2. afin de développer un estimateur d'état à horizon glissant. Le principe de la méthode est donné par la figure 7.

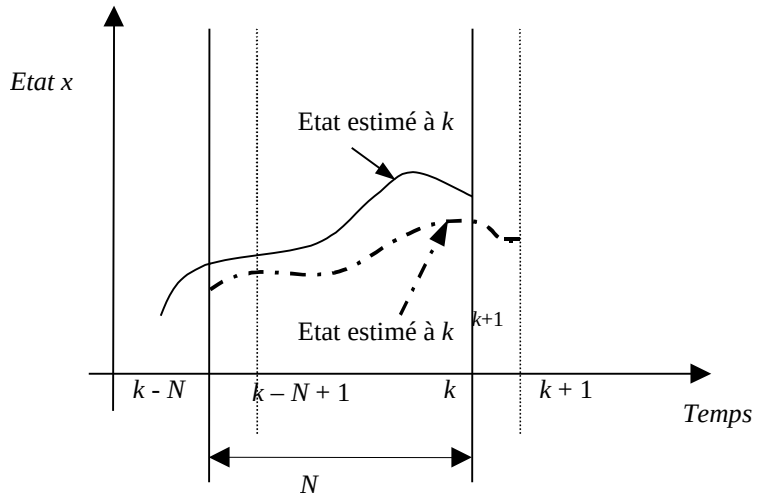

Figure 7 : Principe de l'estimation d'état avec un horizon glissant

Sur la figure 7, on note par *N* la taille de l'horizon, on suppose alors que le nombre de mesures N' vérifie $N' \gg N$. Le principe de l'estimateur proposé dans ce paragraphe consiste à utiliser à chaque instant les *N* dernières mesures pour estimer l'état sur cet horizon.

Cette approche à horizon glissant a été utilisée dans un contexte à erreurs bornées pour des systèmes à temps discret dans [FA03], l'observateur proposé dans [FA03] consiste donc à combiner l'approche prédiction/correction avec le contracteur de Newton. L'utilisation de ce contracteur nécessite une taille de l'horizon fixe et égale à la dimension du vecteur d'état ; en effet, cette dernière restriction permet d'avoir un nombre d'équations égal au nombre de variables à déterminer.

Dans la suite de paragraphe, nous proposons un observateur d'état à horizon glissant pour des systèmes décrits par des EDOs [RRC05]. Il consiste à combiner l'approche prédiction/correction et le contracteur propagation – rétropropagation. L'utilisation de ce contracteur nous permet de considérer une taille de l'horizon pouvant être différente de la dimension du vecteur d'état.

Le principe de l'observateur consiste à utiliser l'algorithme présenté dans la section 3.2. afin de réduire les domaines des variables d'état entre l'instant *k-N* et *k* (où *k* est la fin de l'horizon et *k-N* est son début). Ensuite, dès qu'une nouvelle mesure est disponible, on fait glisser l'horizon d'un pas afin de tenir compte de cette nouvelle information.

L'estimateur est alors donné par l'algorithme 5 [RRC05].

Algorithme 5 : Estimateur d'état à horizon glissant

Entrées : $[\mathbf{x}_0], [\mathbf{y}_1],\ldots,[\mathbf{y}_{N'}]$

1. Pour $i = N$ à N' // On fait glisser la fin de l'horizon
2. Pour $k= i-N$ à i-1, calculer // Propagation
3. $[\mathbf{x}_{k+1}] = \left(\text{IODE}\left(\mathbf{f}, [\mathbf{x}_k]\right)\right) \cap [\mathbf{x}_{k+1}]$
4. $[\mathbf{x}_{k+1}] = \left(\mathbf{g}^{-1}\left([\mathbf{y}_{k+1}]\right)\right) \cap [\mathbf{x}_{k+1}]$
5. Pour $k= i$ à $(i - N + 1)$, calculer // Rétropropagation
6. $[\mathbf{y}_k] = \left(\mathbf{g}\left([\mathbf{x}_k]\right)\right) \cap [\mathbf{y}_k]$
7. $[\mathbf{x}_{k-1}] = \left(\text{IODE}^{-1}\left(\mathbf{f}, [\mathbf{x}_k]\right)\right) \cap [\mathbf{x}_{k-1}]$

Sorties : $[\mathbf{x}_0],\ldots,[\mathbf{x}_{N'}], [\mathbf{y}_1],\ldots,[\mathbf{y}_{N'}]$

On note que dans la première étape de l'algorithme 5, la boucle servant à faire glisser l'horizon commence à l'instant N, i.e. l'algorithme commence dès qu'on a N mesures. Néanmoins, on peut utiliser l'algorithme 1 (prédiction-correction) pour estimer l'état aux instants (1,…, N-1). Par souci de clarté on a appliqué les phases de propagation et rétro-propagation une seule fois, en pratique on effectue plusieurs allers/retours afin de réduire le pessimisme.

3.3.2. Application à un bioprocédé

Un bioprocédé est un procédé constitué par des bactéries ou des champignons ; il est difficile de trouver un modèle reproduisant d'une manière satisfaisante le comportement de ce type de systèmes. Les méthodes ensemblistes constituent donc une approche intéressante pour caractériser l'incertitude sur ces modèles. Dans ce contexte, un estimateur d'état pour des bioprocédés décrits par des équations différentielles a été proposé dans [RFC97]. Néanmoins, on ne peut en aucun cas garantir que les approximations extérieures calculées contiennent la vraie solution. D'autres estimateurs utilisant une discrétisation temporelle du modèle continu ont été proposés dans [FA03] [GFA03]. Dans la suite de ce paragraphe, nous allons étudier un bioprocédé à l'aide de l'estimateur à horizon glissant présenté dans la section 3.3.1.

L'exemple étudié est tiré de la littérature [BH01]. Le procédé considéré est constitué de biomasse, de glucose, de glutamine et de lactate. Les réactions chimiques en présence sont les suivantes :

Croissance : $$a_1 \cdot G ln \xrightarrow{\varphi_c} X \tag{13}$$

Maintenance : $$G + a_2 \cdot X \xrightarrow{\varphi_m} a_2 X + a_3 L \tag{14}$$

Les variables *X*, *G*, *Gln* et *L* représentent respectivement les concentrations en biomasse, en glucose, en glutamine et en lactate ; a_1, a_2 et a_3 sont les coefficients stœchiométriques ; φ_c et φ_m sont les taux de croissance et de maintenance dans la réaction. A l'aide des réactions (13) et (14) on obtient l'ensemble des équations régissant l'évolution des quatre éléments [BCH99] [BH01] :

$$\begin{aligned}
\frac{dX(t)}{dt} &= \varphi_c(X(t), G(t), G ln(t), L(t)) \\
\frac{dG(t)}{dt} &= -\varphi_m(X(t), G(t), G ln(t), L(t)) \\
\frac{dG ln(t)}{dt} &= -a_1 \cdot \varphi_c(X(t), G(t), G ln(t), L(t)) \\
\frac{dL(t)}{dt} &= a_3 \cdot \varphi_m(X(t), G(t), G ln(t), L(t))
\end{aligned} \tag{15}$$

Les modèles cinétiques servant à décrire les taux de réactions φ_c et φ_m sont donnés par [Bogaerts et al., 1999] :

$$\begin{aligned}
\varphi_c(X, G, G ln, L) &= \alpha_c X^{\gamma_{c,X}} G^{\gamma_{c,G}} G ln^{\gamma_{c,G ln}} e^{-\beta_{c,X} X} e^{-\beta_{c,G} G} e^{-\beta_{c,G ln} G ln} e^{-\beta_{c,L} L} \\
\varphi_m(X, G, G ln, L) &= \alpha_m X^{\gamma_{m,X}} G^{\gamma_{m,G}} G ln^{\gamma_{m,G ln}} e^{-\beta_{m,X} X} e^{-\beta_{m,G} G} e^{-\beta_{m,G ln} G ln} e^{-\beta_{m,L} L}
\end{aligned} \tag{16}$$

Les paramètres figurant dans (16) ont été identifiés dans [BCH99] et on obtient alors les expressions suivantes pour φ_c et φ_m :

$$\begin{aligned}
\varphi_c(X, G, G ln, L) &= 0.129 X^{0.33} G ln^{0.069} e^{-0.006 G} \\
\varphi_m(X, G, G ln, L) &= 0.043 X^{1.006} G^{0.114} e^{-0.09 X}
\end{aligned} \tag{17}$$

On note par $\mathbf{x} = (x_1, x_2, x_3, x_4)^T = (X, G, Gln, L)^T$ le vecteur d'état. Les équations décrivant le système sont données par :

$$\begin{cases}
\dot{x}_1 = 0.129 x_1^{0.33} x_3^{0.069} e^{-0.006 x_2} \\
\dot{x}_2 = -0.043 x_1^{1.006} x_2^{0.114} e^{-0.09 x_1} \\
\dot{x}_3 = -0.030444 x_1^{0.33} x_3^{0.069} e^{-0.006 x_2} \\
\dot{x}_4 = 0.064586 x_1^{1.006} x_2^{0.114} e^{-0.09 x_1} \\
\mathbf{y} = (0,1,1,1)^T \mathbf{x} + \mathbf{v}
\end{cases} \tag{18}$$

Notre but est de donner une approximation extérieure de la concentration de la biomasse en supposant que celles du glucose, de la glutamine et du lactate sont mesurées. Dans ce paragraphe, les données mesurées sont obtenues en simulant le modèle (18) avec l'état initial $\mathbf{x}_0 = (6\times10^5, 20.5, 2.4, 1)^T$ (dans la suite, la première composante du vecteur d'état sera fixée à 6 et non pas à 6×10^5). Ces pseudo-mesures sont ensuite bruitées par un bruit numérique uniforme d'amplitude 1%. On suppose aussi qu'une mesure est réalisée toutes les 6 heures ; ce qui est assez réalisable compte tenu des coûts et du temps effectif de mesure.

Etude de cas :

On suppose que l'état initial est $[\mathbf{x}_0] = [4,8]\times[18,23]\times[1,3]\times[0.5,1.5]$, la taille de l'horizon est $N = 6$ et le pas d'intégration de l'EDO est $h = 2$, ce qui revient à effectuer deux intégrations avant chaque correction. D'autre part, comme on effectue une mesure toutes les 6 heures, l'étape de correction est réalisée seulement quand une nouvelle mesure est disponible. Enfin, pour améliorer les performances, on a partitionné le pavé contenant le vecteur d'état au début de l'horizon en 10 boîtes.

L'estimateur à horizon glissant génère l'ensemble des pavés tracés sur la figure 8 ; on remarque alors que la taille de ces pavés n'augmente pas systématiquement même avec une incertitude importante sur l'état initial, et l'erreur est de l'ordre de 0.4 (i.e. 0.4×10^5 Cell/ml), soit entre 7 et 3%.

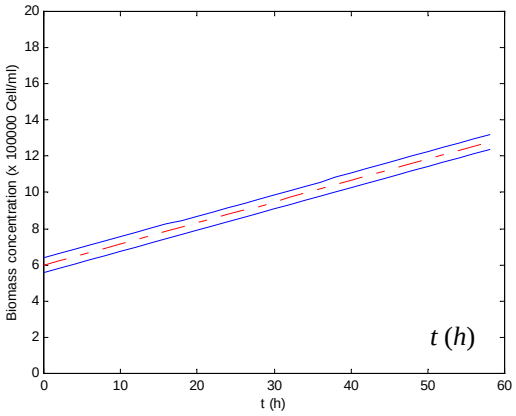

Figure 8 : Concentration de la Biomasse estimée avec l'estimateur à horizon glissant, l'état initial est $[x_0] = [4,8]\times[18,23]\times[1,3]\times[0.5,1.5]$, la taille de l'horizon $N = 6$ et avec 10 bissections. En ligne continue : estimée ; ligne interrompue : vraies valeurs de l'état.

Analyse de la taille de l'horizon

Pour montrer l'influence de la taille de l'horizon, nous considérons de nouveau le même exemple. On décide d'appliquer uniquement les étapes de prédiction et correction, i.e. on ne fait pas de propagation – rétropropagation. Les bornes des approximations obtenues sont tracées sur la figure 9. On remarque que la taille de ces intervalles est plus importante que celle obtenue avec un horizon $N = 6$. Ceci s'explique par le fait que le contracteur propagation – rétropropagation permet d'éliminer des parties du domaine du vecteur d'état à l'instant t_k

qui ne sont pas consistantes avec une des mesures sur tout l'horizon. De plus, l'utilisation de plusieurs mesures permet d'exploiter plus d'informations.

Supposons maintenant que la taille de l'horizon est $N = 1$ (on utilise deux mesures), on obtient alors une erreur sur la concentration initiale de la biomasse de 0.8, cette erreur décroît pour atteindre 0.68 à la fin du processus d'estimation (à t = 58 heures).

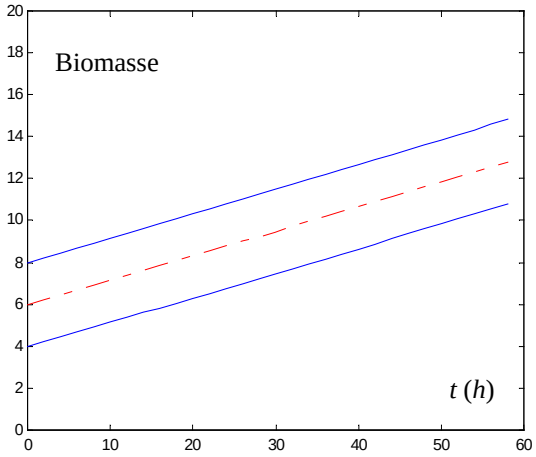

Figure 9 : Concentration de la Biomasse estimée avec l'estimateur prédicteur/correcteur, l'état initial est $[x_0] = [4,8]\times[18,23]\times[1,3]\times[0.5,1.5]$ et avec 10 bissections. En ligne continue : estimée ; ligne interrompue : vraies valeurs de l'état.

Cette exemple montre que la stratégie d'horizon glissant permet d'obtenir des approximations extérieures du vecteur d'état moins pessimistes que celles obtenues par l'utilisation de l'approche prédiction/ correction. Les performances de l'observateur à horizon glissant est d'autant meilleure que la taille de l'horizon est grande.

Influence du nombre de bissections

Dans ce paragraphe, nous avons fixé la taille de l'horizon à $N = 3$ et on a fait varier le nombre de bissections. Nous avons tracé sur la figure 10.a. la trajectoire de la concentration en biomasse estimée par l'observateur à horizon glissant en bissectant le vecteur d'état au début de l'horizon en deux ; l'erreur est de 2, soit de 33 à 16%. Maintenant, en effectuant 4 bissections, l'erreur est de 1, soit de 16 à 8% (voir Figure 9.b). Enfin, pour 10 bissections, l'observateur à horizon glissant génère les bornes inférieure et supérieure de la concentration en biomasse tracées sur la figure 10.c ; l'erreur est de 0.4, soit de 6 à 3%.

On constate d'après ces résultats que la bissection du vecteur d'état au début de l'horizon contribue à l'élimination des parties du vecteur d'état qui ne sont pas consistantes avec les mesures, ceci permet donc de réduire l'erreur sur l'état estimé. Néanmoins, on sait par ailleurs que le temps de calcul augmente considérablement lorsqu'on fait recours aux bissections. Mais ceci ne pose pas de problème dans notre application étant donné qu'on n'effectue qu'une seule mesure toutes les 6 heures.

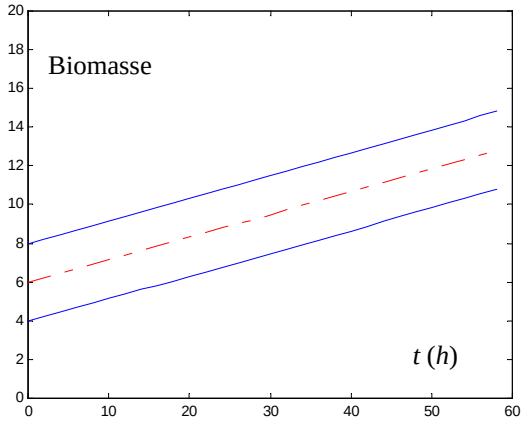

Figure 10.a

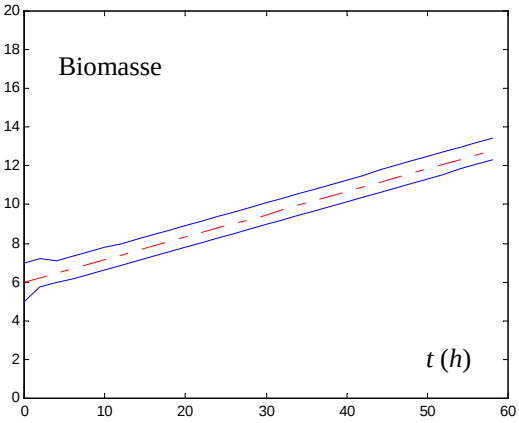

Figure 10.b

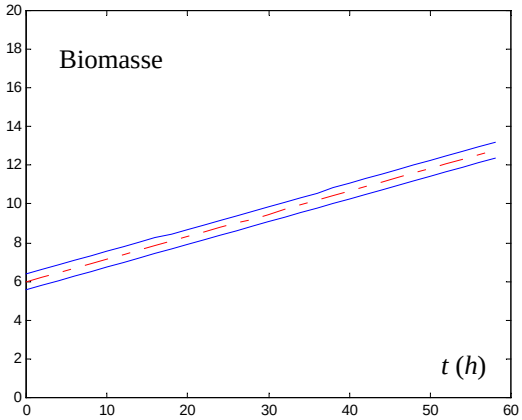

Figure 10.c

Figure 10 : Concentration en biomasse estimée à l'aide de l'observateur à horizon glissant avec $N = 3$ et en faisant varier le nombre de bissections :10.a avec 2 bissections, 10.b avec 4 bissections et 10.c avec 10 bissections.

Conclusion :

Dans cette section nous avons proposé un observateur d'état à horizon glissant pour des systèmes non-linéaires. L'observateur est basé sur l'arithmétique d'intervalles et l'inversion ensembliste. Les performances de cet observateur ont été étudiées sur un exemple de bioprocédé. Nous avons constaté que l'approximation extérieure du vecteur d'état est moins pessimiste lorsqu'on utilise, d'une part, une taille de l'horizon grande et de l'autre part, des bissections du vecteur d'état.

4. Estimation de paramètres pour des systèmes décrits par des équations différentielles non linéaires

Dans ce paragraphe, on va étudier le problème d'estimation garantie de paramètres dans des modèles décrits par des équations de la forme suivante :

$$\begin{cases} \dot{\mathbf{x}} = \mathbf{f}(\mathbf{p}, \mathbf{x}(t)) \\ \mathbf{y} = \mathbf{g}(\mathbf{p}, \mathbf{x}(t), \mathbf{v}) \\ \mathbf{x}(t_0) \in [\mathbf{x}_0] \\ \mathbf{p} \in [\mathbf{P}_0] \end{cases} \tag{19}$$

où $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{p} \in \mathbb{R}^m$ et $\mathbf{y} \in \mathbb{R}^p$ sont respectivement les vecteurs d'état, de paramètres à identifier et de sortie. Le vecteur **v** représente le bruit de mesure supposé borné et de borne connue $\overline{e}$. Les fonctions **f** et **g** sont supposées non linéaires.

L'estimateur donné par l'algorithme 2 peut être étendu afin d'estimer simultanément les vecteurs d'état et de paramètres. Dans ce cas, on construit un vecteur d'état étendu contenant l'état original et l'ensemble des paramètres à identifier ; cette approche a été étudiée dans [Lju79] dans un contexte statistique et a été utilisée dans un contexte à erreurs bornées dans [KJW02]. Cette idée a également été exploitée dans [WK03] dans le cadre de l'estimation de paramètres pour une classe particulière (systèmes coopératifs) de systèmes décrits par des EDOs.

Dans la suite, le problème d'estimation de paramètres est écrit sous la forme d'un problème d'inversion ensembliste. Il est alors résolu en utilisant l'algorithme SIVIA [Jau93]. D'autre part, le modèle n'est pas donné par une fonction explicite des paramètres inconnus, mais par une équation différentielle. L'idée proposée est alors d'utiliser une approximation numérique garantie de l'EDO.

L'ensemble des vecteurs de paramètres consistants avec les mesures et avec les bornes d'erreurs est donné par :

$$\mathbb{S} = \left\{ \mathbf{p} \in [\mathbf{P}_0] \;\middle|\; \forall t_j \in \{t_1, t_2, \ldots, t_N\}, \left(\dot{\mathbf{x}}(t_j) = \mathbf{f}(\mathbf{p}, \mathbf{x}(t_j))\right) \wedge \left(\mathbf{g}(\mathbf{x}(t_j, \mathbf{p}), \mathbf{p}) \in [\mathbf{y}_j]\right) \right\} \tag{20}$$

où $[\mathbf{P}_0]$ est un ensemble *a priori* contenant tous les vecteurs de paramètres solutions de (19). Pour trouver une approximation extérieure de l'ensemble $\mathbb{S}$ défini par (20), il suffit alors d'utiliser SIVIA avec un test d'inclusion.

On suppose que la solution de l'EDO ($\dot{\mathbf{x}} = \mathbf{f}(\mathbf{p}, \mathbf{x}(t))$) aux instants $t_j \in \{t_1, t_2, \ldots, t_N\}$ est disponible ; alors :

- Pour prouver qu'un pavé $[\mathbf{p}] \subseteq [\mathbf{P}_0]$ ne contient aucune solution, il suffit de vérifier que :

$$\exists t_j \in \{t_1, t_2, \ldots, t_N\}, \quad \mathbf{g}([\mathbf{p}],[\mathbf{x}_j]) \cap [\mathbf{y}_j] = \varnothing \tag{21}$$

- Pour prouver que $\forall \mathbf{p} \in [\mathbf{p}]$, $\mathbf{p}$ est une solution, il suffit de montrer que :

$$\forall t_j \in \{t_1, t_2, \ldots, t_N\}, \quad \mathbf{g}([\mathbf{p}],[\mathbf{x}_j]) \subseteq [\mathbf{y}_j] \tag{22}$$

- Si aucune des deux conditions précédentes n'est vérifiée, alors le pavé **[p]** est dit indéterminé.

Le test d'inclusion est alors donné par l'algorithme 6 [RRC03c].

Algorithme 6 : Test

Entrées : $[\mathbf{p}]$, $[\mathbf{x}_0]$, $[\mathbf{y}_1],\ldots,[\mathbf{y}_N]$, **f**, **g**

Test_cont = 0;

1. Pour k = 0 à N-1,
2. $[\mathbf{x}_{k+1}]^+ = \mathrm{IODE}(\mathbf{f}, [\mathbf{x}_k])$;
3. Si $\mathbf{g}([\mathbf{x}_{j+1}]^+) \cap [\mathbf{y}_{j+1}] = \varnothing$, retourner *non_ faisable_*;
4. Si $\mathbf{g}([\mathbf{x}_{j+1}]^+) \subseteq [\mathbf{y}_{j+1}]$, Test_cont = Test_cont +1;
5. Fin de la boucle ;
6. Si (Test_cont = N), retourner *_faisable_*;
   retourner *_indéterminé_*;

Sorties : [*_faisable _|_non_faisable_|_indeterminé_*])

La boucle (Pour k = 0 à N-1) est utilisée afin de tester si les sorties du modèle sont consistantes avec les données mesurées.

Exemple 3 : On considère le modèle de Lotka-Volterra présenté dans l'exemple 1 ; dans cet exemple on suppose que les paramètres a et c sont connus (a = 1, c = 1) mais b et d sont inconnus. Les pseudo-mesures sont obtenues en simulant le modèle avec $x_0 = (50, 50)^T$ et la vraie valeur du vecteur de paramètres $\mathbf{p}^* = (0.01,0.02)^T$. Le pas d'échantillonnage est h =

0.005, le nombre de points est $N = 1400$, les hypothèses sur le bruit de mesure sont les mêmes que dans l'exemple 1 et l'espace initial de recherche est $[\mathbf{P}_0] = [-1,1] \times [-1,1]$.

En supposant que l'état initial est parfaitement connu, SIVIA avec le test d'inclusion défini ci-dessus génère en 3 minutes sur un Pentium IV, 1.6GHz l'ensemble des pavés tracés sur la figure 11. L'erreur absolue est de 2.5%, soient :

$$p_1 \in [0.00975, 0.01025]$$

$$p_2 \in [0.0195, 0.0205]$$

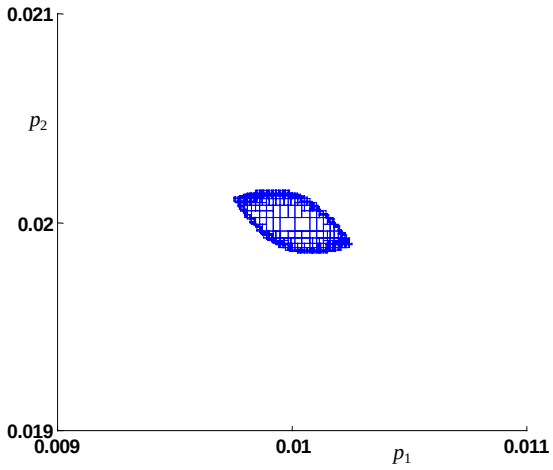

Figure 11 : Ensemble des pavés générés par SIVIA avec un état initial supposé connu

$\mathbf{p}^* = (0.01,0.02)^T$

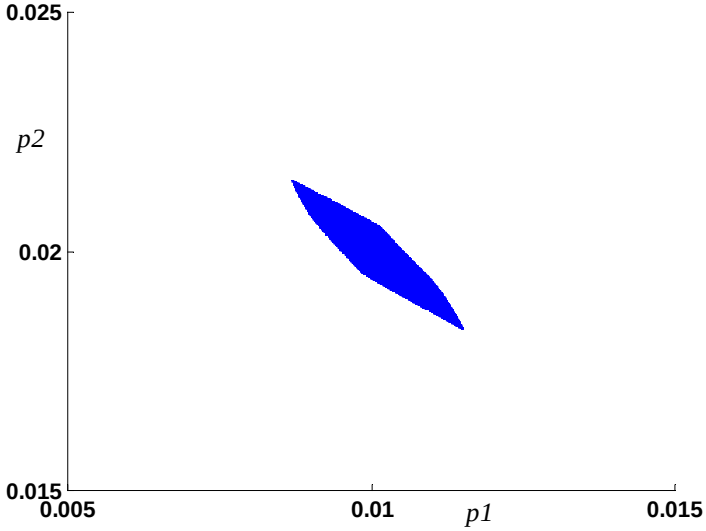

Figure 12 : Ensemble des pavés générés par SIVIA avec un état initial incertain

$\mathbf{p}^* = (0.01,0.02)^T$, $[\mathbf{x}_0] = [49;51] \times [49;51]$

Sur la figure 12, nous avons tracé l'ensemble des pavés générés par SIVIA en 1 heure, mais en supposant que l'état initial n'est pas parfaitement connu, $[\mathbf{x}_0] = [49;51]\times[49;51]$. On remarque alors que le temps de calcul augmente considérablement lorsque l'état initial est incertain. L'erreur absolue est de 14% sur *c* et de 7.7% sur *d*, soient :

$$p_1 \in [0.0086, 0.0114]$$

$$p_2 \in [0.01846, 0.02154]$$

Conclusion :

Dans cette section, nous avons combiné SIVIA avec les méthodes d'intégrations garanties des EDOs pour l'estimation des paramètres de modèles décrits par des EDOs. Les performances de la méthode ont été testées sur un exemple numérique. Nous avons alors constaté que le temps de calcul est d'autant petit que l'état initial est connu. L'incertitude sur les paramètres estimés est relativement petite.

5. Limitations

On considère le banc d'essai présenté dans le chapitre 3 ($2^{ème}$ partie) et dédié à la mesure des propriétés thermophysiques de matériaux. Le modèle utilisé dans le chapitre 3 est basé sur la représentation «quadripôles». Il permet de donner une forme explicite de la fonction de transfert liant la température de sortie à celle d'entrée. Le fait d'avoir une formulation explicite permet d'utiliser les algorithmes d'inversion ensembliste et les contracteurs.

Dans ce paragraphe, on va utiliser un modèle décrit par une équation différentielle, il est obtenu à partir d'une discrétisation spatiale, par la méthode des différences finies, de l'équation de la chaleur.

5.1. Modèle

5.1.1. Equation de la chaleur

On considère un matériau homogène, de propriétés isotropes, caractérisé par sa conductivité thermique λ, sa masse volumique ρ et sa capacité thermique massique C_p. L'équation de la chaleur est alors donnée par :

$$\rho C_p \frac{\partial T}{\partial t} = \lambda \frac{\partial}{\partial x}\left(\frac{\partial T}{\partial x}\right) \tag{23}$$

Dans la suite, nous allons utiliser la méthode des différences finies pour la discrétisation spatiale de l'équation (23) pour chaque couche du banc d'essai présenté dans le chapitre 3.

5.1.2. Discrétisation 1D

La discrétisation spatiale unidirectionnelle du dispositif étudié est présentée sur la figure 13.

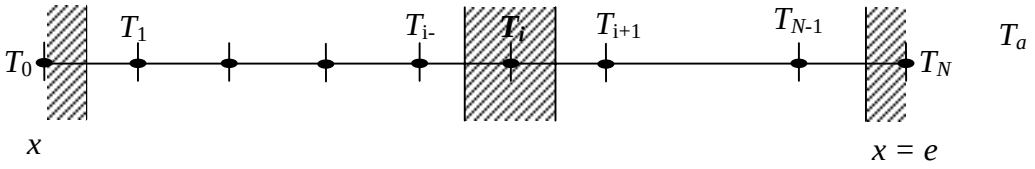

Figure 13 : Discrétisation 1D du dispositif

<u>Discrétisation dans la couche k</u>

Dans chaque couche, l'équation discrétisée de la conservation de l'énergie s'écrit sous la forme suivante :

$$\rho_k C_{pk} \Delta x_k \frac{\partial T_i}{\partial t} = \frac{\lambda_k}{\Delta x_k}\left(T_{i+1} - 2T_i + T_{i-1}\right) \tag{24}$$

où k est l'indice de la couche correspondante :

$k = 1$: pour le laiton

$k = g$: pour la graisse

$k = 2$: pour l'échantillon

$k = 3$: pour le cuivre

On obtient ainsi pour les différentes couches homogènes :

$$\begin{cases} \text{couche laiton} : \rho_1 C_{p1} \Delta x_1 \dfrac{\partial T_i}{\partial t} = \dfrac{\lambda_1}{\Delta x_1}\left(T_{i+1} - 2T_i + T_{i-1}\right) \\ \text{échantillon} : \rho_2 C_{p2} \Delta x_2 \dfrac{\partial T_i}{\partial t} = \dfrac{\lambda_2}{\Delta x_2}\left(T_{i+1} - 2T_i + T_{i-1}\right) \\ \text{couche cuivre} : \rho_3 C_{p3} \Delta x_3 \dfrac{\partial T_i}{\partial t} = \dfrac{\lambda_3}{\Delta x_3}\left(T_{i+1} - 2T_i + T_{i-1}\right) \end{cases} \tag{25}$$

<u>A l'interface entre deux couches de propriétés thermophysiques différentes</u>

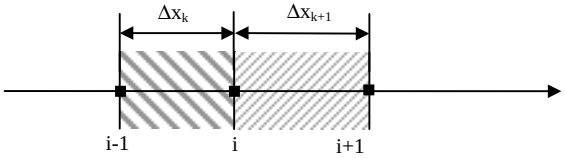

Figure 14 : Discrétisation à l'interface de deux couches

La discrétisation à l'interface de deux couches est donnée par :

$$\frac{1}{2}\left(\rho_k C_{pk}\Delta x_k + \rho_{k+1} C_{pk+1}\Delta x_{k+1}\right)\left(\frac{\partial T_i}{\partial t}\right) = \frac{\lambda_{k+1}}{\Delta x_{k+1}} T_{i+1} - \left(\frac{\lambda_{k+1}}{\Delta x_{k+1}} + \frac{\lambda_k}{\Delta x_k}\right) T_i + \frac{\lambda_k}{\Delta x_k} T_{i-1} \qquad (26)$$

Aux différentes interfaces de couches de propriétés différentes, on obtient :

Interface laiton-graisse :

$$\frac{1}{2}\left(\rho_1 C_{p1}\Delta x_1 + \rho_g C_{pg}\Delta x_g\right)\left(\frac{\partial T_i}{\partial t}\right) = \frac{\lambda_1}{\Delta x_1} T_{i-1} - \left(\frac{\lambda_1}{\Delta x_1} + \frac{\lambda_g}{\Delta x_g}\right) T_i + \frac{\lambda_g}{\Delta x_g} T_{i+1} \qquad (27)$$

Interface graisse-échantillon :

$$\frac{1}{2}\left(\rho_2 C_{p2}\Delta x_2 + \rho_g C_{pg}\Delta x_g\right)\left(\frac{\partial T_i}{\partial t}\right) = \frac{\lambda_g}{\Delta x_g} T_{i-1} - \left(\frac{\lambda_2}{\Delta x_2} + \frac{\lambda_g}{\Delta x_g}\right) T_i + \frac{\lambda_2}{\Delta x_2} T_{i+1} \qquad (28)$$

Interface échantillon-graisse :

$$\frac{1}{2}\left(\rho_2 C_{p2}\Delta x_2 + \rho_g C_{pg}\Delta x_g\right)\left(\frac{\partial T_i}{\partial t}\right) = \frac{\lambda_2}{\Delta x_2} T_{i-1} - \left(\frac{\lambda_2}{\Delta x_2} + \frac{\lambda_g}{\Delta x_g}\right) T_i + \frac{\lambda_g}{\Delta x_g} T_{i+1} \qquad (29)$$

Interface graisse-cuivre :

$$\frac{1}{2}\left(\rho_3 C_{p3}\Delta x_3 + \rho_g C_{pg}\Delta x_g\right)\left(\frac{\partial T_i}{\partial t}\right) = \frac{\lambda_g}{\Delta x_g} T_{i-1} - \left(\frac{\lambda_3}{\Delta x_3} + \frac{\lambda_g}{\Delta x_g}\right) T_i + \frac{\lambda_3}{\Delta x_3} T_{i+1} \qquad (30)$$

<u>Conditions aux limites :</u>

Les conditions aux limites sont données par les relations suivantes :

- Pour x = 0 : on a T(0,t) = T_0 (excitation imposée)
- Pour x = e : on a T(e,t) = T_N (température de la face arrière)

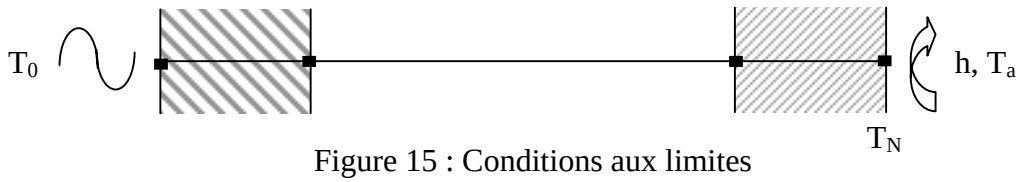

Figure 15 : Conditions aux limites

La discrétisation à *x = e* est donnée par :

$$\rho_3 C_{p3}\Delta x_3\left(\frac{\partial T_N}{\partial t}\right) = h\left(T_a - T_N\right) + \frac{\lambda_3}{\Delta x_3}\left(T_{N-1} - T_N\right) \qquad (31)$$

soit :

$$\rho_3 C_{p3} \Delta x_3 \left(\frac{\partial T_N}{\partial t} \right) = \frac{\lambda_3}{\Delta x_3} T_{N-1} - \left(\frac{\lambda_3}{\Delta x_3} + h \right) T_N + h T_a \tag{32}$$

où h est le coefficient d'échange global en face arrière et T_a est la température de l'environnement.

Les équations précédentes peuvent s'écrire sous forme matricielle :

$$\mathbf{C}\dot{\mathbf{T}} = \mathbf{A}\mathbf{T} + \mathbf{B}\mathbf{E} \tag{33}$$

où **C** est une matrice diagonale dont les éléments non nuls sont donnés par :

$$\begin{cases} C_{ii} = \dfrac{\rho_k C_{pk} \Delta x_k}{2} & \text{aux extrémités} \\ C_{ii} = \dfrac{\rho_k C_{pk} \Delta x_k + \rho_{k+1} C_{pk+1} \Delta x_{k+1}}{2} & \text{aux interfaces} \\ C_{ii} = \rho_k C_{pk} \Delta x_k & \text{aux interfaces} \end{cases}$$

A est une matrice tridiagonale symétrique :

$$\mathbf{A} = \begin{pmatrix} \frac{\lambda_1}{\Delta x_1} & \frac{\lambda_1}{\Delta x_1} & 0 & \cdots & & \\ \frac{\lambda_1}{\Delta x_1} & -2\frac{\lambda_1}{\Delta x_1} & \frac{\lambda_1}{\Delta x_1} & \ddots & & \\ 0 & \ddots & \ddots & \ddots & & \\ \vdots & \ddots & \frac{\lambda_1}{\Delta x_1} & -\left(\frac{\lambda_1}{\Delta x_1} + \frac{\lambda_2}{\Delta x_2} \right) & \frac{\lambda_2}{\Delta x_2} & \\ & & & \frac{\lambda_2}{\Delta x_2} & -2\frac{\lambda_2}{\Delta x_2} & \ddots \\ & & & & \ddots & \ddots \end{pmatrix}$$

et

$$\mathbf{B} = \begin{pmatrix} 1 & 0 \\ 0 & \vdots \\ \vdots & 0 \\ 0 & h \end{pmatrix}, \qquad \mathbf{E} = \begin{pmatrix} T_0 \\ T_a \end{pmatrix}$$

5.2. Estimation

Dans cette section, nous allons tenter d'identifier les paramètres $p_1 = \lambda_2$ et $p_2 = \rho_2 C_{p2}$ dans un contexte à erreurs bornées par inversion ensembliste en utilisant l'algorithme SIVIA avec le test d'inclusion défini par l'algorithme 6.

Pour réaliser le test d'inclusion, nous devons procéder à l'intégration numérique garantie de l'équation différentielle (33) où les paramètres p_1 et p_2 sont représentés par des intervalles. Lors de cette réalisation, plusieurs problèmes ont été rencontrés et ont conduit à une divergence de la taille des pavés contenant la trajectoire du vecteur d'état.

Ces problèmes sont dus d'une part à un choix inadapté du pas d'intégration et d'autre part à un problème de dépendance. Nous abordons ces questions dans la suite.

Choix du pas d'intégration

Le choix du pas de temps lors de l'intégration d'une équation différentielle est délicat. Dans certains cas, un mauvais choix peut entraîner la divergence de la solution ; dans le cas des méthodes garanties, on peut obtenir des pavés de taille infinie. Or, dans le cas de l'estimation d'état et de paramètres, ce problème est récurrent étant donné qu'aux instants où les mesures sont effectuées, doivent correspondre des temps intégrations de l'EDO. Deux approches sont alors possibles.

Une première approche, simple et intuitive, consiste à utiliser un pas d'intégration fixe correspondant au pas d'échantillonnage des données expérimentales. Pour tester la faisabilité de cette approche, nous avons intégré l'EDO (33) en supposant l'état initial et les paramètres p_1 et p_2 parfaitement connus. Tous les algorithmes d'intégration présentés dans le chapitre 4 ont divergé au bout de quelques pas de temps ; ils génèrent un ensemble de pavés dont la taille diverge au bout de quelques pas. Cette divergence s'explique par le fait qu'à certains endroits de la trajectoire, le pas de temps est trop petit par rapport à la dynamique de la réponse du système. La trajectoire est donc ″plate″, ce qui rend très pessimistes les fonctions d'inclusion centrées utilisées pour l'évaluation des coefficients de Taylor.

La seconde approche consiste à utiliser un pas d'intégration variable. Ce dernier est calculé lors de la première étape des algorithmes d'intégration, au moment de la vérification de l'existence de la solution de l'EDO. Un test de faisabilité de cette approche, dans les mêmes conditions que précédemment, état initial et paramètres p_1 et p_2 parfaitement connus, a fourni des résultats satisfaisants : la taille des pavés contenant la trajectoire de l'état ne diverge pas. Mais pour pouvoir procéder à l'estimation d'états, nous devons calculer une solution de l'EDO aux pas de temps de mesure des données expérimentales. La solution que nous avons retenue consiste à procéder à une autre intégration numérique garantie entre un instant d'intégration et l'instant de mesure, comme illustré sur la figure 16.

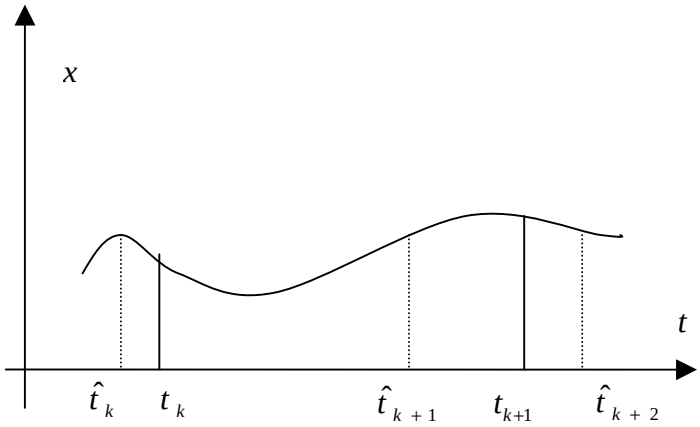

Figure 16 : Trajectoire d'une variable d'état ; $\hat{t}_k$: correspond à un temps intégration et t_k à un temps de mesure.

On note par $\hat{t}_k$ (k = 1,…, N) les instants déterminés par l'algorithme d'intégration de l'EDO et par t_k les instants qui correspondent aux mesures. La prédiction de l'état à l'instant t_k est alors obtenue en effectuant une interpolation garantie à l'aide d'un développement de Taylor de la solution de t_k à $\hat{t}_k$.

Problème de dépendance

La stabilité d'un schéma d'intégration numérique d'une EDO n'est possible que lorsque cette EDO est stable au sens de Lyapunov.

Dans notre cas, le modèle décrit par l'équation (33) est linéaire par rapport à l'état, il est alors stable si les pôles de la matrice **A** appartiennent au demi-plan complexe gauche. Supposons alors que le vecteur de paramètres **p** défini ci-dessus est incertain et est donné par :

$$\mathbf{p} \in \left[\underline{\mathbf{p}}, \overline{\mathbf{p}}\right] = \left[0.99 \cdot \mathbf{p}^*, 1.01 \cdot \mathbf{p}^*\right]$$

où $\mathbf{p}^*$ est la vraie valeur de **p.** On démontre facilement qu'il existe des valeurs de $\mathbf{p} \in \left[0.99 \cdot \mathbf{p}^*, 1.01 \cdot \mathbf{p}^*\right]$ telles que la matrice **A** est instable (présence de pôles positifs). Pour vérifier cela, il suffit de choisir $\mathbf{p} = \underline{\mathbf{p}}$ dans les éléments diagonaux de **A** et $\mathbf{p} = \overline{\mathbf{p}}$ dans les éléments non diagonaux. Cette instabilité est due au problème de dépendance, en effet, chacun des paramètres p_1 et p_2 apparaît plusieurs fois dans la matrice **A**, mais les éléments de **A** sont considérés indépendamment l'un de l'autre.

6. Conclusion

Dans ce chapitre, nous avons utilisé les techniques d'intégration numérique garantie des équations différentielles ordinaires afin de développer des estimateurs d'états pour des systèmes non linéaires à temps continu.

Nous avons proposé en premier lieu un estimateur basé sur l'approche *prédiction/correction* et semblable au filtre de Kalman. La phase de prédiction consiste à intégrer l'EDO à un instant t_{j+1} en partant de la solution à t_j ; cette étape produit alors un pavé $[\mathbf{x}_{j+1}]^+$. La correction consiste à contracter ce dernier pavé en utilisant la mesure disponible à t_{j+1}. On obtient alors un pavé $[\mathbf{x}_{j+1}]$ contenant les valeurs du vecteur d'état consistantes avec $[\mathbf{x}_j]$, la mesure à l'instant t_{j+1} et la borne d'erreur fixée *a priori*. Les performances de cet estimateur ont été testées sur un exemple numérique.

Dans un second temps, nous avons proposé un estimateur d'état à horizon glissant. Il est construit en introduisant des techniques de consistance dans l'estimateur proposé dans la première partie du chapitre. Les performances de cet algorithme ont été testées sur un système de bio-procédé. Les données utilisées sont obtenues à l'aide de simulations.

Dans la dernière partie du chapitre, nous avons étudié le cas où des paramètres inconnus doivent être estimés. Nous avons alors associé les techniques d'intégrations d'EDOs avec SIVIA. L'exemple numérique étudié montre que cette méthode permet de résoudre le problème d'estimation avec un temps de calcul raisonnable lorsque l'état initial est connu.

Par ailleurs, dans le cadre de l'estimation des paramètres thermophysiques du dispositif présenté dans le chapitre 3, nous avons constaté que le phénomène de dépendance rend le modèle instable. Il est alors nécessaire d'améliorer le modèle en tenant compte de ce problème.

Conclusions et perspectives

Ce travail a été dédié à l'estimation d'état et de paramètres de systèmes non-linéaires dans un contexte à *erreurs bornées*. L'estimation de paramètres (ou d'état) dans un tel contexte consiste à caractériser l'ensemble des valeurs *admissibles* du vecteur de paramètres. Une valeur est dite *admissible*, si la sortie du modèle calculée en utilisant cette valeur reste à l'intérieur des domaines des sorties mesurées.

Deux types de modèles ont été considérés : les modèles donnés par des fonctions non-linéaires explicites et les modèles décrits par des équations différentielles ordinaires. Dans les deux cas, les méthodes développées sont basées sur l'analyse par intervalles. Ce dernier outil, initialement développé par Moore [Moo66] pour le contrôle des erreurs d'arrondi dans les calculateurs numériques, est utilisé dans le domaine de l'estimation et de l'identification afin de calculer des solutions *globales* et *garanties*.

La première partie de cette thèse a été consacrée à l'identification de paramètres de modèles décrits par des fonctions non-linéaires à variables complexes. Ce problème est résolu en utilisant l'algorithme d'inversion ensembliste SIVIA [JW93a] [JKDW01]. Néanmoins, ce dernier nécessite l'évaluation de la sortie du modèle ; ceci peut être réalisé de deux manières différentes. Lorsque la fonction à variables complexes peut être décomposée à l'aide du calcul symbolique en deux parties réelle et imaginaire, l'évaluation de la sortie est effectuée à l'aide de l'arithmétique des intervalles réels. L'avantage de cette approche vient du fait qu'il est possible d'utiliser les bibliothèques mathématiques dédiées à l'analyse par intervalles réels ; par conséquence, aucun développement informatique lourd n'est nécessaire. Par ailleurs, les paramètres des fonctions, obtenues par décomposition explicite du modèle en parties réelle et imaginaire, sont souvent multi-occurrents, même lorsque ceux du modèle initial ne le sont pas. Le contracteur *propagation – rétropropagation*, utilisé afin de réduire les domaines de ces paramètres, n'est alors plus optimal.

La deuxième approche consiste à utiliser l'arithmétique des intervalles complexes afin d'évaluer la sortie du modèle ; ceci impose donc le choix d'une représentation géométrique de ces intervalles. D'autre part, étant donné que les modèles traités dans cette thèse sont fortement non-linéaires, nous avons opté pour l'utilisation de la représentation polaire où une donnée complexe incertaine est représentée par un *secteur*. Notre choix découle du fait que cette forme s'adapte mieux aux modèles non-linéaires. Néanmoins, cette représentation n'est pas fréquemment utilisée et on ne dispose pas de bibliothèques mathématiques utilisant cette forme.

Dans le chapitre 2, nous avons introduit la représentation polaire des intervalles complexes. Nous avons alors montré que l'opération de multiplication et de division de deux secteurs est définie d'une manière exacte ; d'où l'intérêt de l'employer lors de l'évaluation de fonctions

fortement non-linéaires. Mais l'inconvénient vient du fait que l'addition et la soustraction ne sont plus exactes.

La contribution du chapitre 2 est de présenter des algorithmes permettant d'effectuer l'addition de deux secteurs sans passer par une autre représentation. Le résultat obtenu est minimal, c'est à dire qu'il n'existe pas de secteur plus petit contenant la somme de *Minkowski* de ces deux secteurs. L'opération d'addition a été traitée comme un problème d'optimisation sous contraintes. Dans le chapitre 2, ce problème d'optimisation a été résolu analytiquement ; ceci est possible puisque le nombre de variables est assez réduit et les fonctions à minimiser sont simples. Les algorithmes proposés sont simples à mettre en œuvre. La mise à disposition de la bibliothèque développée est prévue.

Dans le chapitre 3, nous avons étudié deux applications pour lesquelles les modèles sont décrits par des fonctions explicites à variables complexes. Dans la première application, consistant à estimer des paramètres diélectriques, une décomposition analytique de la sortie du modèle est disponible ; nous avons donc comparé les deux approches : intervalles réels et intervalles complexes. Nous avons constaté que l'utilisation des intervalles complexes permet d'évacuer le problème de *dépendance*. Ainsi, le contracteur *propagation – rétropropagation* est optimal. Néanmoins, la représentation d'un ensemble par un secteur introduit un phénomène d'enveloppement. L'évaluation de la fonction d'inclusion naturelle de la sortie du modèle est alors pessimiste. On rappelle par ailleurs que ce problème ne s'est pas posé en utilisant uniquement des intervalles réels. Nous proposons alors, dans une étude ultérieure, de combiner ces deux approches.

D'autre part, l'identification des paramètres diélectriques du modèle d'Havriliak-Negami est un problème numériquement difficile et la caractérisation de l'approximation intérieure de l'ensemble solution n'a pas été réalisée. Nous proposons, dans une étude ultérieure, d'associer des techniques de recherche ponctuelle aux méthodes ensemblistes afin de pouvoir caractériser une telle approximation.

La seconde application étudiée dans le chapitre 3, concerne l'identification de paramètres thermophysiques de matériaux. Le modèle utilisé est fortement non-linéaire et à variable complexe et une décomposition en une partie réelle et une partie imaginaire à l'aide du calcul symbolique est très difficile à atteindre. L'utilisation de l'arithmétique des intervalles complexes est alors indispensable. Dans la thèse [Bra02], la représentation rectangulaire a été employée étant donné sa simplicité de mise en œuvre. Dans ce chapitre, nous avons montré qu'il est possible d'utiliser la bibliothèque développée dans le chapitre 2. Nous proposons alors, dans une étude ultérieure, d'associer cette bibliothèque aux algorithmes de projection ensembliste développés dans [Bra02] afin de caractériser le dispositif expérimental.

La seconde partie de cette thèse a été consacrée à l'estimation *garantie* d'état et de paramètres pour des systèmes décrits par équations différentielles ordinaires. Dans le chapitre 5, nous avons proposé un observateur d'état de structure semblable à celle du filtre de Kalman. Il comporte deux phases : la première, appelée prédiction, consiste à prédire l'état à un instant t_{j+1} sachant que l'état à l'instant t_j appartient à un pavé connu. Cette étape est réalisée en effectuant une intégration numérique garantie de l'équation différentielle d'état à l'aide de l'un des schémas d'intégration présentés dans le chapitre 4. La seconde étape consiste à

trouver l'ensemble des valeurs du vecteur d'état compatibles avec la mesure à l'instant t_{j+1} et avec la borne d'erreur fixée *a priori*. Elle est formulée comme étant un problème d'inversion ensembliste qui peut être résolu d'une manière garantie à l'aide de SIVIA. L'étape de correction consiste donc à contracter le domaine du vecteur d'état obtenu dans la phase de prédiction.

Dans une seconde partie du chapitre 5, nous avons proposé un observateur d'état à horizon glissant. Il est construit en introduisant des techniques de consistance dans un observateur fondé sur une approche prédiction/correction. Les performances de cet algorithme ont été testées sur un système de bio-procédés.

Dans la dernière partie du chapitre 5, les techniques d'inversion ensembliste ont été associées aux méthodes d'intégration garantie des équations différentielles afin d'étudier le cas où des paramètres inconnus doivent être estimés. Nous avons montré la faisabilité de la méthode sur un exemple numérique.

Par ailleurs, dans le cadre de l'identification des paramètres thermophysiques du dispositif présenté dans le chapitre 3, un second modèle décrit par des équations différentielles a été utilisé. Nous avons alors constaté que le phénomène de dépendance rend le modèle instable. Actuellement, nous étudions une amélioration de la méthode d'identification afin de tenir compte du phénomène de dépendance et éviter ainsi la divergence des résultats du modèle.

Bibliographie

[AH83] G. Alefeld and J. Herzberger. *Introduction to Interval Computations*. Academic Press, New York, 1983.

[BBC90] G. Belforte, B. Bona and V. Cerone. Parameter estimation algorithms for a set-membership description of uncertainty. *Automatica*, 26 : 887-898, 1990.

[BCH99] P. Bogaerts, J. Castillo and R. Hanus. A general mathematical modelling technique for bioprocesses in engineering applications. *Systems analysis modelling simulation*, 35 : 87-113, 1999.

[Bec03] Y. Becis - Aubry. *Contribution à l'estimation ensembliste des systèmes linéaires et non linéaires*. PhD dissertation, Université Henri Poincaré, Nancy, France 2003.

[BG97] F. Benhamou and L. Granvilliers. Automatic generation of numerical redundancies for nonlinear constraint dolving. Reliable Computing, 3(3): 335-344, 1997.

[BH99] E. W. Bai and Y. F. Huang. Convergence of optimal sequential outer bounding sets in bounded error parameter estimation. *Mathematics and Computers in Simulation*. 49(6) : 307-317, 1999.

[BH01] P. Bogaerts and R. Hanus. On-line state estimation of bioprocesses with full horizon observers. *Mathematics and computers in Simulation*, 56 : 425-441, 2001.

[BJKRW03] I. Breams, L. Jaulin, M. Kieffer, N. Ramdani and E.WALTER. Reliable parameter Estimation in Presence of Uncerainty. *13th IFAC Symposium on System Identification, SYSID2003*, Rotterdam, 2003.

[Bly79] A R. Blythe. *Electrical Properties of Polymers*. Cambridge Solid State Science Series, Cambridge University Press, 1979.

[BM98] M. Berz; K. Makino. Verified Integration of ODEs and Flows Using Differential Algebraic Methods on High-Order Taylor Models. Reliable Computing, 4(4) : 361-369, 1998.

[BMH94] F. Benhamou, D. McAllester and P. van Hentenryck. CLP (intervals) resvisited, In : M. Bruynooghe (ed.), *Proceedings of the International Logic Programming Symposium*, Ithaca, NY, 124-138.

[Bou03] A. Boudenne. *Etude expérimentale et theorique des propriétés thermophysiques de matériaux composites à matrice polymère*. PhD dissertation, Université Paris XII – Val de Marne, France, 2003.

[Bra02] I. Braems. *Méthodes ensemblistes garanties pour l'estimation de grandeurs physiques*. PhD dissertation, Université Paris-Sud, Orsay, France, 2002.

[BRG01] V. M. Becerra, P. D. Roberts and G. W. Griffths. Applying the extended Kalman filter to systems described by nonlinear differential-algebraic equations. *Control Engineering Practice*, 9 : 267-281, 2001.

[BRKW03] I. Braems, N. Ramdani, M. Kieffer et E. Walter. Caractérisation garantie d'un dispositif de mesure de grandeurs thermiques. *Journal Européen des Systèmes Automatisés*, 37(9) : 1129-1143, 2003.

[CA93] R. Coelho and B. Aladenize. *Les diélectriques*. Hermès, Paris, 1993.

[CGZ96] L. Chisci, A. Garulli, and G. Zappa. Recursive state bounding by parallelotopes. *Automatica*, 32 : 1049-1056, 1996.

[CGVZ98] L. Chisci, A. Garulli, A. Vicino and G. Zappa. Block recursive parallelotopic bounding in set membership identification. *Automatica*, 34 : 15-22, 1998.

[Cle87] J. C. Cleary. Logical arithmetic. *Future Computing Systems*, 2(2) : 125-149, 1987.

[CR97] T. Cendes and D. Ratz. Subdivision direction selection in interval methods for global optimization. *SIAM Journal of numerical Analysis*, 34: 922-938, 1997.

[CRRI04] Y. Candau, T. Raïssi, N. Ramdani and L. Ibos. Complex interval arithmetic using polar form. *Reliable Computing*, soumis en Juillet 2004.

[CWS97] G. Chen, J. Wang and L. S. Shieh. Interval Kalman filtering. *IEEE Transaction on Aerospace and Electronic Systems*, 33(1) : 250-258, 1997.

[Dar1876] G. Darboux. Sur les développements en série des fonctions d'une seule variable. *Journal des Mathématiques pures et appliquées*, 3ème série, t. II, 291-312, 1876.

[Dav87] E. Davis. Constraint propagation with interval labels. Artificial Intelligence, 32(3): 281-331, 1987.

[DC50] D. W. Davidson and R. H. Cole. Dielectric relaxation in glycerol. *J. Chem. Phys*., 18, 1950.

[DPW96] C. Durieu, B. Polyak and E. Walter. Trace versus determinant in ellipsoidal outer bounding with application to state estimation. *In Proceedings of the 13th IFAC World Congress*, Vol. I, San Francisco, pp 43-48.

[DWP00] C. Durieu, E. Walter and B. Polyak. Set-estimation with the trace criterion made simpler than with the determinant criterion. In *12th IFAC Symposium on System Identification*, Cd-Rom, Santa Barbara, Californie, 2000.

[DWP01] C. Durieu, E. Walter and B. Polyak. Multi-Input Multi-Output Ellipsoidal State Bounding. Journal of Optimization Theory and Applications, 111(2) : 273-303, 2001.

[Eij81] P. Eijgenraam. *The Solution of Initial Value Problems Using Interval Arithmeti*c. Mathematical Centre Tracts No. 144. Stichting Mathematisch Centrum, Amsterdam, 1981.

[FA03] J. M. Flaus et O. Adrot. Estimation d'état sûre pour procédés non linéaires par méthodes ensemblistes – Application à un procédé biotechnologique. *Journal Européen des Systèmes Automatisés*, 37(9) : 1145-1161, 2003.

[FH82] E. Fogel and Y. F. Huang. *On the value of information in system identification – bounded noise case*. *Automatica*, 18 : 229-238.

[GCHK97] A. Griewank, G. F. Corliss, P. Hanneberger, G. Kirlinger, F. A. Potra and H. J. Stetter. High-order stiff ODE solvers via automatic differentiation and rational prediction. In: *Lecture Notes in Computer Science*, 1196 : 114-125, Spriger, Berlin, 1997.

[GFA03] H. V. Gonzàlez, J. M. Flauss and G. Acuna. Moving horizon state estimation with global convergence using interval techniques: application to biotechnological processes. *Journal of Process Control*, 13 : 325-336, 2003.

[GH72] I. Gargantini and P. Henrici. Circular arithmetic and the determination of polynomial zeros. *Numer. Math.* 18 : 305--320, 1972.

[GRH00] J. L. Gouzé, A. Rapaport and M. Z. Hadj-Sadok. Interval observers for uncertain biological systems. *Ecological Modelling*, 45–56, 2000.

[Han92] E. R. Hansen, *Global optimization using interval analysis.*, Marcel Dekker, New York, 1992.

[Her1878] Ch. Hermite. Extrait d'une lettre de M. Ch. Hermite à M. Borchardt sur la formule d'interpolation de Lagrange. *Journal de Crelle*, œuvres, tome III, 84(70): 432-443.

[HG01] M. Z. Hadj-Sadok and J. L. Gouzé. Estimation of uncertain models of activated sludge processes with interval observers. *Journal of Process Control*, 11 : 299-310, 2001.

[HH97] J. S. Havriliak and S. J. Havriliak. *Dielectric and Mechanical Relaxation in Materials*, Hanser, Munich, 1997.

[Ibo00] L. Ibos. *Contribution à l'étude de la pyroélectricité dans les polymères ferroélectriques pour capteurs intégrés*. PhD dissertation, Université Paul Sabatier, Toulouse, France, 2000.

[Jau00] L. Jaulin. Interval constraint propagation with application to bounded-error estimation. *Automatica*, 36 : 1547-1552, 2000.

[Jau02] L. Jaulin. Nonlinear bounded-error state estimation of continuous-time systems**.** *Automatica*, 38(2) : 1079-1082, 2002.

[JKBW01] L. Jaulin, M. Kieffer, I. Braems and E. Walter. Guaranteed nonlinear estimation using constraint propagation on sets. *International Journal of Control*, 74(18) : 1772-1782, 2001.

[JKDW01] L. Jaulin, M. Kieffer, O. Didrit and E. Walter. *Applied Interval Analysis*. Springer-Verlag, London, 2001.

[Jon83] A.K. Jonscher. *Dielectric relaxation in solids*. Ed. Chelsea dielectric Press, London (1983)

[JW93a] L. Jaulin and E. Walter. Guaranteed nonlinear parameter estimation from bounded-error data via interval analysis. *Mathematics and computers in simulation*, 35 : 1923-1937.

[JW93b] L. Jaulin and E. Walter. Set-inversion via interval analysis for nonlinear bounded-error estimation. *Automatica*, 29 : 1053-1064.

[Kah68] W. M. Kahan. A more complete interval arithmetic. *Lecture notes for a summer course* at the University of Michigan, 1968.

[Kal60] R.E. Kalman. A new approach to linear filtering and prediction problems. *Trans. ASME, J. Basic Engineering*, 35-45, 1961.

[Kea96] R. B. Kearfott. *Rigorous Global Search: Continuous Problems*. Kluwer, Dordrecht, Netherlands, 1996.

[KHN91] R. B. Kearfott, C. Hu and M. Nova. A review of preconditioners for interval Gauss-Seidel method. *Interval computations*, 1: 59-85, 1991.

[Kie98] M. Kieffer. *Estimation ensembliste par analyse par intervalles, application à la localisation d'un véhicule*. PhD dissertation, Université Paris-Sud, Orsay, France, 1998.

[KJW02] M. Kieffer, L. Jaulin and E. Walter. Guaranteed recursive non-linear state bounding using interval analysis. *International Journal of Adaptive Control and Signal Processing*, 16 : 193-218, 2002.

[Kra69] R. Krawczyk. Newton-Algorithm zur Bestimmung von Nullstellen mit Fehlershranken. Computing, 4 : 187-201, 1969.

[Kri74] N. Krier. Komplexe Kreisarithmetik. Z. Angew. Math. Mech. 54 : 225-226, 1974.

[LAG99] P. Lagonotte, Y. Bertin and J. B. Saulnier. Analyse de la qualité de modèles nodaux réduits à l'aide de la méthode des quadripôles. *Int. J. Therm. Sci*, 38 : 51-65, 1999.

[LB02] S. Lesecq et A. Barraud. Une approche factorisée et numériquement stable pour l'estimation ensembliste ellipsoïdale. *Journal Européen des Systèmes Automatisés*, 36(4) : 505-518, 2002.

[LG85] R. Lohner, J. W. v. Gudenberg. Complex Interval Division with Maximum Accuracy. In *Proceedings of the 7th IEEE Symposium on Computer Arithmetic*, 332-336, 1985.

[Lju79] L. Ljung. Asymptotic behavior of the extended kalman filter as a parameter estimator for linear systems. *IEEE Transactions on Automatic Control*, 24(1) : 36-50, 1979.

[Loh88] R. J. Lohner. Einschließung der Lösung gewöhnlicher Anfangs–und Randwertaufgaben und Anwendungen. PhD thesis, Universität Karlsruhe, 1988.

[Loh94] R. J. Lohner. *AWA: software for the computation of guaranteed bounds for solutions of ordinary initial value problems*. Institut fur Angewandte Mathematik, Universitat Karlsruhe, 1994. Available on :

ftp://iamk4515.mathematik.uni-karlsruhe.de/pub/awa

[Lue71] D. G. Luenberger. An introduction to observers. *IEEE Transaction on Automatic Control*, 16(2) : 596-602, 1971.

[MM95] H. Michalska and D. Q. Mayne. Moving horizon observers and observer-based control. *IEEE Transaction on Automatic Control*, 40 : 995–1006, 1995.

[MN02] D. G. Maksarov and J. P. Norton. Computationally efficient algorithms for state estimation with ellipsoidal approximations. *International Journal of Adaptive Control and Signal Processing*, 16(6) : 411-434, 2002.

[Moo62] R. E. Moore. *Interval arithmetic and automatic error analysis in digital computing*. PhD dissertation, Department of Computer Science, Stanford University, 1962.

[Moo66] R. E. Moore. *Interval analysis*. Prentice-Hall, Englewood Cliffs, NJ, 1966.

[Moo79] R. E. Moore. *Methods and Applications of Interval Analysis*. SIAM, Philadelphia, PA, 1979

[Moo88] R. E. Moore. *Reliability in computing*. Academic Press, San Diego, 1988.

[Ned99] N. S. Nedialkov. *Computing rigorous bounds on the solution of an initial value problem for an ordinary differential equation*. PhD thesis, University of Toronto, 1999.

[NJ99] N. S. Nedialkov and K. R. Jackson. An Interval Hermite-Obreschkoff Method for Computing Rigorous Bounds on the Solution of an Initial Value Problem for an Ordinary Differential EquationBy: Nedialko S. Reliable Computing, 5(3) : 289-310, 1999.

[NJP01] N. S. Nedialkov, K. R. Jackson and J. D. Pryce. An Effective High-Order Interval Method for Validating Existence and Uniqueness of the Solution of an IVP for an ODE. Reliable Computing, 7(6) : 449-465, 2001.

[Neu90] A. Neumaier. *Interval methods for systems of equations*. Cambridge University Press, Cambridge, 1990.

[Neu03] A. Neumaier. Taylor forms - use and limits. *Reliable Computing*, 9 : 43-79, 2003.

[Nic69] K. Nickel. Triplex-Algol and its applications. In E. Hansen, editor, *Topics in Interval Analysis,* pages 10-24, Clarendon Press, Oxford, 1969.

[PP98] M. S. Petkovic and L. D. Petkovic. Complex interval arithmetic and its applications. Wiley-VCh, Berlin, 1998.

[PT93] L. D. Petkovic and M. Trajkovic. On some optimal inclusion approximations by disks. Interval Computation, 1 : 34-50, 1993.

[PNDW04] B. T. Polyak, S. A. Nazin, C. Durieu and E. Walter . Ellipsoidal parameter or state estimation under model uncertainty. *Automatica*, 40(7) : 1171-1179, 2004.

[Ral81] L. B. Rall. Automatic differentiation: Techniques and Applications. Lecture Notes in *Computer Science*. Springer-Verlag, Berlin, 1981.

[RC96] L. B. Rall and G. F. Corliss. Introduction to automatic differentiation, *In: Computational Differentiation: Techniques, Applications, and Tools*. SIAM: Philadelphia, 1-18, 1996.

[RAT92] D. Ratz, Automatische Ergebnisverifikation bei globalen Optimierungsproblemen, PhD dissertation, Université Karlsruhe, 1992.

[RFC97] G. Romero, J. M. Flaus and A. Cheruy. Semiqualitative modelling of bioprocesses. *Journal of Mathematical Modelling,* 3(3) : 246-264, 1997.

[Rih94] R. Rihm. Interval methods for initial value problems in ODEs. In *Topics in Validated computations: proceedings of the IMACS-GAMM International Workshop on Validated Computations*. University of Oldenburg, J. Herzberger, ed. Elsevier Studies in Computational Mathematics. Elsevier, Amsterdam, New York, 1994.

[RIRC03] T. Raïssi, L. Ibos, N. Ramdani and Y. Candau. Analyse de spectres de relaxation diélectrique par arithmétique d'intervalles. *Numelec'2003,* octobre 2003, Toulouse.

[RIRC05] T. Raïssi, L. Ibos, N. Ramdani and Y. Candau. Analyse de spectres de relaxation diélectrique par inversion ensembliste : une première approche. *Revue Internationale de Génie Electrique. Numéro spécial,* à paraître.

[RIRC04] T. Raïssi, L. Ibos, N. Ramdani and Y. Candau. Guaranteed method for the estimation of dielectric relaxation models parameters. *8th IEEE International Conference on Solid Dielectrics*, 111-114, 2004.

[RL71] J. Rokne and P. Lancaster. Complex interval arithmetic. *ACM,* 14 : 111-112, 1971.

[RR88] H. Ratschek and J. Rokne, *New computer methods for global optimization*. Ellis Horwood, Chichester, UK, 1988.

[RRC03a] T. Raïssi, N. Ramdani and Y. Candau. Estimation d'état pour des systèmes décrits par des équations différentielles non linéaires dans un contexte à erreurs bornées. *Journées Doctorales d'automatique,* Valenciennes, 2003.

[RRC03b] T. Raïssi, N. Ramdani and Y. Candau. State estimation for non-linear continuous systems in a bounded error context. *13th IFAC Symposium on System Identification, SYSID-2003***,** pp. 1725-1730, Rotterdam, 2003.

[RRC03c] T. Raïssi, N. Ramdani and Y. Candau. Parameter estimation for non-linear continuous-time systems in a bounded error context. *42nd IEEE Conference on Decision and Control*, pp. 2240-2245, Hawaii, 2003.

[RRC04] T. Raïssi, N. Ramdani and Y. Candau. Set membership state and parameter estimation for systems described by nonlinear differential equations. *Automatica*, 40 : 1771–1777, 2004.

[RRC05] T. Raïssi, N. Ramdani and Y. Candau. Bounded error moving horizon state estimator for non-linear continuous-time systems : application to a bioprocess system. *Journal of Process Control*, à paraître.

[RRIC04] T. Raïssi, N. Ramdani, L. Ibos and Y. Candau. Analyse de propriétés diélectriques dans un contexte à erreurs bornées. *CIFA 2004, Session invitée Méthodes ensemblistes pour l'Automatique*, 22-24 Novembre 2004, Douz, Tunisie.

[RRL01] C. V. Rao, J. B. Rawlings, and J. H. Lee. Constrained linear state estimation – a moving horizon approach. *Automatica*, 37(10) : 1619-1628, 2001.

[RUM88] S. M. Rump. *Algorithm for verified inclusions – theory and practice*. In [MOO88] : 109-126, 1988.

[Sch68] F. C. Schweppe. Recursive state estimation: unknown but bounded errors and system inputs. *IEEE Trans. On Automatic Control*, 13(1) : 22-28, 1968.

[Sch94] G. Schaumburg. Modern measurement techniques in Broadband Dielectric Spectroscopy. *Dielectric Newsletter*, Issue March 1994.

[TK98] E. Tang-Kwor. Contribution au développement de méthodes périodiques de mesure de propriétés thermophysiques des matériaux opaques. PhD Dissertation, Université Paris XII, Créteil, 1998.

[Wal75] D. L. Waltz. Generating semantic descriptions from drawing of scenes with shadows. In P. H. Winston, editor, *The psychology of computer vision*, 19-91, New York, 1975.

[Wan77] G. Wanner. On the integration of stiff differential equations. In*: Proceedings of the Colloquium on Numerical Analysis*, volume 37 of Intern. Ser. Numer. Math., pages : 209-226.

[WDM02] H. Wang, A. Degiovanni and C. Moyne. Periodic thermal contact : a quadrupole model and an experiment. *Int. J. Therm. Sci.*, 41 : 125-135, 2002.

[Wen64] R. Wengert. A simple automatic derivative evaluation program. *Communications of the ACM*, 7: 463-464.

[WK03] E. Walter and M. Kieffer. Interval analysis for guaranteed nonlinear parameter estimation, *13th IFAC Symposium on System Identification, SYSID-2003***,** pp. 259-270, Rotterdam, 2003.

[WP94] E. Walter et L. Pronzato. *Identification de modèles paramétriques*. Masson, Paris, 1994.

Résumé :

Cette thèse est dédiée au développement et à l'application de méthodes ensemblistes pour l'estimation d'état et de paramètres pour des systèmes non-linéaires. Deux types de modèles sont considérés : modèles donnés par des fonctions explicites à variables complexes et modèles décrits par des équations différentielles ordinaires (EDOs).

L'identification de paramètres de modèles décrits par des fonctions explicites est réalisée à l'aide des techniques d'inversion ensembliste par analyse par intervalles. Par ailleurs, les modèles utilisés sont à variables complexes ; dans ce cas l'évaluation de la sortie se fait à l'aide d'intervalles complexes. Dans ce travail, nous avons développé une arithmétique des intervalles complexes basée sur la représentation polaire. La multiplication et la division sont des opérations exactes, mais la somme et la différence ne le sont pas. Pour réduire le pessimisme introduit par ces dernières opérations, nous avons développé des algorithmes assurant les propriétés de *minimalité*.

Cette bibliothèque a été associée aux méthodes d'inversion ensembliste dans le cadre de l'estimation de paramètres de modèles diélectriques, d'une part, et pour l'identification de paramètres thermophysiques d'autre part.

Dans la deuxième partie de cette thèse, des algorithmes d'estimation d'état pour des systèmes décrits par des équations différentielles sont présentés. Ils permettent de fournir, à chaque instant, un ensemble contenant d'une manière garantie, toutes les valeurs du vecteur d'état compatibles avec les mesures et avec les bornes d'erreurs. Ces estimateurs sont basés sur des méthodes d'intégration garantie d'EDOs et sur l'inversion ensembliste.

Enfin, une technique d'estimation de paramètres de modèles décrits par des EDOs est proposée.

Abstract

This work is dedicated to the development and the application of set-membership methods for state and parameter estimation for non-linear systems. Two classes of models are investigated: models given by closed-forms expressions of complex variables and models described by ordinary differential equations (ODE).

For models described by closed-forms expressions, parameter identification is achieved by set inversion techniques through interval analysis. Furthermore, a complex interval arithmetics using polar forms is developed where multiplication and division operations are exact but no longer addition and subtraction. Thus, in order to ensure *minimality* property, new algorithms based on analytical constrained optimization are given. This new polar interval arithmetic toolbox is associated with set inversion and used for parameter estimation in the bounded error context, for the dielectric and thermal analyses of materials.

The second part of this work deals with set membership state and parameter estimation for non-linear systems described by ODEs. A new state estimator based on a predictor/corrector approach similar to the Kalman filtering, is given. This estimator relies on guaranteed numerical integration techniques and set inversion. Furthermore, a moving horizon state estimator is proposed. Finally, a parameter estimation technique is suggested for systems described by ODEs.