HAL
open science

Algorithmes numériques pour l'analyse topologique : Analyse par intervalles et théorie des graphes.

Nicolas Delanoue

To cite this version:

Nicolas Delanoue. Algorithmes numériques pour l'analyse topologique : Analyse par intervalles et théorie des graphes.. Automatique / Robotique. Université d'Angers, 2006. Français. NNT : . tel-00340999

HAL Id: tel-00340999

https://theses.hal.science/tel-00340999v1

Submitted on 24 Nov 2008

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Université d'Angers
ISTIA

Année 2006
N° d'ordre .

Algorithmes numériques pour l'analyse topologique.

Analyse par intervalles et théorie des graphes.

Thèse de doctorat

Spécialité : Automatique et Informatique Appliquée

École Doctorale d'Angers

soutenue le

le 14 décembre 2006

à l'ISTIA - Université d'Angers

par Nicolas Delanoue

devant le jury ci-dessous :

	Président du jury	
Michel Petitot	Rapporteur	*Professeur, LIFL - Laboratoire d'Informatique Fondamentale de Lille, Université de Lille.*
Frédéric Benhamou	Rapporteur	*Professeur, LINA - Laboratoire d'Informatique de Nantes Atlantique, Université de Nantes.*
Adam Parusinski	Examinateur	*Professeur, LAREMA - Laboratoire Angevin de Recherche en Mathématiques, Université d'Angers.*
Bertrand Cottenceau	Examinateur	*Maître de Conférences, LISA - Laboratoire d'Ingénierie des Systèmes Automatisés, Université d'Angers*
Luc Jaulin	Examinateur	*Professeur, E3I2 - Extraction et Exploitation de l'Information en Environnements Incertains, ENSIETA Brest*

Directeur de thèse : **Luc Jaulin, Bertrand Cottenceau**

Algorithmes numériques pour l'analyse topologique.

Analyse par intervalles et théorie des graphes.

soutenue le 14 décembre 2006

Nicolas Delanoue

Décembre 2006.

Résumé

Le travail présenté dans cette thèse concerne d'une part, l'étude qualitative d'ensembles et d'autre part, celui de l'étude de la stabilité d'un système dynamique. Les méthodes numériques proposées combinent le calcul par intervalles et la théorie des graphes.

De nombreux problèmes, comme l'étude de l'espace des configurations d'un robot, se ramènent à une étude qualitative d'ensembles. Ici, la "taille" de l'ensemble importe peu, ce qui compte, c'est sa "topologie". Les méthodes proposées calculent des invariants topologiques d'ensembles. Les ensembles considérés sont décrits à l'aide d'inégalités $\mathcal{C}^\infty$. L'idée maîtresse est de décomposer un ensemble donné en parties contractiles et d'utiliser l'homologie de Čech.

La seconde partie de la thèse concerne l'étude de point asymptotiquement stables des systèmes dynamiques (linéaires ou non). Plus largement, on propose une méthode pour approcher le bassin d'attraction d'un point asymptotiquement stable. Dans un premier temps, on utilise la théorie de Lyapunov et le calcul par intervalle pour trouver effectivement un voisinage inclus dans le bassin d'attraction d'un point prouvé asymptotiquement stable. Puis, on combine, une fois de plus, la théorie des graphes et les méthodes d'intégration d'équations différentielles ordinaires pour améliorer ce voisinage et ainsi construire un ensemble inclus dans le bassin d'attraction de ce point.

Mots-clés : Etude topologique d'ensembles, analyse par intervalles, méthodes numériques garantie, bassin d'attraction.

Abstract

This dissertation proposes, in a first place, numerical methods to compute qualitative properties of sets and, in a second place, algorithms to estimate the attraction domain of an equilibrium state. All the proposed approaches combine interval analysis and graph theory.

Many problems, as studing the configuration space of a robot, amount to analysing topological properties of a given set. In this thesis, we define methods to compute

topological invariants of a given set. Those sets are defined by $\mathcal{C}^\infty$ inequalities. The main idea is to decompose the given set into subsets that are proven contractible and to use Čech homology.

The second part of the thesis presents a method to estimate the attraction domain of an asymptotically stable equilibrium state x_∞. First, one uses interval analysis and Lyapunov theory to compute a neighboorhood of x_∞ included in the attraction domain. Then, we combine graph theory and inclusion methods of O.D.E. to improve this neighboorhood and estimate the attraction domain.

Keywords : topological analysis, interval analysis, reliable algorithms, attraction domain.

Remerciements

Luc Jaulin a accepté d'être mon directeur de thèse. Il m'a accompagné lors de ces premiers pas de chercheur qui constituent les trois années de thèse. Je l'en remercie très sincèrement. Je tiens aussi à exprimer toute ma reconnaissance à Bertrand Cottenceau pour son encadrement, son soutien constant tout au long de ma thèse.

Michel Petitot m'a fait l'honneur d'accepter d'être président de mon jury de thèse et rapporteur de celle-ci. Pour cela, ainsi que pour ses commentaires sur mon mémoire, je lui exprime ma profonde gratitude. J'ai eu la chance de lui rendre visite dans son laboratoire, cette entrevue a été riche d'enseignement. Je remercie Frédéric Benhamou d'avoir accepté d'être rapporteur de ma thèse, ainsi que pour ses jugements très pertinents sur mon manuscrit.

Je remercie Adam Parusinski pour avoir accepté de faire partie de mon jury de thèse. J'ai beaucoup appris en géométrie semi-algébrique réelle grâce à lui. C'est, en particulier, en suivant ses conseils que j'ai participé au trimestre sur la géométrie réelle[1].

Je remercie tous les chercheurs, enseignants et membres du personnel du laboratoire LISA pour leur amitié et leur aide pendant ces trois années de thèse. Notamment les gens qui ont partagés mes repas et mon bureau : Sébastien Lagrange, Laurent Houssin et Xavier Baguenard. Merci également à eux qui ont fait et font encore la richesse des discussions de la pause café. Je remercie vivement Olivier Le Gal que j'ai eu l'occasion de rencontrer lors des Rencontres Doctorales Mathématiques, pour les échanges de courriers électroniques concernant les géométries o-minimales.

Enfin, merci à Julie pour avoir toujours été à mes côtés.

[1] TRIMESTER ON REAL GEOMETRY September 12th - December 16th, 2005, Centre Emile Borel, institut Henri Poincaré.

Table des notations

Nous utiliserons les notations suivantes :

$\mathbb{IR}$	L'ensemble des intervalles compacts de $\mathbb{R}$.
$[x]$	un élément de $\mathbb{IR}$.
$\underline{x}$	la borne inférieure d'un intervalle $[x]$ de $\mathbb{IR}$.
$\overline{x}$	la borne supérieure d'un intervalle $[x]$ de $\mathbb{IR}$.
2^E	L'ensemble des parties de E.
Df	la différentielle de f.
$GL(\mathbb{R}^n)$	l'ensemble des matrices carré inversibles de taille $n \times n$.

Cette page est restée volontairement vierge.

Table des matières

1 Introduction **1**

2 Analyse par intervalles **3**

2.1 Introduction 3

2.1.1 Le calcul flottant et ses inconvénients 4

2.1.2 Le calcul par intervalles 8

2.2 Intervalles 10

2.2.1 Ensembles ordonnés 10

2.2.2 Intervalles de $\mathbb{R}^n$ 16

2.2.3 Fonction d'inclusion 20

2.3 Résolution de systèmes d'équations 28

2.3.1 Introduction 28

2.3.2 Méthode de bissection 28

2.3.3 Méthode de Newton par intervalles 30

2.4 Positivité 39

2.4.1 Introduction 39

2.4.2 Positivité stricte 39

2.4.3 Positivité ≥ 0 42

2.5 Conclusion 46

3 Invariants topologiques **47**

3.1 Rappels topologiques 47

3.1.1 Triangulation 51

3.1.2 Quelques invariants 55

3.1.3 Le groupe fondamental 56

3.2 Etat de l'art et contributions 57

3.2.1 Etat de l'art. 57

3.3 Connexité - Algorithme C.I.A. 67
3.3.1 Introduction 67
3.3.2 Condition suffisante pour qu'un ensemble soit étoilé. 68
3.3.3 Discrétisation 72
3.3.4 Algorithme - Nombre de composantes connexes par arcs. . . . 76
3.3.5 Limites de cette méthode 80
3.4 Type d'homotopie - Algorithme H.I.A. 81
3.4.1 Introduction 81
3.4.2 Discrétisation 81
3.4.3 Un algorithme 83
3.4.4 Limites de cette méthode 88
3.5 Applications 89
3.5.1 Exemple de robotique 90
3.5.2 Espace des configurations 90
3.5.3 Topologie de l'espace des configurations 92
3.5.4 Planification de trajectoires 93
3.5.5 Applications de HIA 96
3.5.6 Calcul par intervalles sur les variétés 99
3.6 Conclusion 102

4 Systèmes dynamiques 105
4.1 Introduction 105
4.2 Rappels - Systèmes dynamiques 106
4.3 Preuve de la stabilité 109
4.3.1 Théorie de Lyapunov 109
4.3.2 Vérification de l'existence d'une fonction de Lyapunov. 111
4.3.3 Algorithme 112
4.3.4 Exemple illustratif 113
4.4 Bassin d'attraction 115
4.4.1 Fonction d'inclusion du flot 115
4.4.2 Discrétisation 120
4.4.3 Algorithme 120
4.5 Conclusion 123

5 Conclusion et perspectives 125

Chapitre 1

Introduction

Classiquement, la preuve d'un résultat n'est obtenue que via une démarche intellectuelle. Il a fallu attendre 1976 pour que la conjecture des quatre couleurs[1] soit prouvée en partie de façon algorithmique. Ce résultat ainsi démontré a été élevé au rang de théorème des quatre couleurs. Cette preuve a été très controversée, la plupart des mathématiciens n'étant pas capables de vérifier l'exactitude des programmes utilisés. Tous les paramètres étaient représentés de façon exacte, ce qui rendait les résultats nécessairement garantis et conférait à la méthode le statut de preuve.

Ce n'était cependant pas le cas pour bon nombre d'algorithmes confrontés aux approximations liées à la représentation des réels. Ce n'est en effet qu'en 1998 qu'un tel algorithme a permis à Hales de prouver la conjecture de Kepler, on parle dès lors de *méthodes numériques garanties*. Cette avancée a été possible grâce notamment au calcul par intervalles. Cet outil permet de manipuler l'incertitude sur les nombres réels tout en garantissant le résultat. Le calcul par intervalles a aussi été employé pour montrer algorithmiquement l'existence de l'attracteur de Lorentz. Preuve qui a valu à Tucker le prix R.E Moore [5].

Ainsi, dans cette thèse, nous utilisons le calcul par intervalles afin d'élaborer des algorithmes garantis pour la planification de trajectoires, et plus généralement pour l'étude de propriétés topologiques d'ensembles décrits par des inégalités. Le lecteur notera que le calcul par intervalles ne se réduit pas à remplacer les opérations arithmétiques effectuées sur les nombres réels par leur équivalent sur les intervalles. Il est nécessaire de développer des algorithmes propres à cet outil.

[1] Cette conjecture fut publiée pour la première fois par Cayley en 1878.

L'originalité de cette thèse est de combiner la calcul par intervalle à la théorie des graphes. Dans toutes les méthodes présentées ici, on décompose l'espace de recherche et on crée un graphe dont les noeuds sont les pièces de ce recouvrement. Puis on exploite ce graphe pour en extraire des propriétés. Dans le chapitre 3, on crée par exemple un graphe qui contient autant de composantes connexes qu'une partie de $\mathbb{R}^n$ décrite par des inégalités. Une fois le processus accompli, compter le nombre de composantes connexes de cette partie de $\mathbb{R}^n$ se réduit à compter le nombre de composantes connexes du graphe ; ce qui est aisé algorithmiquement. Le processus s'apparente ainsi à une discrétisation. Dans le chapitre 4, le graphe créé est orienté, il nous permet de discrétiser le comportement d'un système dynamique continu décrit par une équation différentielle ordinaire. Les informations contenues dans ce graphe sont suffisamment riches pour construire un ensemble inclus dans le bassin d'attraction d'un point.

Cette thèse commence par une présentation du calcul par intervalles ainsi que quelques algorithmes qui seront les briques élémentaires pour les méthodes présentées dans les chapitres suivants.

Chapitre 2

Analyse par intervalles

2.1 Introduction

Il existe différentes manières de représenter les nombres pour les systèmes informatiques. En mettant de coté les problèmes de la taille des mémoires (que l'on considère toujours finie), n'importe quel nombre entier ou nombre rationnel peut être codé sur une machine. Les nombres réels sont quant à eux le plus souvent approchés au moyen d'une représentation à virgule flottante.

Les représentations à virgule flottante diffèrent seulement par le nombre de chiffres significatifs et la base choisis. Dans ce chapitre, on présente les limites du calcul classique avec des nombres à virgule flottante. On commence par montrer qu'il est possible de majorer l'erreur commise lors de calculs simples comme l'addition ou la multiplication. On verra par exemple, que lorsque l'on s'autorise uniquement l'utilisation de nombres à virgules flottantes, l'erreur commise pour évaluer $a + b \times c$, avec a, b et c des réels, est majorée par :

$$\epsilon(\epsilon + 2)(|a| + |b||c| + \epsilon(\epsilon^2 + 5\epsilon + 3)|b||c|) + \epsilon(\epsilon^2 + 5\epsilon + 3)|b||c| \tag{2.1}$$

où ϵ est un paramètre qui dépend de l'ensemble des flottants considérés. Etant donnés trois réels a, b et c, on est donc "capable" de majorer l'erreur commise lors de l'évaluation de $a + b \times c$. Maintenant, si de nouveau on ne s'autorise que l'utilisation des nombres flottants pour évaluer l'erreur 2.1, il ne faut pas faire d'erreur...

trouver une expression qui majore l'erreur est une chose
la calculer sans erreur en est une autre.

Cette dernière remarque nous confronte donc à une sorte de cercle vicieux. En effet, pour calculer l'erreur sans erreur, il faudrait faire une majoration de l'erreur lors du calcul de l'erreur...En définitive, la seule utilisation des nombres flottants ne suffit pas pour effectuer des calculs *garantis*.

Le plan pour ce chapitre est donc le suivant : on commence par montrer comment on peut calculer l'erreur (2.1). Puis cette étude nous mène tout naturellement à l'introduction du calcul par intervalle et à l'étude de ses propriétés dans la section 2.1.2.

2.1.1 Le calcul flottant et ses inconvénients

On commence par illustrer le calcul avec les nombres à virgule flottante à l'aide d'un exemple tiré du livre de Hansen [8]. Soit f la fonction rationnelle $f : \mathbb{R}^2 \to \mathbb{R}$ définie par l'expression suivante :

$$f(x, y) = 333.75y^6 + x^2(11x^2y^2 - y^6 - 121y^4 - 2) + 5.5y^8 + \frac{x}{2y}. \tag{2.2}$$

Le calcul de $f(77\,617, 33\,096)$ avec différents outils conduit :

- à l'aide de Matlab : $f \simeq 7.005 \times 10^{39}$,
- à l'aide de Mupad : $f \simeq -5.764607523 \times 10^{17}$,
- à l'aide d'un programme écrit en C :
  - Simple précision : $f \simeq 1.172603\ldots$,
  - Double précision : $f \simeq 1.1766039400531\ldots$,
  - Précision étendue : $f \simeq 1.176603940053178\ldots$.

Toutes ces évaluations sont fausses. Comme f est rationnelle, il est facile d'obtenir la valeur correcte :

$$f(77617, 33096) = -\frac{54\,767}{66\,192} \simeq -0.8273960599. \tag{2.3}$$

Pour comprendre pourquoi chacun de ces précédents calculs est faux, nous allons étudier comment les machines à calculer manipulent habituellement les réels[1]. Ils représentent une partie finie des réels avec ce que l'on appelle les *nombres à virgule flottante* ou les nombres flottants. Ces machines effectuent le plus souvent ces calculs

[1] Tous les calculateurs ne fonctionnent pas forcément avec des flottants.

en base 2, mais par souci de clarté, nous allons présenter ces nombres, en base 10 :

$$\begin{aligned} F_{n,q}^{10} \quad = \{ \quad & x \in \mathbb{R}, \;\; x = \pm m \cdot 10^p \\ & m \text{ un décimal sous la forme } 0, a_0 a_1 \ldots a_n, \\ & \text{où } a_i \in \{0, \ldots, 9\} \text{ et } 10^{-1} \leq m < 1 \\ & \text{et } p \text{ un entier à } q \text{ chiffres}\} \cup \{0\} \end{aligned}$$

Le nombre entier $a_0 a_1 \ldots a_n$ est appelé la mantisse, l'entier p l'exposant.

Exemple 2.1.1

Regardons l'ensemble $F_{2,1}$.

1. le nombre $17 \in F_{2,1}$; en effet $17 = 0.17 \cdot 10^2 (m = 0.17, p = 2)$.
2. le nombre $-0.2 \in F_{2,1}$; en effet $-0.2 = -0.20 \cdot 10^0 (m = -0.20, p = 0)$.
3. le nombre $0.01 \in F_{2,1}$; en effet $0.01 = 0.10 \cdot 10^{-1} (m = 0.10, p = -1)$.
   On aurait aussi pu l'écrire, $0.01 = 0.01 \cdot 10^0$, mais alors $10^{-1} \not\leq m$. Le fait d'imposer $10^{-1} \leq m$ permet d'obtenir l'unicité de représentation d'un réel par un flottant.
4. le nombre $0.01 \cdot 10^{-9} \notin F_{2,1}$; en effet, on a : $10^{-1} \not\leq m$.
5. le nombre $0.1 \cdot 10^{-9} \in F_{2,1}$; c'est le plus petit élément de $F_{2,1}$ strictement positif.
   Si l'on cherche à représenter un nombre plus proche de 0, on obtient une exception appelée `underflow`.
6. le nombre $990\,000\,000 \in F_{2,1}$, $990\,000\,000 = 0.99 \cdot 10^9 (m = 0.99, p = 9)$; c'est le plus grand nombre de $F_{2,1}$.
7. le nombre $990\,000\,000 - 17 = 989\,999\,983 \notin F_{2,1}$ bien que $990\,000\,000, 17 \in F_{2,1}$, $989\,999\,983 \simeq 0.99 \cdot 10^9 (m = 0.99, p = 9)$.
   Question : que se passe-t-il si l'on réitère ce calcul n fois ?
8. le nombre $990\,000\,000 + 17 \notin F_{2,1}$; on cherche à représenter un nombre trop grand. C'est un cas d'`overflow`.

La figure 2.1 montre la répartition des nombres flottants. En balayant les réels de 0 à 13 dans le sens croissant, on peut dire de façon peu rigoureuse que chaque fois que l'on passe un exposant de la base b, les flottants sont b fois moins nombreux.

FIG. 2.1 – Représentation des nombres flottants entre 0 et 13.

Remarque 2.1.2
La plupart des calculateurs actuels utilise la base 2 pour les nombres flottants (norme IEEE 754). Souvent, la mantisse contient 23 chiffres (un chiffre est soit 0, soit 1) *i.e.* 23 bits, 8 chiffres pour l'exposant (8 bits), et un bit pour le signe (0 pour positif et 1 pour négatif) :

signe	exposant	mantisse
1 bit	8 bits	23 bits

Avec les notations précédentes : l'ensemble des flottants utilisés par ces calculateurs est noté $F^2_{23,8}$. On peut tout de suite remarquer que cet ensemble ne contient pas $0,1$. De la même manière que $\frac{1}{3}$ n'est pas un élément de $F^{10}_{2,1}$. Montrons le par l'absurde, si $\frac{1}{3} \in F^{10}_{2,1}$, alors il existe m et p tels que $\frac{1}{3} = m \cdot 10^p$ avec $m = 0,a_1a_2..$ Par conséquent, il existe m' et $p' > 0$ tels que $\frac{1}{3} = \frac{m'}{10^{p'}}$, donc $10^{p'} = 3m'$, et donc $10^{p'} = 0$ modulo 3, ce qui est absurde (car $10^{p'} = 99\ldots99 + 1$).

Comme illustré par la figure 2.1, lorsque l'on approche un nombre réel x par un flottant x', l'erreur commise dépend de x. On obtient donc une erreur relative d'approximation donnée par la proposition suivante.

Proposition 2.1.3 (Erreur lors de l'approximation d'un réel à l'aide d'un flottant)
L'approximation d'un nombre réel x par un flottant x' de $F_{n,q}$ est à précision relative ;

$$\frac{|x - x'|}{|x|}\frac{\Delta x}{|x|} \leq 10^{1-n}. \tag{2.4}$$

Preuve : Voir l'Annexe C.

□

On notera par $\epsilon = 10^{1-n}$. Dans la suite, on se propose de majorer l'erreur sur une somme et sur un produit lorsque l'on utilise des flottants.

Propriétés 2.1.4 (Erreur lors de la somme de flottants)
Soient x et y deux nombres réels flottants, on note par $x +' y$ la somme flottantes, et par $\Delta(x + y)$ la distance qui sépare $x + y$ de $x +' y$. On a :

$$\Delta(x + y) \leq \epsilon(|x| + |y|). \tag{2.5}$$

Preuve : Voir Annexe C.

□

En général, les réels x et y ne sont pas des flottants, ils sont approchés par des nombres flottants x' et y'. Mais on peut tout de même majorer l'erreur commise en les additionnant avec des flottants.

Propriétés 2.1.5 (Erreur de la somme de deux réels en utilisant des flottants)
Si x et y sont des réels, l'erreur commise en les additionnant avec des flottants est donnée par

$$\Delta(x+y) \leq \epsilon(\epsilon+2)(|x|+|y|) \tag{2.6}$$

Preuve : Voir Annexe C.

□

Propriétés 2.1.6 (Erreur lors du produit de deux flottants)
Soient x et y deux nombres réels flottants, on note par $x \times' y$ le produit flottant, et par $\Delta(x \times y)$ la distance qui sépare $x \times y$ de $x \times' y$. On a :

$$\Delta(x \times y) \leq \epsilon|x||y|. \tag{2.7}$$

Propriétés 2.1.7 (Erreur lors du produit de deux réels en utilisant des flottants)
Soient x et y deux nombres réels avec une mantisse à n chiffres,

$$\Delta(x \times y) \leq \epsilon(\epsilon^2+5\epsilon+3)|x||y|. \tag{2.8}$$

Preuve : Voir Annexe C.

□

Exemple 2.1.8 (Premier calcul avec les flottants $F_{2,1}$)
Dans cet exemple, nous nous intéressons au calcul de f donné par l'expression : $f = 12,01 + 27,14 \times 7,026$. Il est possible de majorer l'erreur de ce calcul formellement en combinant les propositions précédentes. Avec a, b et c des réels, on a :

$$\begin{aligned}
a + b \times c &= a + b' \times' c' + e_{(b\times c)} \\
&= a' +' b' \times' c' + e_{(a+b'\times' c')} + e_{(b\times c)} \\
\Delta(a+b\times c) &\leq \underbrace{|e_{(a+b'\times' c')}|}_{\leq\epsilon(\epsilon+2)(|a|+|b'\times' c'|)} + \underbrace{|e_{(b\times c)}|}_{\leq\epsilon(\epsilon^2+5\epsilon+3)|b||c|} \\
&\leq \epsilon(\epsilon+2)(|a|+|b'\times' c'|) + \epsilon(\epsilon^2+5\epsilon+3)|b||c| \\
&\leq \epsilon(\epsilon+2)(|a|+|b||c|+e_{(b\times c)}) + \epsilon(\epsilon^2+5\epsilon+3)|b||c| \\
&\leq \epsilon(\epsilon+2)(|a|+|b||c|+\epsilon(\epsilon^2+5\epsilon+3)|b||c|) + \epsilon(\epsilon^2+5\epsilon+3)|b||c|
\end{aligned}$$

$$\Delta(a+b\times c) \leq \epsilon(\epsilon+2)|a| + \epsilon(\epsilon^4+7\epsilon^3+14\epsilon^2+12\epsilon+5)|b||c| \tag{2.9}$$

Le calcul, avec les flottants $F_{2,1}$, se déroule de la façon suivante :

	$12,01 + 27,14 \times 7,026$
- étape 1 : chacun des nombres réels est remplacé par son flottant le plus proche :	$12 + 27 \times 7,0.$
- étape 2 : on effectue la multiplication :	$12 + 189.$
- étape 3 : le résultat obtenu est remplacé par son flottant le plus proche :	$12 + 190$ car $189 \notin F_{2,1}$.
- étape 4 : on effectue la somme :	$202.$
- étape 5 : le résultat obtenu est remplacé par son flottant le plus proche :	200 car $202 \notin F_{2,1}$.

Le résultat obtenu avec les flottants est relativement satisfaisant sachant que l'évaluation correcte de f est : $202,695\ 64$. Néanmoins, si nous n'avions pas le moyen de connaître f correctement, nous aurions dû utiliser (2.9) pour garantir que l'erreur faite sur le calcul de f est inférieure à :

$$\Delta f = 123,552\ 182\ 6 = \epsilon(\epsilon+2)|12,01| + \epsilon(\epsilon^4 + 7\epsilon^3 + 14\epsilon^2 + 12\epsilon + 5)|27,14||7,026| \tag{2.10}$$

Autrement dit, en effectuant le calcul sur $F_{2,1}$, le seule garantie que nous ayons est

$$\begin{array}{rcl} f & \in & [\ 200 - \Delta f\ ;\ 200 + \Delta f\ ]\ , \\ i.e.\ f & \in & [\ 76,4478174\ ;\ 323,5521826\ ]\ . \end{array} \tag{2.11}$$

Remarque 2.1.9
Cet exemple illustre la remarque qui avait été faite dans l'introduction. Ici, nous avons pu calculer Δf correctement, mais en pratique comment le calculer sans faire d'erreurs (par défaut) ? La section suivante propose une méthode qui permet de passer outre ce problème.

2.1.2 Le calcul par intervalles

Comme l'a montrée la section précédente 2.1.1, l'étude de l'erreur de manipulation des nombres réels avec des nombres flottants est une étude fastidieuse. De plus, elle ne se suffit pas : l'utilisation *seule* des nombres flottants ne permet pas de faire du calcul garanti. Il existe au moins un autre moyen d'obtenir un résultat *garanti* qui a les particularités suivantes :

- il nous apporte en général un meilleur résultat (un Δf plus petit).
- il ne nécessite pas d'étudier l'erreur à priori.

Appliquons et détaillons cette méthode sur l'exemple précédent et calculons un encadrement pour $f = 12,01 + 27,14 \times 7,026$:

– étape 1 : chacun des nombres réels est encadré par deux flottants de $F_{2,1}$:

$$\begin{array}{ccccc} 12 & \leq & 12,01 & \leq & 13 \\ 27 & \leq & 27,14 & \leq & 28 \\ 7,0 & \leq & 7,026 & \leq & 8,0 \end{array} \tag{2.12}$$

– étape 2 : on effectue la multiplication, on en déduit :

$$\begin{array}{ccccc} 27 \times 7,0 & \leq & 27,14 \times 7,026 & \leq & 28 \times 8,0 \\ 189 & \leq & 27,14 \times 7,026 & \leq & 224 \end{array} \tag{2.13}$$

– étape 3 : le résultat obtenu est remplacé par son flottant le plus proche, par défaut à gauche et par excès à droite :

$$\begin{array}{ccccc} 189 & \leq & 27,14 \times 7,026 & \leq & 224 \\ 180 & \leq & 27,14 \times 7,026 & \leq & 230 \end{array} \tag{2.14}$$

– étape 4 : on effectue la somme, on en déduit :

$$\begin{array}{ccccc} 12 + 180 & \leq & 12,01 + 27,14 \times 7,026 & \leq & 13 + 230 \\ 192 & \leq & 12,01 + 27,14 \times 7,026 & \leq & 243 \end{array} \tag{2.15}$$

– étape 5 : le résultat obtenu est remplacé par son flottant le plus proche, par défaut à gauche et par excès à droite :

$$190 \leq 12,01 + 27,14 \times 7,026 \leq 250 \tag{2.16}$$

On en déduit donc que : $f \in [190, 250]$.

Une autre façon de présenter ces calculs consiste à manipuler directement les intervalles. On s'autorisera donc, par exemple, l'écriture $[\ 27\ ;\ 28\ ] \times [\ 7,0\ ;\ 8,0\ ]$ qui correspond à l'étape 2. Le tableau suivant présente les différents calculs en utilisant directement les intervalles.

	Réel	Intervalle
étape 1	$12,01$	$[\,12\,;13\,]$
	$27,14$	$[\,27\,;28\,]$
	$7,026$	$[\,7,0\,;8,0\,]$
étape 2	$27,14 \times 7,026$	$[\,27\,;\,28\,] \times [\,7,0\,;\,8,0\,]$
		$= [\,189\,;\,224\,]$
étape 3	$27,14 \times 7,026$	$[\,180\,;\,230\,]$
étape 4	$12,01 + 27,14 \times 7,026$	$[\,12\,;\,13\,] + [\,180\,;\,230\,]$
		$= [\,192\,;\,243\,]$
étape 5	$f = 12,01 + 27,14 \times 7,026$	$[\,190\,;\,250\,]$

(2.17)

Les intérêts de cette méthode sont nombreux. Avec le calcul précédent, on peut affirmer que $f \in [190, 250]$. D'un point de vue plus général, le calcul par intervalles nous donne une approximation garantie de l'image directe d'un intervalle $[x]$ par la fonction $f : \mathbb{R} \to \mathbb{R}$. En définitive, on construit une fonction $[f]$ qui à un intervalle $[x]$ associe un intervalle $[f]([x])$ tel que

$$\forall [x] \text{ intervalle de } \mathbb{R}, f([x]) \subset [f]([x]) \tag{2.18}$$

Une telle fonction est qualifiée de *fonction d'inclusion* pour f. Dans la section suivante, on présente le calcul par intervalles et ses propriétés générales.

2.2 Intervalles

Un intervalle est classiquement défini comme un convexe de $\mathbb{R}$. La section présente étend, d'un certain point de vue, la notion d'intervalle aux ensembles qui admettent une structure de *treillis*. Après avoir ordonné $\mathbb{R}^2$, on pourra par exemple parler d'un intervalle de $\mathbb{R}^2$. On présentera aussi une façon de construire naturellement des intervalles sur un espace produit, et la topologie classiquement posée sur l'ensemble des intervalles de $\mathbb{R}^n$. On finira par discuter de différentes fonctions d'inclusion et de leurs qualités.

2.2.1 Ensembles ordonnés

Définition 2.2.1

Un *treillis* $(X, \leq)$ est un ensemble muni d'une relation d'ordre vérifiant :

$\forall\ x, y \in X, x \vee y \in X$ et $x \wedge y \in X$, où $x \wedge y$ est la borne inférieure de x et y, et

$x \vee y$ est la borne supérieure. Voir [36] et [37] pour plus de détails sur les treillis et les ensembles ordonnés.

Exemple 2.2.2

1. l'intervalle compact $[0, 1]$ de $\mathbb{R}$ muni de la relation d'ordre naturelle $\leq$ est un treillis.
2. l'ensemble des nombres réels $\mathbb{R}$ muni de cette même relation est aussi un treillis.
3. avec $n \in \mathbb{N}$, l'ensemble $\{1, \ldots, n\}$ muni de la relation $\leq$ est un treillis.
4. soit E un ensemble, l'ensemble des parties de E muni de l'inclusion $\subset$ est un treillis. Il est souvent noté 2^E ou $\mathcal{P}(E)$. Dans ce cas, les lois de composition internes $\cap$ et $\cup$ coïncident avec $\vee$ et $\wedge$. La figure 2.2 montre l'ensemble des parties de $\{a, b, c, d\}$ et illustre la relation d'ordre avec des arcs.

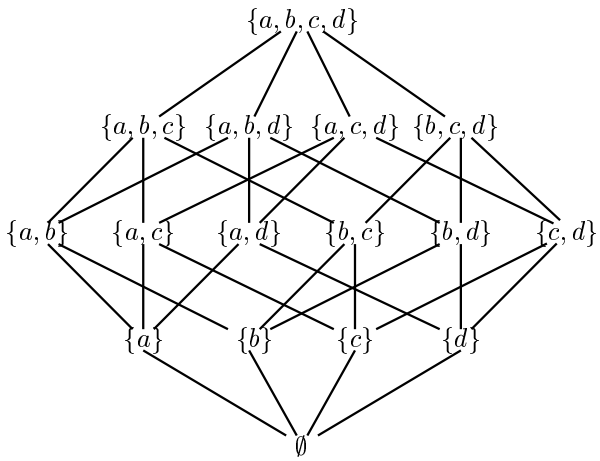


FIG. 2.2 – L'ensemble $2^{\{a,b,c,d\}}$ muni de l'ordre $\subset$ est un treillis.

Définition 2.2.3

Soit $(E, \leq)$ un treillis, et $\underline{x}, \overline{x}$ deux éléments de E tels que $\underline{x} \leq \overline{x}$, on note par $[\underline{x}, \overline{x}]$ le sous ensemble de E défini par

$$\{x \in E \mid \underline{x} \leq x \leq \overline{x}\}. \tag{2.19}$$

On dit que $[\underline{x}, \overline{x}]$ est un intervalle dont les *bornes* sont $\underline{x}, \overline{x}$. On note par $\mathbb{I}E$ l'ensemble des intervalles de E. C'est un sous-ensemble de 2^E.

Exemple 2.2.4

1. Avec $E = \mathbb{R}$, $[1, 2]$ est un intervalle.

2. $\{1, 2, 3\}$ est un intervalle de $\mathbb{N}$ alors que $\{1, 3\}$ non.
3. Dans $2^{\{a,b,c,d\}}$, l'intervalle $[\{b\}, \{a, b, d\}]$ est constitué des 4 éléments $\{b\}, \{a, b\}, \{b, d\}$ et $\{a, b, d\}$.
4. Les singletons sont des intervalles.

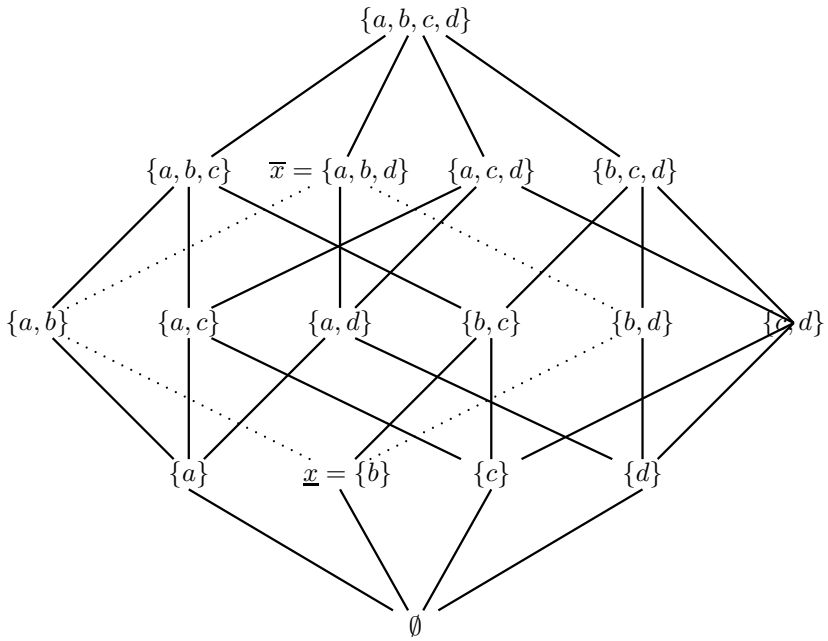


FIG. 2.3 – L'ensemble $[\underline{x}, \overline{x}] = [\{b\}, \{a, b, d\}]$ est un intervalle du treillis $(2^{\{a,b,c,d\}}, \subset)$ composé des éléments $\{b\}, \{a, b\}, \{b, d\}$ et $\{a, b, d\}$.

La proposition suivante permet de construire des intervalles sur un ensemble produit $E_1 \times E_2$. Cette construction nécessite que nous ayons défini au préalable des intervalles sur E_1 et sur E_2.

Proposition 2.2.5 *Soient $(E_1, \leq_1)$ et $(E_2, \leq_2)$ deux treillis, la relation d'ordre $\leq$ définie sur $E_1 \times E_2$ par :*

$$(x_1, x_2) \leq (y_1, y_2) \Leftrightarrow x_1 \leq_1 y_1 \text{ et } x_2 \leq_2 y_2$$

vérifie $\mathbb{I}(E_1 \times E_2) = \mathbb{I}(E_1) \times \mathbb{I}(E_2)$.

Preuve : Voir [36] ou [37].

□

Exemple 2.2.6

Ici, on s'intéresse à l'ensemble $E = \{1,2\} \times \{1,2\}$, avec $\{1,2\}$ muni de l'ordre naturel. Les figures 2.4 et 2.5 représentent respectivement l'ensemble des parties de E muni de l'inclusion et les intervalles de E. Par souci de clarté, l'élément (i,j) de E est renommé a_{ij}.

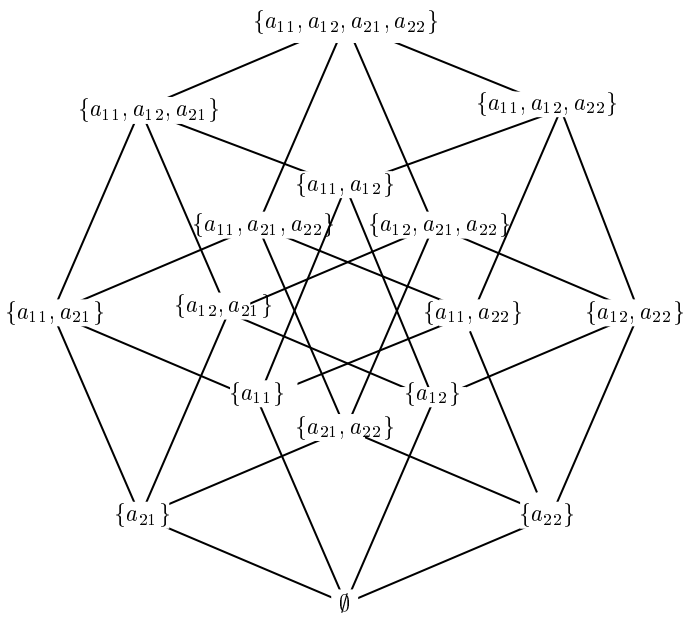


FIG. 2.4 – L'ensemble 2^E.

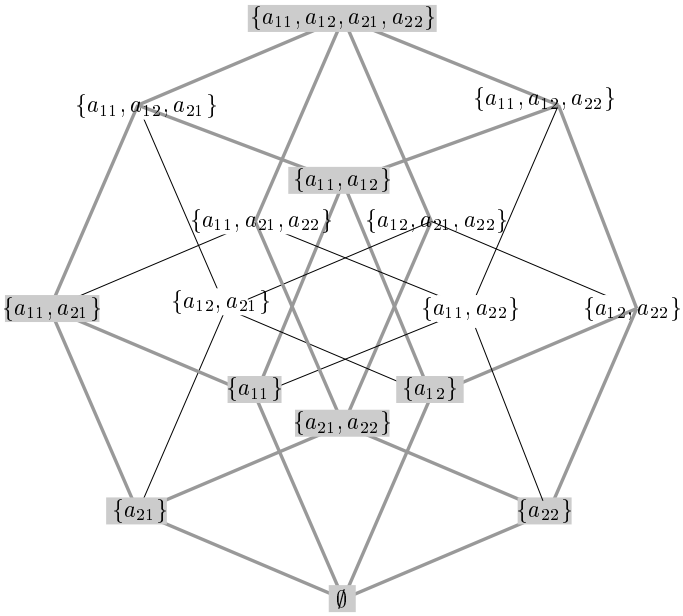


FIG. 2.5 – L'ensemble des intervalles de $\{1, 2\} \times \{1, 2\}$ est grisé. C'est un sous-ensemble de 2^E et un treillis.

Exemple 2.2.7

Soit E un ensemble. Un graphe sur E est une relation d'équivalence sur E, c'est donc une partie de $E \times E$. Soit G l'ensemble de tous les graphes [34] sur E, G est un treillis pour la relation d'ordre :

$$g_1 \leq g_2 \Leftrightarrow g_1 \subset g_2 \text{ avec } g_1, g_2 \in G. \tag{2.20}$$

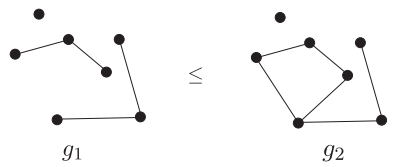


FIG. 2.6 – Le graphe g_1 est plus petit que g_2 dans $(\mathcal{G}, \leq)$.

En utilisant les notations de la figure 2.6, on peut dire que $[g_1, g_2]$ est un intervalle de $(\mathcal{G}, \leq)$, il contient 4 éléments :

FIG. 2.7 – Eléments de l'intervalle $[g_1, g_2]$ de $(\mathcal{G}, \leq)$.

Dans la suite de cette thèse, on utilisera cette représentation pour désigner l'intervalle $[g_1, g_2]$:

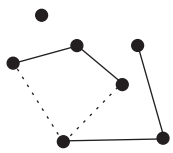


FIG. 2.8 – Représentation de l'intervalle $[g_1, g_2]$.

Comme nous avons pu le constater, la définition des intervalles d'un ensemble E est étroitement liée à la relation d'ordre que l'on considère sur E. L'exemple suivant illustre ce lien.

Exemple 2.2.8

On suppose ici que E est le plan Euclidien, et on considère les éléments $x = (x_1, x_2)$ et $y = (y_1, y_2)$ de E, on pose sur E les relations d'ordre suivantes :

$$x \leq_2 y \Leftrightarrow \begin{cases} x_1 \leq y_1 \\ x_2 \leq y_2 \end{cases} \tag{2.21}$$

$$x \leq_3 y \Leftrightarrow \begin{cases} x_1 + x_2 \leq y_1 + y_2 \\ x_1 - x_2 \leq y_1 - y_2 \end{cases} \tag{2.22}$$

Les intervalles définis par la relation $\leq_2$ et ceux définis par la relation $\leq_3$ sont différents et représentés sur la figure 2.9.

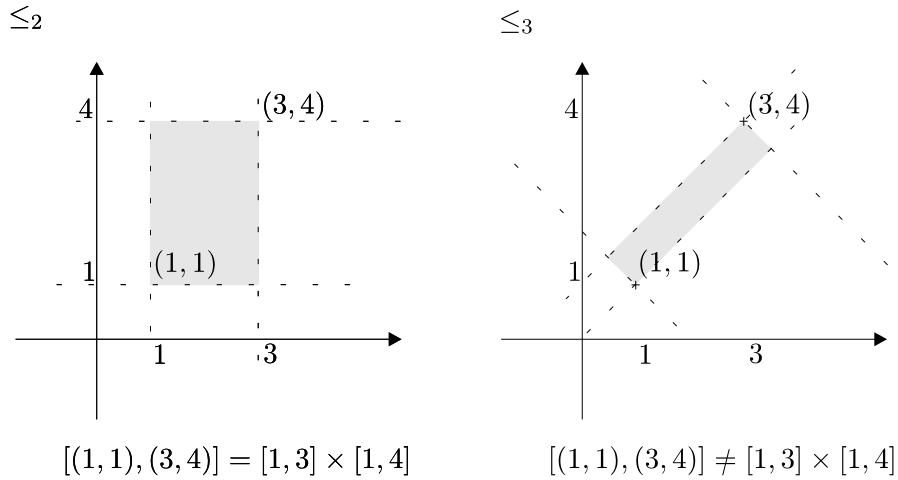


FIG. 2.9 – Exemples d'intervalles de $\mathbb{R}^2$.

Dans la suite, on considérera toujours que les intervalles de $\mathbb{R}^n$ sont ceux créés grâce à la relation définie par la proposition 2.2.5. Cette classe d'intervalles fera l'objet d'une étude plus approfondie dans la section 2.2.2. On verra, par exemple, que $\mathbb{IR}^n$ peut être muni d'une métrique ce qui nous permettra de parler de continuité, de vitesse de convergence de certains algorithmes ...

2.2.2 Intervalles de $\mathbb{R}^n$

Dans cette section, on présente les intervalles de $\mathbb{R}^n$. Ces intervalles sont classiquement appelés pavés. Ce sont les plus couramment utilisés. On note par $\mathbb{IR}$ l'ensemble des intervalles (compacts) de $\mathbb{R}$:

$$\mathbb{IR} = \{[\underline{x}, \overline{x}] \mid \underline{x}, \overline{x} \in \mathbb{R}, \ \underline{x} \leq \overline{x}\} \tag{2.23}$$

Par la suite, on prendra l'habitude de noter entre crochets les variables qui sont des intervalles, on dira par exemple : soit $[x]$ un intervalle de $\mathbb{IR}$. Ces sous-ensembles de $\mathbb{R}$ sont compacts et convexes et l'on pourra donc d'utiliser quelques résultats d'analyse convexe.

D'un point de vue géométrique, les intervalles réels peuvent être vus comme les éléments du demi-plan défini par $x_2 \geq x_1$ comme l'illustre la figure 2.10.

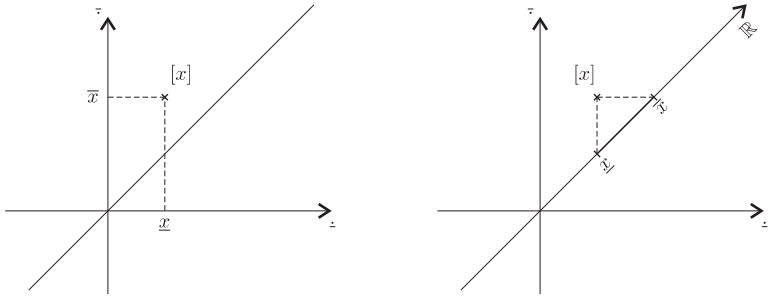


FIG. 2.10 – Les intervalles $\mathbb{IR}$ peuvent être vus comme les éléments du demi-plan défini par $x_2 \geq x_1$.

Exemple 2.2.9

L'intervalle $[1, \pi]$ est un élément de $\mathbb{IR}$ alors que $[2, 1]$ et $[1, \infty[$ n'en sont pas[2].

L'ensemble $\mathbb{IR}$ est muni de la topologie induite par la distance $q : \mathbb{IR} \times \mathbb{IR} \to \mathbb{R}^+$ définie par $q([a], [b]) = \sup\{|\underline{a} - \underline{b}|, |\overline{a} - \overline{b}|\}$. La distance q n'est rien d'autre que la distance de Hausdorff définie sur l'ensemble des compacts de $\mathbb{R}$ restreinte aux éléments de $\mathbb{IR}$. La figure suivante montre la boule de centre $[x]$ et de rayon ϵ.

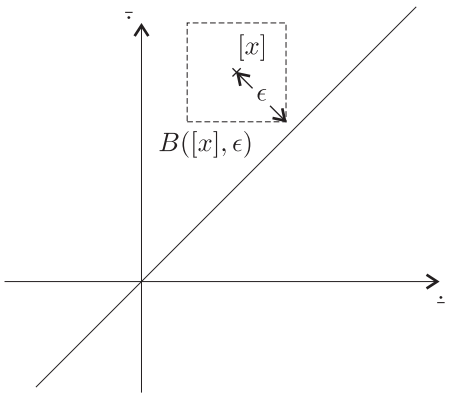


FIG. 2.11 – Boule de centre $[x]$ et de rayon ϵ.

En tant que partie de $2^{\mathbb{R}}$, l'ensemble $\mathbb{IR}^n$ hérite de la relation d'ordre $\subset$. La figure 2.12 illustre cette relation.

[2]L'intervalle noté $[2, 1]$ peut paraître dénué de sens, c'est pourtant l'opposé de l'intervalle $[-2, -1]$, si on s'autorise ces intervalles impropres alors l'ensemble des intervalles muni de l'opération $+$ devient un groupe ...Cette théorie est classiquement appelée théorie des intervalles modaux [52], [53], [54] et [55].

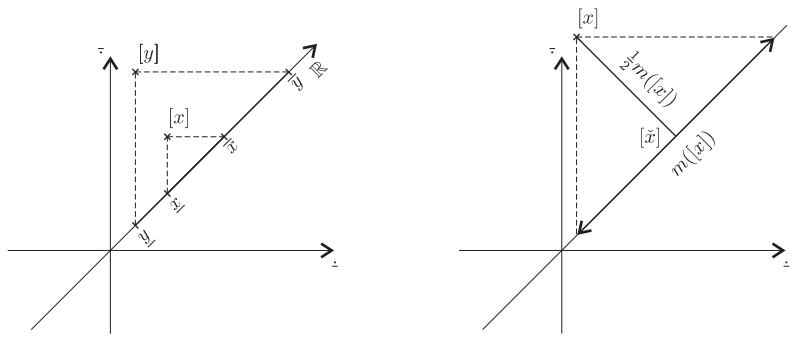


FIG. 2.12 – La figure de gauche illustre la relation d'ordre donnée par l'inclusion et celle de droite, la relation qui existe entre la taille d'un intervalle et la distance qui le sépare de son centre.

La proposition suivante donne une autre expression pour la distance q.

Proposition 2.2.10 *Si $[x]$ et $[y]$ sont deux intervalles de $\mathbb{IR}$, la distance $q([x],[y])$ est aussi donnée par :*

$$\inf\{q \in \mathbb{R}^+ \mid [\underline{x}, \overline{x}] \subset [\underline{y} - q, \overline{y} + q], [\underline{y}, \overline{y}] \subset [\underline{x} - q, \overline{x} + q]\}. \tag{2.24}$$

Proposition 2.2.11 *Une suite $\{[x]_n\}_{n\in\mathbb{N}} \in \mathbb{IR}^{\mathbb{N}}$ converge vers $[x] \in \mathbb{IR}$ si et seulement si les suites $(\underline{x}_n)_{n\in\mathbb{N}}$ et $(\overline{x}_n)_{n\in\mathbb{N}}$ de $\mathbb{R}^{\mathbb{N}}$ convergent respectivement vers $\underline{x}$ et $\overline{x}$.*

Preuve : La distance qui sépare deux intervalles $[x_1, x_2]$ et $[y_1, y_2]$ est donnée par la distance d_∞ qui sépare les points (x_1, x_2) et (y_1, y_2) du demi-plan. Donc $[x_n]$ converge vers $[x]$ si et seulement si les suites réelles $(\underline{x}_n)_{n\in\mathbb{N}}$ et $(\overline{x}_n)_{n\in\mathbb{N}}$ convergent respectivement vers $\underline{x}$ et $\overline{x}$.

□

Définition 2.2.12

Soit $\|.\|$ une norme sur $\mathbb{R}^n$ et $[a], [b]$ deux éléments de $\mathbb{IR}^n$. On considère la distance sur $\mathbb{IR}^n$, $q : \mathbb{IR}^n \times \mathbb{IR}^n \to \mathbb{R}^+$ définie par $q([a],[b]) = \sup\{\|\underline{a} - \underline{b}\|, \|\overline{a} - \overline{b}\|\}$. La figure 2.13 illustre cette distance en proposant la boule de centre $[v] \in \mathbb{IR}^2$ et de rayon r.

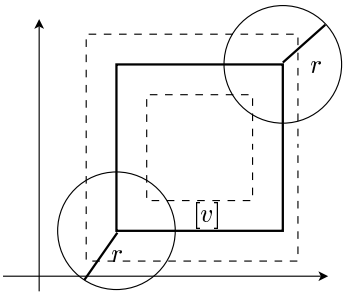


FIG. 2.13 – La boule de centre $[v] \in \mathbb{IR}^2$ et de rayon r contient tous les intervalles $[x]$ qui vérifient $\|\underline{x} - \underline{v}\| < r$ et $\|\overline{x} - \overline{v}\| < r$.

Remarque 2.2.13
Soit $\|.\|_\infty$ la norme sur $\mathbb{R}^n$ définie par $\|x - y\|_\infty = \max\limits_{i \in \{1,\ldots,n\}} |x_i - y_i|$ et $[a], [b]$ deux éléments de $\mathbb{IR}^n$. La fonction $q_\infty : \mathbb{IR}^n \times \mathbb{IR}^n \to \mathbb{R}^+$ définie par $q_\infty([a], [b]) = \sup\{\|\underline{a} - \underline{b}\|_\infty, \|\overline{a} - \overline{b}\|_\infty\}$ est une distance sur $\mathbb{IR}^n$. Cette distance q_∞ coïncide avec la distance de Hausdorff définie sur l'ensemble des compacts de $\mathbb{R}^n$. La figure 2.14 illustre ce fait. Comme dans la preuve de la proposition précédente, on considère les intervalles de $\mathbb{IR}^n$ $([\underline{x}, \overline{x}])$ comme des éléments de $\mathbb{R}^{2n}$ contraints par la relation $\underline{x} \leq \overline{x}$. De plus, lorsque $\mathbb{IR}^n$ est muni de la métrique d_∞, la boule fermée de centre $[v]$ et de rayon r est un treillis.

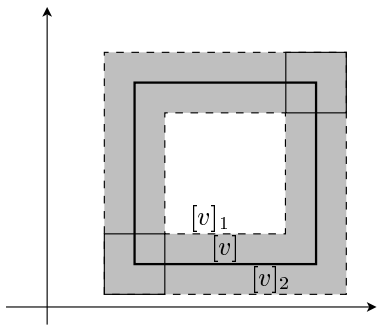


FIG. 2.14 – La boule de centre $[v]$ et de rayon r coïncide avec l'ensemble des intervalles compris (au sens de l'inclusion) entre $[v]_1$ et $[v]_2$.

Proposition 2.2.14 *Une suite* $\{[x]_n\}_{n \in \mathbb{N}} \in \mathbb{IR}^{n^{\mathbb{N}}}$ *converge vers* $[x] \in \mathbb{IR}^n$ *si et seulement si les suites* $(\underline{x}_n)_{n \in \mathbb{N}}$ *et* $(\overline{x}_n)_{n \in \mathbb{N}}$ *de* $\mathbb{R}^{n^{\mathbb{N}}}$ *convergent respectivement vers* $\underline{x}$ *et* $\overline{x}$.

Preuve : identique à la preuve de la proposition 2.2.11.

□

2.2.3 Fonction d'inclusion

Soit f une fonction de E dans F. Cette fonction induit naturellement une fonction de 2^E dans 2^F, encore notée f par :

$$\begin{array}{rccl} f: & 2^E & \rightarrow & 2^F \\ & X & \mapsto & f(X) = \{y \in F \mid y = f(x), x \in X\} \end{array} \tag{2.25}$$

On dit que $f(X)$ est l'*image directe* de X par f. Lorsque E contient une infinité d'éléments il est en général difficile de calculer algorithmiquement $f(X)$. Néanmoins, il est souvent possible de calculer un sous-ensemble de F qui contient $f(X)$.

Définition 2.2.15
Une fonction $[f] : \mathbb{I}E \rightarrow \mathbb{I}F$ qui vérifie

$$\forall [x] \in \mathbb{I}E, f([x]) \subset [f]([x]) \tag{2.26}$$

est qualifiée de *fonction d'inclusion* pour f.

La figure 2.15 nous montre l'évaluation d'un intervalle $[x]$ par la fonction f et par une fonction $[f]$ d'inclusion pour f. On a bien $f([x]) \subset [f]([x])$.

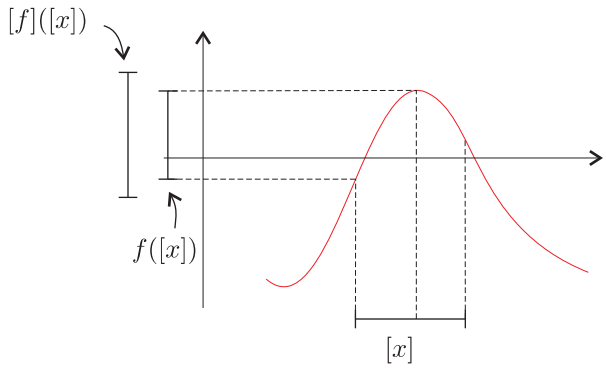


FIG. 2.15 – Evaluation de l'intervalle $[x]$ via une fonction $[f]$ d'inclusion pour f.

Exemple 2.2.16

- La fonction constante $[f] : \mathbb{IR} \rightarrow \mathbb{IR}$, définie par $[f] : [x] \mapsto [-1; 1]$ est une fonction d'inclusion pour la fonction sin.

– La fonction $[g] : \mathbb{IR} \to \mathbb{IR}$, définie par $[g] : [x] \mapsto [\mathrm{e}^{\underline{x}}; \mathrm{e}^{\overline{x}}]$ est une fonction d'inclusion pour la fonction : $\mathbb{R} \ni x \mapsto \mathrm{e}^x \in \mathbb{R}$.

Remarque 2.2.17
Il est possible d'ordonner partiellement l'ensemble des fonctions d'inclusion pour une fonction $f : E \to F$. Cela se fait en posant

$$[f]_1 \subset [f]_2 \Leftrightarrow \forall [x] \in \mathbb{I}E, [f]_1([x]) \subset [f]_2([x]). \tag{2.27}$$

Pour une fonction donnée f, la plus petite des fonctions d'inclusion de f est appelée *fonction d'inclusion minimale.*
Il est facile de définir la fonction d'inclusion minimale, par contre il est algorithmiquement difficile de la construire. Par construction algorithmique, on entend une méthode, qui étant donnés un intervalle $[x]$ et une fonction f, calcule le plus petit intervalle $[y]$ tel que $f([x]) \subset [y]$. Ce problème se ramène à m problèmes de maximisation et minimisation. Dans la pratique, on se contente de fonctions d'inclusion beaucoup plus grossières. Néanmoins, les fonctions d'inclusion que nous allons utiliser ont un comportement sympathique au voisinage des singletons comme le montrera la proposition 2.2.24.

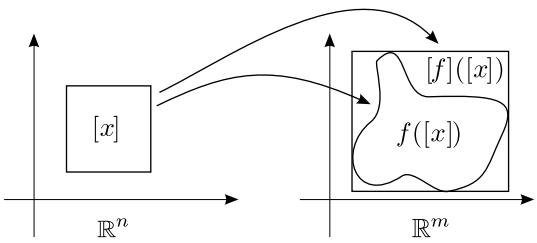


FIG. 2.16 – Fonction d'inclusion minimale.

Lemme 2.2.18 *Soient E et F deux ensembles ordonnés. Si f est une fonction croissante de E dans F, la fonction $[f]$ définie par*

$$[f]([\underline{x}, \overline{x}]) = [f(\underline{x}), f(\overline{x})] \tag{2.28}$$

est une fonction d'inclusion pour f.

Preuve : évident.

□

En pratique, les intervalles les plus utilisés sont les intervalles de $\mathbb{R}^n$. Dans cette section, nous rappelons quelques fonctions d'inclusion pour les fonctions usuelles comme l'addition, la soustraction ...Cette famille de fonctions d'inclusion forme ce qui est appelée l'*arithmétique des intervalles*[3].

Si $\star \in \{+, -, \times, \div\}$ et $[a], [b] \in \mathbb{IR}$ alors $[a] \star [b]$ est l'élément de $\mathbb{IR}$

$$\{a \star b, a \in [a] \text{ et } b \in [b]\} \tag{2.29}$$

Les différentes fonctions d'inclusion sont explicitées ici :

$$\begin{array}{rcl} [a] + [b] & = & [\underline{a} + \underline{b}; \overline{a} + \overline{b}] \\ [a] - [b] & = & [\underline{a} - \overline{b}; \overline{a} - \underline{b}] \\ [a] \times [b] & = & [\min\{\underline{ab}, \underline{a}\overline{b}, \overline{a}\underline{b}, \overline{a}\overline{b}\}; \max\{\underline{ab}, \underline{a}\overline{b}, \overline{a}\underline{b}, \overline{a}\overline{b},] \\ [a] \div [b] & = & [a] \times [1/\overline{b}; 1/\underline{b}] \end{array} \tag{2.30}$$

Remarque 2.2.19
Si $0 \in [b]$ alors $[a] \div [b]$ n'est pas définie, en effet l'ensemble $\{a \star b, a \in [a] \text{ et } b \in [b]\}$ n'est en général pas un élément de $\mathbb{IR}$.

Remarque 2.2.20
Il n'est pas difficile de construire des fonctions d'inclusion pour les fonctions usuelles telles que cos, sin, tan ...

On termine ce paragraphe avec un résultat qui nous permet de construire algorithmiquement une fonction d'inclusion $[f]$ lorsque f est définie par une expression.

Lemme 2.2.21 *Soient $f : E \to F$ et $g : F \to G$ deux fonctions. Si $[f]$ et $[g]$ sont, respectivement, deux fonctions d'inclusion pour f et g alors la fonction définie par $[f] \circ [g]$ est une fonction d'inclusion pour $f \circ g$.*

Preuve : évident.

□

En définitive, lorsqu'une fonction f est donnée par une expression, *e.g.* $f(x) = (\sin x - x^2 + 1)\cos x$, pour créer une fonction d'inclusion, on remplace chaque fonction qui compose f par son équivalent intervalle, *i.e.* $[f]([x]) = (\sin[x] - [x]^2 + 1)\cos[x]$.

[3] La paternité du calcul par intervalles est difficilement attribuable, on peut citer : R. C. Young (1931), M. Warmus (1956), T. Sunaga (1958), R. E. Moore (1959) Interval Analysis (1966).

Exemple 2.2.22

Cet exemple montre comment on calcule un intervalle qui contient l'image directe de $[0, \frac{1}{2}]$ par f où $f(x) = (\sin x - x^2 + 1)\cos x$.

$$\begin{aligned}
[f]([0;\tfrac{1}{2}]) &= (\sin[0;\tfrac{1}{2}] - [0;\tfrac{1}{2}]^2 + 1)\cos[0;\tfrac{1}{2}] \\
&= (\sin[0;\tfrac{1}{2}] - [0;\tfrac{1}{4}] + 1)\cos[0;\tfrac{1}{2}] \\
&= (\sin[0;\tfrac{1}{2}] + [-\tfrac{1}{4};0] + 1)\cos[0;\tfrac{1}{2}] \\
&= ([0;\sin\tfrac{1}{2}] + [\tfrac{3}{4};1])[\cos\tfrac{1}{2};1] \\
&= [\tfrac{3}{4};1+\sin\tfrac{1}{2}] \times [\cos\tfrac{1}{2};1] \\
&= [\tfrac{3}{4}\cos\tfrac{1}{2};1+\sin\tfrac{1}{2}] \\
&\subset [0.65818;1.4795]
\end{aligned} \tag{2.31}$$

C'est ce genre de calcul qui a donné naissance au *calcul par intervalles*. Au niveau de l'implémentation, on utilise des arbres pour représenter les fonctions. La fonction $(\sin x - x^2 + 1)\cos x$ est par exemple représentée grâce à l'arbre suivant :

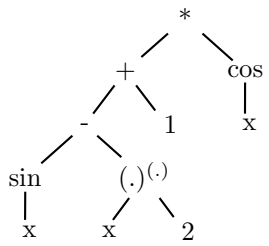


FIG. 2.17 – Représentation de la fonction f avec un arbre.

Chaque noeud de cet arbre représente une fonction. Ces noeuds sont aussi appelés les *atomes* de f.

Remarque 2.2.23

La dernière étape du calcul précédent a été réalisée avec des nombres à virgule flottante. On a veillé à arrondir le calcul par défaut pour la borne inférieure de l'intervalle et à arrondir par excès pour l'autre borne. Cette pratique est qualifiée d'*arrondi extérieur* (*outward rounding* en anglais). Il existe de nombreuses librairies écrites dans différents langages qui permettent de faire ce genre de calculs que le lecteur peut trouver sur le World Wide Web[4].

Dans l'exemple 2.2.22, on voit que lorqu'une fonction f est décrite par une expression formelle, il est possible de créer une fonction d'inclusion pour f connais-

[4] `http://www.cs.utep.edu/interval-comp/intsoft.html`

sant une fonction d'inclusion pour chacun des atomes qui la composent. La fonction d'inclusion, créée via ce processus, est appelée fonction d'inclusion *naturelle*.

On parle de *la* fonction d'inclusion naturelle de f mais c'est un abus de langage. La fonction d'inclusion naturelle de f n'est pas uniquement déterminée par la fonction f. En effet, comme on a pu le remarquer, la fonction d'inclusion naturelle de f dépend de l'expression de f. La proposition : *Si f admet deux expressions, alors les fonctions d'inclusions sont les mêmes* est, en général, fausse. Il suffit de prendre $f_1(x) = 0$ et $f_2(x) = x - x$, l'évaluation naturelle donne $[f_1]([0,1]) = [0,0]$ alors que $[f_2]([0,1]) = [-1,1]$. Dans ce cas, on préférera l'expression donnée par f_1. Ce phénomène est classiquement appelé le *problème de dépendance*. Dans l'évaluation par intervalles de f_2, on utilise la fonction d'évaluation de la différence $x-y$ définie sur $\mathbb{IR} \times \mathbb{IR}$, et on l'applique à la seule variable x. Le calcul par intervalles "oublie" que les variables x et y sont liées (ici $x = y$), d'où le nom de problème de dépendance.

Comme on l'a déjà précisé, le calcul par intervalles n'est pas capable, en général, de calculer exactement l'image directe d'un intervalle par une fonction. Il est seulement capable de calculer un intervalle qui contient cette image directe. Le résultat suivant donne le comportement asymptotique de l'erreur $q(f([x]), [f]([x]))$ qui est commise au voisinage des singletons.

Proposition 2.2.24 *Si $f : [D] \to \mathbb{R}$ est une fonction décrite par une expression finie ayant pour atomes $+$, $-$, $*$,* sin, cos, tan *et $[f]$ la fonction d'inclusion naturelle pour f, alors il existe un réel λ tel que*

$$\forall [x] \subset [D], \forall [x^0] \subset [x], q(f([x]), [f]([x])) \leq \lambda q([x], [x^0]). \tag{2.32}$$

Preuve : Voir le livre de Neumaier [30].

□

Il existe d'autres manières de créer algorithmiquement une fonction d'inclusion pour f. En particulier, il existe une méthode basée sur le théorème des valeurs intermédiaires qui nécessite une fonction d'inclusion de la différentielle Df. En pratique, on utilise la fonction d'inclusion naturelle comme fonction d'inclusion pour Df et par conséquent, cette fois encore, la fonction d'inclusion dépend de l'expression de f.

Proposition 2.2.25 *Soit $f : \mathbb{R}^n \to \mathbb{R}$ une fonction continuement différentiable et $[Df]$ une fonction d'inclusion pour la différentielle Df alors la fonction $[f] : \mathbb{IR}^n \to$*

$\mathbb{IR}$ *définie par*

$$[f] : [x] \mapsto f(x^0) + [Df]([x]) \cdot ([x] - x^0) \tag{2.33}$$

avec x^0 *un point de* $[x]$ *est une fonction d'inclusion pour* f*. Cette fonction d'inclusion est appelée en anglais* mean value form*. Dans le cas, où* x^0 *est le centre de* $[x]$*, cette fonction d'inclusion est appelée* forme centrée*.*

Preuve : Il suffit de montrer que l'ensemble $x^0 + [Df]([x]) \cdot ([x] - x^0)$ contient $f([x])$. La fonction f est supposée continuement différentiable, par conséquent pour tout $x \in [x]$, il existe ξ sur le segment délimité par x^0 et x tel que $f(x) = f(x^0) + Df(\xi)(x - x^0)$, on en déduit que $f(x) \in [f]([x])$.

□

Le comportement asymptotique de l'erreur $q(f([x]), [f]([x]))$ au voisinage des singletons est meilleur si $[f]$ est la forme centrée que si $[f]$ est la fonction d'inclusion naturelle. Ceci est donné par la proposition suivante qui a été conjecturée par Moore.

Proposition 2.2.26 *On note par* $[x^0]$ *un singleton inclus dans l'intervalle* $[x]$*. Si* f *est une fonction* $\mathcal{C}^1(\mathbb{R}^n, \mathbb{R})$ *décrite par une expression finie ayant pour atomes* $+$*,* $-$*,* $*$*,* sin*,* cos*,* tan*, et* $[Df]$ *la fonction d'inclusion naturelle pour* Df *telle qu'il existe* n *réels* λ_i *vérifiant* $q(Df_i([x]), [Df]_i([x])) \leq \lambda_i q([x], [x^0])$*, alors il existe un réel* λ *tel que*

$$q(f([x]), [f]([x])) \leq \lambda q^2([x], [x^0]). \tag{2.34}$$

Preuve : Il existe plusieurs preuves de ce résultat. Ici nous reprenons une preuve de Stahl [1] qui est sans doute la plus élémentaire. L'idée maîtresse de la preuve est de montrer l'encadrement :

$$f([x]) \subset [f]([x]) \subset f([x]) + [-1, 1]\lambda q^2([x], [x^0]). \tag{2.35}$$

On définit par *smig* la fonction $\mathbb{IR} \to \mathbb{R}$:

$$\begin{array}{rcl} smig & : & \mathbb{IR} \quad \to \quad \mathbb{R} \\ & & [x] \quad \mapsto \quad \begin{cases} \underline{x} & \text{si} \quad \underline{x} > 0 \\ \overline{x} & \text{si} \quad \overline{x} < 0 \\ 0 & \text{si} \quad 0 \in [x] \end{cases} \end{array} \tag{2.36}$$

Evidemment, on a pour tout $[x]$ de $\mathbb{IR}$:

$$[x] \subset smig([x]) + [-1, 1]m([x]) \tag{2.37}$$

On commence par montrer que pour tout $[x]$ de $\mathbb{IR}^n$ et pour tout x^0 de $[x]$ on a :

$$f(x^0) + smig([Df]([x]))([x] - x^0) \subset f([x]) \tag{2.38}$$

On doit donc montrer

$$\begin{cases} \overline{f(x^0) + smig([Df]([x]))([x] - x^0)} \leq \overline{f([x])} \\ \underline{f(x^0) + smig([Df]([x]))([x] - x^0)} \geq \underline{f([x])} \end{cases} \tag{2.39}$$

On montre seulement la première inégalité, la preuve de la seconde est basée sur le même raisonnement. Soit

$$\begin{array}{l} I^+ = \{i \mid \frac{df}{dx_i}([x]) \subset \mathbb{R}^{+*}, i \in \{1, \ldots, n\}\} \\ I^- = \{i \mid \frac{df}{dx_i}([x]) \subset \mathbb{R}^{-*}, i \in \{1, \ldots, n\}\} \end{array} \tag{2.40}$$

Si on note par $\{e_i\}_{i \in \{1,\ldots,n\}}$ une base de $\mathbb{R}^n$ et si i est un élément de I^+ alors pour tout $x \in [x]$, on a $\langle \nabla f(x), e_i \rangle > 0$. Autrement dit, la fonction f croit dans la direction e_i. L'idée pour montrer $\overline{f(x^0) + smig([Df]([x]))([x] - x^0)} \leq \overline{f([x])}$ est de prouver l'existence d'un $x \in [x]$ tel que

$$\overline{f(x^0) + smig([Df]([x]))([x] - x^0)} \leq f(x) \tag{2.41}$$

On pose $x \in \mathbb{R}^n$ défini par les coordonnées suivantes :

$$x_i = \begin{cases} \overline{x}_i & \text{si } i \in I^+ \\ \underline{x}_i & \text{si } i \in I^- \\ x_i^0 & \text{autrement.} \end{cases} \tag{2.42}$$

Par construction, x est un élément de $[x]$. On a alors $\overline{f(x^0) + smig([Df]([x]))([x] - x^0)}$

$$\begin{aligned} &= f(x^0) + \sum_{i=1}^{i=n} \overline{smig([Df]_i([x]))([x]_i - x_i^0)} \\ &= f(x^0) + \sum_{i \in I^+} smig([Df]_i([x]))(\overline{x}_i - x_i^0) + \sum_{i \in I^-} smig([Df]_i([x]))(\underline{x}_i - x^0) \end{aligned}$$

Donc $\forall \xi \in [x]$, on a $\overline{f(x^0) + smig([Df]([x]))([x] - x^0)}$

$$\leq \quad f(x^0) + \sum_{i \in I^+} Df_i(\xi)(\overline{x}_i - x_i^0) + \sum_{i \in I^-} Df_i(\xi)(\underline{x}_i - x_i^0) \tag{2.43}$$

D'après le théorème de Rolle, on en déduit pour tout $x \in [x]$ l'existence d'un $\xi_x \in [x]$ tel que $f(x) = f(x^0) + Df(\xi)(x - x^0)$. En particulier, pour $\xi = \xi_x$, on a :

$$\overline{f(x^0) + smig([Df]([x]))([x] - x^0)}$$

$$\begin{aligned}
&\leq f(x^0) + \sum_{i \in I^+} Df_i(\xi_x)(\overline{x}_i - x^0) + \sum_{i \in I^-} Df_i(\xi_x)(\underline{x}_i - x^0) && (2.44)\\
&= f(x^0) + \sum_{i=1}^{i=n} Df_i(\xi_x)(x - x^0) && (2.45)\\
&= f(x) && (2.46)\\
&\leq \overline{f([x])} && (2.47)
\end{aligned}$$

On montre maintenant l'encadrement 2.35,

$$\begin{aligned}
[f]([x]) &= f(x^0) + \sum_{i=1}^{i=n} [Df]_i([x])([x]_i - x_i^0)\\
&\subset f(x^0) + \sum_{i=1}^{i=n} \left(smig([Df]_i([x])) + [-1,1]m([Df]_i([x]))\right)([x]_i - x_i^0)\\
&\subset f(x^0) + \sum_{i=1}^{i=n} smig([Df]_i([x]))([x]_i - x_i^0) + [-1,1]m([Df]_i([x]))([x]_i - x_i^0)\\
&= \underbrace{f(x^0) + \sum_{i=1}^{i=n} smig([Df]_i([x]))([x]_i - x_i^0)}_{\subset f([x])} + \sum_{i=1}^{i=n} [-1,1]m([Df]_i([x]))([x]_i - x_i^0)\\
&\subset f([x]) + \sum_{i=1}^{i=n} [-1,1]m([Df]_i([x]))([x]_i - x_i^0)\\
&\subset f([x]) + \sum_{i=1}^{i=n} [-1,1]q([Df]_i([x]), [Df]_i([x^0]))q([x]_i, [x^0]_i)\\
&= f([x]) + \sum_{i=1}^{i=n} [-1,1]q([Df]_i([x]), Df_i([x^0]))q([x]_i, [x^0]_i)\\
&= f([x]) + \sum_{i=1}^{i=n} [-1,1]q([Df]_i([x]), Df_i([x]))q([x]_i, [x^0]_i)\\
&\subset f([x]) + \sum_{i=1}^{i=n} [-1,1]\lambda_i q^2([x]_i, [x^0]_i)
\end{aligned}$$

Par conséquent, il existe $\lambda \in \mathbb{R}$ tel que

$$[f]([x]) \subset f([x]) + [-1,1]\lambda q^2([x], [x^0])$$

On en conclut l'inégalité souhaitée.

□

Afin de motiver l'introduction du calcul par intervalles, nous allons étudier quelques algorithmes. En particulier, nous allons nous intéresser dans la section 2.3 à la résolution de systèmes à n équations avec m inconnues. Pour chacun de ces algorithmes, nous supposons l'existence d'une fonction d'inclusion pour toutes les fonctions rencontrées.

2.3 Résolution de systèmes d'équations

2.3.1 Introduction

Dans cette section, nous présentons deux algorithmes qui peuvent être utilisés pour chercher les zéros d'une fonction $f : \mathbb{R}^n \to \mathbb{R}^m$. Ce problème est équivalent à la résolution d'un système de m équations avec n inconnues. Les algorithmes décrits dans cette section sont basés sur un processus de découpage de l'espace de recherche (un peu à la manière de la division barycentrique des triangulations [48]). Le premier algorithme proposé, appelé *algorithme de bissection*, n'emploie que cette technique de découpage.

Dans un second temps, nous présentons un *algorithme de Newton par intervalles*. Son champ d'application est moins large que la méthode de bissection puisqu'il ne s'applique qu'aux fonctions $f : \mathbb{R}^n \to \mathbb{R}^n$ continuement différentiables. Néanmoins, on verra que cette méthode, appelé méthode de Newton par intervalles, est souvent capable d'isoler les zéros de f. Ces deux algorithmes seront utilisés dans la suite de la thèse. En particulier, l'algorithme de Newton par intervalles sera utilisé pour montrer l'existence et l'unicité d'un point d'équilibre d'un système dynamique au chapitre 4. Le premier algorithme sera employé au chapitre 3 pour montrer qu'un ensemble décrit par des inégalités est vide. L'idée n'est pas de présenter les meilleurs algorithmes mais de montrer leur essence, leurs propriétés et leurs limites.

2.3.2 Méthode de bissection

Soit f une fonction $f : [x] \to \mathbb{R}$, avec $[x]$ un intervalle de $\mathbb{R}^n$. On s'intéresse au problème suivant : approcher les zéros de f, c'est à dire approcher cet ensemble :

$$V_f = \{x \in [x] \mid f(x) = 0\}. \tag{2.48}$$

Par hypothèse, on a à notre disposition une fonction d'inclusion pour f. Si $[x_i]$ est un intervalle contenu dans $[x]$, deux situations peuvent alors se réaliser :

– soit $0 \notin [f]([x_i])$, alors $0 \notin f([x_i])$ et on peut conclure que $V_f \cap [x_i] = \emptyset$.

– soit $0 \in [f]([x_i])$, alors on ne peut rien en déduire ; l'équation $f(x) = 0, x \in D$ admet éventuellement une solution (ou plusieurs). Dans cette situation, l'intervalle courant $[x_i]$ est divisé en deux intervalles qui seront traités ultérieurement.

Nous allons maintenant décrire un algorithme qui crée une famille finie d'intervalles $\{[x_j]\}_j$ telle que $V_f \subset \bigcup_j [x_j]$ et telle que $\forall j, m([x_j]) < \epsilon$. Les zéros de f sont englobés dans de petits intervalles. Cet algorithme est présenté dans le cas où f est une fonction $\mathbb{R} \to \mathbb{R}$. Il n'est pas difficile de le généraliser au cas de fonctions $\mathbb{R}^n \to \mathbb{R}^m$.

Alg. 1 Algorithme de bissections.

Entrée: – un intervalle $[x]$ de $\mathbb{R}$,
– une fonction d'inclusion pour f, $[f] : \mathbb{IR} \to \mathbb{IR}$,
– Une précision ϵ.

Sortie: Une famille finie d'intervalles $\mathcal{P} = \{[x_j]\}_j$ telle que

$$V_f \subset \bigcup [x_j], \text{ et } \forall j, m([x_j]) < \epsilon \tag{2.49}$$

1: Initialisation : $\mathcal{P} \leftarrow \emptyset$, $\mathcal{P}_\Delta \leftarrow \{[x]\}$,
2: **tant que** $\mathcal{P}_\Delta \neq \emptyset$ **faire**
3: $[x_t] \leftarrow [x]$ où $[x] \in \mathcal{P}_\Delta$.
4: $\mathcal{P}_\Delta \leftarrow \mathcal{P}_\Delta - \{[x_t]\}$.
5: **si** $0 \in [f]([x_t])$ **alors**
6: **si** $m([x_t]) < \epsilon$ **alors**
7: $\mathcal{P} \leftarrow \mathcal{P} \cup \{[x_t]\}$;
8: **sinon**
9: $[x_{t_1}] \leftarrow [\ \underline{x_t}\ ;\ 0.5 \cdot (\underline{x_t} + \overline{x_t})\ ]$. et $[x_{t_2}] \leftarrow [\ 0.5 \cdot (\underline{x_t} + \overline{x_t})\ ;\ \overline{x_t}\ ]$.
10: $\mathcal{P}_\Delta \leftarrow \mathcal{P}_\Delta \cup \{[x_{t_1}]\} \cup \{[x_{t_2}]\}$;
11: **fin si**
12: **fin si**
13: **fin tant que**

A l'étape 9, on coupe un intervalle en deux intervalles. C'est de cette opération que vient le nom *bissection*. Cet algorithme est voisin d'un processus dichotomique. Dans le cas multidimentionnel $[x] = ([x_1], \ldots, [x_n]) \in \mathbb{IR}^n$, il existe de multiples stratégies pour sectionner un intervalle en deux. Généralement, on prévilégie la direction i pour laquelle $m([x_i])$ est la plus grande.

Cette méthode a l'avantage d'être très générale puisqu'elle peut être généralisée

à toute fonction $f : \mathbb{R}^n \to \mathbb{R}^m$ qui admet une fonction d'inclusion. Cependant, elle ne permet pas d'isoler les zéros de f : en effet, il se peut très bien qu'un intervalle $[x_j]$ fourni par l'algorithme contienne plusieurs solutions de $f(x) = 0$. (Voire même une infinité, voire aucune ...)
Pour certaines fonctions de $\mathbb{R}^n \to \mathbb{R}^n$, il est possible d'isoler les zéros de f grâce à une généralisation de l'algorithme de Newton que nous allons étudier dans la section suivante.

2.3.3 Méthode de Newton par intervalles

Introduction

L'itération de Newton est une méthode numérique classique de recherche des zéros d'un système d'équations $f(x) = 0$ où f est une fonction $\mathcal{C}^1(\mathbb{R}^n, \mathbb{R}^n)$. Si x est une approximation d'un zéro de ce système, la méthode de Newton affine cette approximation en prenant pour nouvelle valeur la solution y de l'équation linéarisée au voisinage de x :

$$f(x) + Df(x)(y - x) = 0 \tag{2.50}$$

Lorsque la différentielle $Df(x)$ est inversible on obtient :

$$y = x - Df(x)^{-1} f(x) \tag{2.51}$$

On appelle opérateur de Newton l'expression ainsi définie :

$$N_f(x) = x - Df(x)^{-1} f(x). \tag{2.52}$$

Il est défini sur l'ensemble des points réguliers de f. L'idée d'améliorer la qualité d'une approximation par ajout d'un terme correctif est fort ancienne. Cette méthode apparaît dans un contexte déjà très général dans *De analysis per aequationes numero terminorum infinitas* de 1669, où Newton considère des équations polynomiales et utilise une technique de linéarisation.

D'autres noms sont associés à cette méthode : Joseph Raphson et Thomas Simpson. En 1690, Raphson publie *Analysis aequationum universalis* dans lequel il présente une nouvelle méthode de résolution des équations polynomiales. En 1740, Simpson introduit une "nouvelle méthode de résolution des équations" en utilisant la "méthode des fluxions" (c'est à dire les dérivées) dans *Essays in Mathmaticks.*
Les premières preuves de convergence de la méthode sont dues à J.R. Mouraille, 1768, puis J. Fourier et A. Cauchy pour le cas des fonctions d'une variable. L'un des meilleurs ouvrages sur cette méthode est sans doute le livre de J.P. Dedieu [13].

Cette section présente une version intervalle de la méthode de Newton. Ici, l'accent est mis sur les propriétés de l'opérateur de Newton par intervalles. L'algorithme présenté en fin de section n'est qu'une première ébauche. En pratique, il nécessite de nombreuses améliorations pour être efficace. Le lecteur pourra consulter par exemple Hansen [8]. Sa structure générale repose sur une méthode de bissection combinée à l'opérateur suivant :

Définition 2.3.1
Soient $[A] \in \mathbb{IR}^{n\times n}$ et $[b] \in \mathbb{IR}^n$, on note par $\Sigma([A],[b])$ un intervalle qui contient de $\mathbb{IR}^n$ qui contient l'ensemble $\{x \in \mathbb{R}^n \mid Ax = b, A \in [A], b \in [b]\}$. Soit f une fonction continuement différentiable de $\mathbb{R}^n$ vers $\mathbb{R}^n$, $[x]$ un intervalle de $\mathbb{R}^n$, et x_1 un élément de $[x]$. Avec $[Df] : \mathbb{IR}^n \to \mathbb{IR}^{n\times n}$ une fonction d'inclusion pour Df, la fonction

$$\begin{array}{rcccl} N & : & \mathbb{IR}^n & \to & \mathbb{IR}^n \\ & & [x] & \mapsto & x_1 - \Sigma([Df]([x]), f(x_1)) \end{array} \tag{2.53}$$

est appelée *opérateur de Newton par intervalles.*

D'un certain point de vue, l'opérateur de Newton par intervalles étend la méthode de la sécante en prenant toutes les pentes $Df([x])$. Cet opérateur jouit de propriétés fortes intéressantes. En effet, comme le montre la proposition 2.3.2, si x^* est un zéro de f appartenant à $[x]$, alors $x^* \in N([x])$. Combinée avec le théorème du point fixe de Brouwer, la proposition 2.3.4 montre que cet opérateur permet de prouver l'existence et l'unicité d'un zéro de f sur un intervalle donné. Finalement, comme l'illustre le théorème 2.3.6, modulo certaines hypothèses, la suite des itérés $\underbrace{N \circ \ldots N}_{n \text{ fois}}([x])$ est convergente.

Proposition 2.3.2 *Soient $f : \mathbb{R}^n \to \mathbb{R}^n$ une fonction continuement différentiable, $[x]$ un intervalle, x_1 et x^* deux points de $[x]$.*

$$f(x^*) = 0 \Rightarrow x^* \in N([x]) \tag{2.54}$$

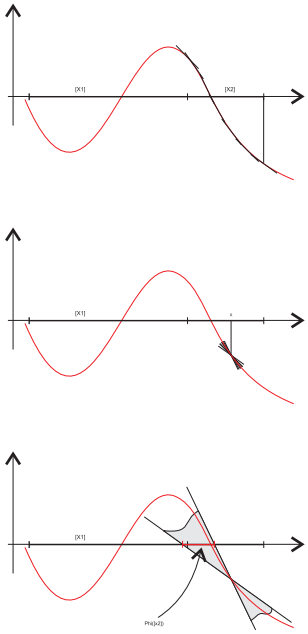


FIG. 2.18 – Illustration graphique de la méthode de Newton par intervalles.

Preuve : On commence par rappeler le théorème de Rolle :
Soit $f : [x] \to \mathbb{R}$ une fonction continuement différentiable sur l'intérieur de $[x]$. Pour tout x, y de $[x]$, il existe ξ sur le segment $[x, y]$ tel que

$$f(x) - f(y) = \nabla f(\xi) \cdot (x - y)$$

Ce théorème n'est plus valable dans le cas des fonctions à valeurs vectorielles. Néanmoins, si on note par f_i les fonctions à valeurs réelles telles que

$$\begin{array}{rcccl} f & : & \mathbb{R}^n & \to & \mathbb{R}^n \\ & & x = (x_1, \ldots, x_n) & \mapsto & (f_1(x), \ldots, f_n(x)) \end{array} \tag{2.55}$$

le théorème de Rolle implique l'existence de n éléments de $[x]$ notés $\xi_1, \ldots, \xi_n$ tels que

$$f(x_1) - f(x^*) = \begin{pmatrix} \nabla f_1(\xi_1) \\ \vdots \\ \nabla f_n(\xi_n) \end{pmatrix} (x_1 - x^*) \tag{2.56}$$

Supposons que x^* annule f. Alors

$$\exists \xi_1, \ldots, \xi_n \in [x], x^* = x_1 - \begin{pmatrix} \nabla f_1(\xi_1) \\ \vdots \\ \nabla f_n(\xi_n) \end{pmatrix}^{-1} f(x_1). \tag{2.57}$$

Comme

$$\begin{pmatrix} \nabla f_1(\xi_1) \\ \vdots \\ \nabla f_n(\xi_n) \end{pmatrix} \in [Df]([x])$$

On en déduit que $x^* \in N([x])$.

□

Remarque 2.3.3
La propriété précédente nous permet de conclure que les zéros de f contenus dans $[x]$ vivent nécessairement dans $N([x])$, par conséquent :

$$x^* \in [x], f(x^*) = 0 \Rightarrow x^* \in [x] \cap N([x]). \tag{2.58}$$

De ce résultat, on peut aussi en déduire par contraposée :

$$[x] \cap N([x]) = \emptyset \Rightarrow \forall x \in [x], f(x) \neq 0 \tag{2.59}$$

Proposition 2.3.4 *On note par $GL(\mathbb{R}^n)$ l'ensemble des matrices $n \times n$ inversibles. Soient f une fonction continuement différentiable sur $[x]$ et $x_1 \in [x]$. Si $[Df]([x]) \subset GL(R^n)$ et si $N([x]) \subset [x]$ alors il existe un unique $x^* \in [x]$ tel que $f(x^*) = 0$.*

Preuve :

– Unicité : Supposons que x^* et x^{**} soient deux zéros de f vivant dans $[x]$, alors on peut écrire :

$$0 = f(x^{**}) - f(x^*) = J(x^*, x^{**})(x^{**} - x^*). \tag{2.60}$$

$$\text{où } J(x^*, x^{**}) = \int_0^1 Df((1-t)x^* + tx^{**})\, \mathrm{d}t. \tag{2.61}$$

or $J(x^*, x^{**}) \in [Df]([x]) \subset GL(R^n)$, donc $J(x^*, x^{**})$ est inversible. On en déduit que : $x^* = x^{**}$.

– Existence : Soit ρ la fonction définie sur $[x]$ par $\rho(x) = x - J(x, x_1)^{-1} f(x)$. Avec la relation $J(x, x_1)(x_1 - x) = f(x_1) - f(x)$, on a :

$$\begin{aligned}
\rho(x) &= x - J(x, x_1)^{-1} f(x_1) + J(x, x_1)^{-1} f(x_1) - J(x, x_1)^{-1} f(x) \\
&= x - J(x, x_1)^{-1} f(x_1) + J(x, x_1)^{-1} (f(x_1) - f(x)) \\
&= x - J(x, x_1)^{-1} f(x_1) + x_1 - x \\
&= x_1 - J(x, x_1)^{-1} f(x_1) \\
\text{donc } \rho(x) &\in N[x] \subset [x].
\end{aligned}$$

On en déduit que ρ est une fonction continue de $[x]$ dans $[x]$. D'après le théorème de Brouwer, il existe $x^* \in [x]$ tel que $\rho(x^*) = x^*$, et donc x^* vérifie la relation $J(x, x_1)^{-1} f(x^*) = 0$, comme $J(x, x_1)$ est un élément de $[Df]([x])$, on en déduit que x^* satisfait $f(x^*) = 0$.

□

Remarque 2.3.5

Pour pouvoir appliquer cette méthode en pratique, il faut déterminer l'ensemble $\Sigma([Df]([x]), f(x_1))$. De plus, $f(x_1)$ n'est souvent connu qu'avec une certaine précision, i.e. $f(x_1) \in [f(x_1)]$. Nous devons donc trouver l'image réciproque de $[f(x_1)]$ via une famille de matrices $[Df]([x])$. Quand cet ensemble est borné, il ressemble à la figure 2.19.

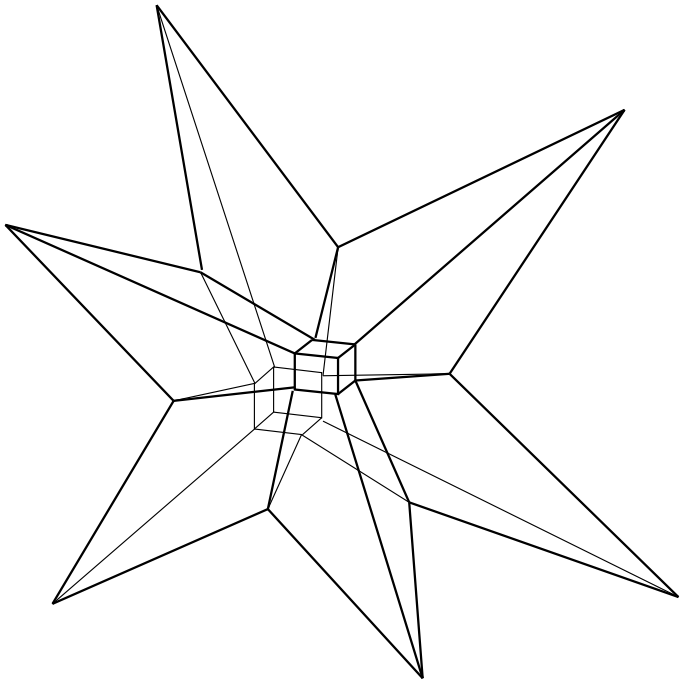


FIG. 2.19 – Représentation de l'image inverse d'un pavé via une famille de matrices qui forme un intervalle de matrices. Cette figure est la première page du livre de Neumaier [30].

En grande dimension, cet ensemble est trop complexe pour être caractérisé en un temps de calcul raisonnable. La plupart du temps, on ne fournit qu'un intervalle de $\mathbb{R}^n$ qui contient cet ensemble. Le livre de Neumaier [30] contient un chapitre entier à la résolution de système d'équations linéaires intervalles.

Avant d'énoncer un théorème de convergence de l'algorithme de Newton par intervalle, on introduit la notation suivante. Si $[x]$ est un élément de $\mathbb{IR}$, alors on note par $|[x]|$ le maximum des $|x|$ tel que $x \in [x]$. On généralise cette notation aux matrices, si $[A]$ est une matrice dont les coefficients sont dans $\mathbb{IR}$, *i.e.* $A \in \mathbb{IR}^{n\times n}$, alors on note par $|[A]|$ la matrice A à coefficients réels tels que $A_{ij} = |[A]_{ij}|$.

Théorème 2.3.6 *Soit $x^* \in [x]$ un zéro de f tel que $Df(x^*) \in GL(\mathbb{R}^n)$ et si le rayon spectral de la matrice $A = |I - \Sigma([Df]([x]), [Df]([x]))|$ est strictement inférieure à 1 alors la suite $\{[x_n]\}_{n\in\mathbb{N}}$ définie par $[x_{n+1}] = N([x_n])$ et $[x_0] = [x]$ converge vers l'intervalle $\{x^*\}$.*

Preuve : La preuve de ce résultat est largement inspirée de la preuve donnée par Ale-

feld [31]. Pour commencer, on s'appuie sur les idées de la démonstration précédente. L'élément $f(x_1)$ vérifie $f(x_1) - f(x^*) = J(x^*, x_1)(x_1 - x^*)$, et de plus $f(x^*) = 0$, par conséquent $f(x_1) = J(x^*, x_1)(x_1 - x^*)$.

On a donc :

$$\begin{array}{rcl} N([x]) - x^* & = & x_1 - x^* - \Sigma([Df]([x]), f(x_1)) \\ N([x]) - x^* & \subset & x_1 - x^* - \Sigma([Df]([x]), I) \cdot f(x_1) \\ N([x]) - x^* & = & x_1 - x^* - \Sigma([Df]([x]), I) J(x^*, x_1) \cdot (x_1 - x^*)) \\ N([x]) - x^* & \subset & x_1 - x^* - \Sigma([Df]([x]), J(x^*, x_1)) \cdot (x_1 - x^*)) \\ N([x]) - x^* & \subset & (I - \Sigma([Df]([x]), [Df][x])) \cdot (x_1 - x^*) \end{array} \tag{2.62}$$

Comme x_1 est un élément de $[x]$, on en déduit que :

$$N([x]) - x^* \subset (I - \Sigma([Df]([x]), [Df]([x])) \cdot ([x] - x^*). \tag{2.63}$$

Ce qui implique

$$q([x_{n+1}], \{x^*\}) \leq \rho(A) q([x_n], \{x^*\}). \tag{2.64}$$

Par induction, on a :

$$q([x_n], \{x^*\}) \leq \rho^n(A) q([x], \{x^*\}). \tag{2.65}$$

On utilise maintenant l'hypothèse $\rho(A) < 1$ pour conclure que la suite $[x]_n$ converge vers $\{x^*\}$.

□

Remarque 2.3.7
Il existe des résultats de vitesse de convergence de cette méthode. Le lecteur intéressé pourra se référer à l'article très clair d'Alefeld [31].

Du théorème 2.3.6, on en déduit le corollaire suivant :

Corollaire 2.3.8 *Si x^* est un zéro de f tel que $[Df](x^*) \in GL(\mathbb{R}^n)$, il existe un intervalle $[x]$ tel que $x^* \in int[x]$, et un entier k tel que $[x]_k \subset [x]$.*

Preuve : La fonction $g : [x] \in \mathbb{IR}^n \mapsto |I - Df^{-1}([x]) Df([x])| \in \mathbb{R}^{+n \times n}$ et la fonction rayon spectral $\rho : \mathbb{R}^{+n \times n} \to \mathbb{R}^+$ sont continues. De plus : $\rho(g(\{x^*\})) = 0$, par conséquent il existe un intervalle $[x]$ contenant x^* tel que $\rho(g([x])) < 1$. D'après le théorème précédent, on sait que la suite des $[x]_n$ converge vers $\{x^*\}$, donc il existe un $k \in \mathbb{N}$ tel que $[x]_k \subset [x]$.

□

Proposition 2.3.9 *Soient f une fonction continuement différentiable sur $[x]$ et $x_1 \in [x]$. Si $[Df]([x]) \subset GL(R^n)$ et s'il existe un entier k tel que $[x]_k \subset [x]$ alors il existe un unique $x^* \in [x]_k$ tel que $f(x^*) = 0$.*

Preuve :

- Unicité : Voir la proposition 2.3.4.
- Existence : l'hypothèse $[x]_k \subset [x]$ implique que la fonction continue N^k définie par $N^k = \underbrace{N \circ \ldots N}_{k \text{ fois}}$ vérifie $N^k([x]) \subset [x]$. Par le théorème de Brouwer, on en déduit l'existence d'un point fixe $x^* \in [x]$ tel que $N^k(x^*) = x^*$.

En définitive, on a l'existence et l'unicité d'un point fixe pour l'application N^k, *i.e.*

$$N^k(x^*) = x^* \tag{2.66}$$

En appliquant la fonction N à chaque membre de cette égalité, on en déduit :

$$\begin{aligned} N(N^k(x^*)) &= N(x^*) \\ N^{k+1}(x^*) &= N(x^*) \\ N^k(N(x^*)) &= N(x^*) \end{aligned}$$

Donc $N(x^*)$ est aussi un point fixe de N^k, par unicité du point fixe de N^k, on en déduit l'existence de $x^* \in [x]$ tel que $N(x^*) = x^*$. Ce qui implique l'existence de $x^* \in [x]$ tel que $f(x^*) = 0$. Par la proposition 2.3.2, on en conclut qu'il existe un unique $x^* \in [x]_k$ vérifiant $f(x^*) = 0$.

□

Grâce à ces propositions, on peut écrire l'algorithme 2 qui combine la bissection et l'opérateur de Newton par intervalles. Cette version naïve de l'algorithme de Newton par intervalles nécessite de nombreuses améliorations avant d'être implémenté. On pourra consulter Hansen [8].

Alg. 2 Algorithme de Newton par intervalles.

Entrée: $[x]$ un intervalle $\mathbb{R}^n$. $f : [x] \to \mathbb{R}^n$.

Sortie: Une famille finie d'intervalles $\mathcal{P} = \{[x_j]\}_j$ telle que $V_f \subset \bigcup [x_j]$ et chaque $[x_j]$ contient un unique zéro de f.

1: Initialisation : $\mathcal{P} \leftarrow \emptyset$, $\mathcal{P}_\Delta \leftarrow \{[x]\}$,
2: **tant que** $\mathcal{P}_\Delta \neq \emptyset$ **faire**
3: $[x_t] \leftarrow [x]$ où $[x] \in \mathcal{P}_\Delta$.
4: $\mathcal{P}_\Delta \leftarrow \mathcal{P}_\Delta - \{[x_t]\}$.
5: **si** $[Df]([x_t]) \subset GL(\mathbb{R}^n)$ et $\rho(|I - \Sigma([Df]([x]), [Df]([x]))|) < 1$ et $0 \in f([x_t])$ **alors**
6: $[x_n] \leftarrow [x_t]$
7: **répéter**
8: $[x_t] \leftarrow N([x_t])$
9: **tant que** $[x_n] \cap N([x_t]) = \emptyset$ **ou** $N([x_t]) \subset [x_n]$.
10: **si** $N([x_t]) \subset [x_n]$ **alors**
11: $\mathcal{P} \leftarrow \mathcal{P} \cup \{[x_t]\}$;
12: **fin si**
13: **sinon**
14: Bissecte $[x_t]$ en deux intervalles $[x_{t_1}]$ et $[x_{t_2}]$.
15: $\mathcal{P}_\Delta \leftarrow \mathcal{P}_\Delta \cup \{[x_{t_1}]\} \cup \{[x_{t_2}]\}$;
16: **fin si**
17: **fin tant que**

Remarque 2.3.10
Il existe différentes manières de vérifier que $\rho(|I - \Sigma([Df]([x]), [Df]([x]))|) < 1$. Comme la matrice $|I - \Sigma([Df]([x]), [Df]([x]))|$ est une matrice à coefficients positifs, on peut utiliser les résultats qui découlent de la théorie de Perron-Frobenius pour évaluer son rayon spectral.

L'algorithme 2 ne termine pas pour toute fonction f différentiable. Par exemple, si on cherche à montrer que la fonction $x \mapsto x^2$ s'annule uniquement en 0 sur l'intervalle $[-1, 1]$, l'algorithme 2 ne termine pas car la condition $[Df]([x_t]) \subset GL(\mathbb{R}^n)$ est fausse pour tout intervalle $[x_t]$ qui contient 0. La proposition suivante montre que cette situation est rare.

Proposition 2.3.11 *Soit f une fonction $\mathcal{C}^\infty(\mathbb{R}^n, \mathbb{R})$ générique. Alors l'algorithme 2 termine. (il fournit un ensemble fini d'intervalles $\{[x]_i\}$ tel que chaque intervalle*

contient un unique zéro de f).

Preuve : La généricité de f entraîne $\{f(x) = 0\} \cap \{Df(x) \notin GL(\mathbb{R}^n)\} = \emptyset$. On peut déduire des travaux d'Alexei Grigoriev (non encore publiés) que l'ensemble des zéros d'une fonction $\mathcal{C}^\infty$ générique définie sur un compact de $\mathbb{R}^n$ est fini.

□

2.4 Positivité

2.4.1 Introduction

Dans le cadre des chapitres suivants, nous aurons besoin de vérifier la positivité d'une fonction donnée. En particulier, on utilisera une des méthodes présentées pour montrer qu'une fonction est de Lyapunov au chapitre 4. Dans les deux sections 2.4.2 et 2.4.3, on se focalise sur le problème de montrer qu'une fonction est positive. On distingue deux situations :

1. le cas où pour tout x, $f(x) > 0$.
2. le cas où pour tout x, $f(x) \geq 0$.

Cette disjonction (qui n'en est pas une) peut paraître étrange. De façon évidente, la première propriété implique la seconde. Modulo quelques hypothèses de continuité sur les fonctions d'inclusion, le calcul par intervalles est capable de montrer que $\forall x, f(x) > 0$. Par conséquent le calcul par intervalle est souvent capable de montrer que $\forall x, f(x) \geq 0$. Il reste donc à étudier le cas où

$$\forall x, f(x) \geq 0 \text{ et } \exists x, f(x) = 0. \tag{2.67}$$

Ce qui est l'objet de la section 2.4.3.

2.4.2 Positivité stricte

Dans cette section, on montre que l'algorithme de bissection est suffisamment puissant pour montrer l'assertion $\forall x, f(x) > 0$. La proposition suivante donne une preuve que l'algorithme de bissection termine si et seulement si $\forall x, f(x) > 0$. On présente le résultat dans le cas monodimentionnel. La preuve du cas multivariables est basée sur le même raisonnement. La preuve de ce résultat, bien que facile d'accès, est introuvable dans la littérature.

Proposition 2.4.1 *Soient* $[w]$ *un intervalle de* $\mathbb{R}$, $f : [w] \to \mathbb{R}$ *continue,* $[f] : \mathbb{IR} \to \mathbb{IR}$ *une fonction d'inclusion continue pour* f *vérifiant* $[f](\{x\}) = \{f(x)\}, \forall x \in [w]$. *Si* $\epsilon = 0$, *alors* $V_f = \emptyset$ *si et seulement l'algorithme de bissection termine.*

Preuve : Supposons $V_f = \emptyset$. L'ensemble $[x]$ est compact et f est continue. On peut donc en déduire que $\forall x \in [w], f(x) > 0$ ou $\forall x \in [w], f(x) < 0$. Supposons, par exemple, que nous sommes dans la première situation pour la fin de la démonstration, l'autre cas étant analogue. La continuité de f, la compacité de $[w]$ et $f(x) > 0$ impliquent l'existence d'un réel δ_0 strictement positif tel que $f(x) \geq \delta_0$. L'ensemble $\mathbb{IR} \cap [w]$ muni de la topologie induite par q est compact. D'après le théorème de Heine, $[f]$ est uniformément continue. Et donc

$$\forall \delta > 0, \exists\, \epsilon > 0, \;\; \forall [x], [y] \in \mathbb{IR} \cap E, q([x], [y]) < \epsilon \Rightarrow q([f]([x]), [f]([y])) < \delta. \quad (2.68)$$

En posant $[y] = \{x'\}$ singleton inclus dans $[x]$, on a :

$$\forall \delta > 0, \exists\, \epsilon > 0, \;\; \forall [x] \subset E, \forall \{x'\} \subset [x], q([x], \{x'\}) < \epsilon \Rightarrow q([f]([x]), [f](\{x'\})) < \delta. \quad (2.69)$$

Or $[f](\{x'\}) = f(\{x'\})$, donc :

$$\forall \delta > 0, \exists\, \epsilon > 0, \;\; \forall [x] \subset E, \forall \{x'\} \subset [x], q([x], \{x'\}) < \epsilon \Rightarrow q([f]([x]), f(\{x'\})) < \delta. \quad (2.70)$$

$\forall \delta > 0, \exists\, \epsilon > 0, \;\; \forall [x] \subset E,$

$$\forall \{x'\} \subset [x], q([x], \{x'\}) < \epsilon \Rightarrow \exists \{x'\} \subset [x], q([f]([x]), f(\{x'\})) < \delta. \quad (2.71)$$

Or $\exists \epsilon_1 > 0 \forall \{x'\} \subset [x], q([x], \{x'\}) < \epsilon_1 \Leftrightarrow \exists \epsilon_2 m([x]) < \epsilon_2$. Et la proposition : $\exists \{x'\} \subset [x], q([f]([x]), f(\{x'\})) < \delta$ est équivalente à $m([f]([x])) < 2\delta$.

Avec $2\delta = \delta_0$, on en déduit l'existence d'un $\epsilon_0 > 0$ tel que

$$\forall [x] \subset E, \;\; m([x]) < \epsilon_0 \Rightarrow \;\; m([f]([x])) < \delta_0 \quad (2.72)$$

La fonction $[f]$ une fonction d'inclusion pour f et $f(x) \leq \delta_0$, et donc nécessairement :

$$\forall [x] \subset E, \;\; \sup[f]([x]) \geq \delta_0 \quad (2.73)$$

Or

$$\inf[f]([x]) = \underbrace{\sup[f]([x])}_{\geq \delta_0} \;\; \underbrace{-m([f]([x]))}_{> -\delta_0 \text{ si } m([x]) < \epsilon_0} \quad (2.74)$$

On en déduit qu'il existe $\epsilon_0 > 0$ tel que :

$$\forall [x] \subset E, m([x]) < \epsilon_0 \Rightarrow \inf[f]([x]) > 0 \quad (2.75)$$

Et donc l'algorithme de bissection termine en fournissant une famille $\mathcal{P}$ vide.

□

Il est difficile de donner une complexité à ce genre d'algorithmes. Le temps de calcul du dernier algorithme dépend cruellement de la valeur δ_0 qui minore f. De plus, donner une estimation du temps de calcul en fonction de cette valeur ne semble pas bien utile. En effet, supposons que l'on connaisse un $\delta_0 \in \mathbb{R}^{+*}$ tel que $\forall x \in [x], f(x) > \delta_0$, on sait alors que $f(x) > 0$, et faire tourner cet algorithme ne sert alors à rien.

De la proposition 2.4.1, on peut déduire le corollaire suivant.

Corollaire 2.4.2 *Soient f, et $\{h_i\}_{i\in\{1,\dots,n\}}$ une famille discrète de fonctions réelles définies sur un intervalle $[x]$ de $\mathbb{R}^n$ dont on dispose de fonction d'inclusion continue vérifiant $[h_i](\{x\}) = \{h_i(x)\}$. Il est possible de résoudre en un temps fini le problème suivant :*
montrer que $\forall x \in [x], f(x) > 0$ sous les contraintes $h_i(x) = 0$ ($i \in \{1, \dots, n\}$).

Preuve : Le problème revient à montrer que

$$\forall\ x \in [x],\ \begin{cases} h_1(x) = 0 \\ \vdots \\ h_n(x) = 0 \end{cases} \Rightarrow f(x) > 0. \tag{2.76}$$

Ce qui est équivalent au problème :

$$\forall\ x \in [x],\ h_1(x) \neq 0 \vee \cdots \vee h_n \neq 0 \vee f(x) > 0. \tag{2.77}$$

ou encore

$$\forall\ x \in [x],\ h_1^2(x) > 0 \vee \cdots \vee h_n^2 > 0 \vee f(x) > 0. \tag{2.78}$$

$$\forall\ x \in [x],\ \max(h_1^2(x), \dots, h_n^2, f(x)) > 0. \tag{2.79}$$

où max est la fonction de $\mathbb{R}^n \to \mathbb{R}$ qui à $(x_1, \dots, x_n)$ associe le réel x_j vérifiant $\forall i \in \{1, \dots, n\},\ x_j \geq x_i$. En posant $g(x) = \max(h_1^2(x), \dots, h_n^2(x), f(x))$ et $[\max] : \mathbb{IR}^n \to \mathbb{IR}, [x] \mapsto [\max(\underline{x}), \max(\overline{x})]$, on retrouve les hypothèses de la proposition 2.4.1.

□

En conclusion, la stricte positivité d'une fonction f peut être obtenue grâce au calcul par intervalles.

2.4.3 Positivité ≥ 0

Cette section présente un théorème qui nous donne une condition suffisante pour vérifier $\forall x \in [x], f(x) \geq 0$. La méthode qui en découle est en partie basée sur le calcul par intervalles et sur le calcul formel.

Définition 2.4.3
Une matrice symétrique $A \in \mathbb{R}^n$ est *définie positive* si $\forall x \in \mathbb{R}^n - \{0\}, x^T Ax > 0$. L'ensemble des matrices $n \times n$ symétriques définies positives est noté S^{n+}.

Théorème 2.4.4 *Soit $f \in \mathcal{C}^\infty([x] \subset \mathbb{R}^n, \mathbb{R})$.*
Si

- $\exists x_0 \in [x]$ *tel que* $f(x_0) = 0$ *et* $\nabla f(x_0) = 0$.
- $\nabla^2 f([x]) \subset S^{n+}$

alors $\forall x \in [x] - x_0,\ f(x) > 0$.

Preuve : La condition $\forall x \in [x], \nabla^2 f(x) \in S^{n+}$ implique que f est une fonction strictement convexe définie sur un ensemble convexe $[x]$. Comme $\nabla f(x_0) = 0$, on en déduit que

$$\forall x \in [x] - \{x_0\}, f(x) > f(x_0) = 0. \tag{2.80}$$

En d'autres termes : $\forall x \in [x] - \{x_0\}, f(x) > 0$.

□

Ce théorème induit une méthode effective pour prouver que f est positive. En effet, si pour un certain $x_0 \in [x]$ les conditions $f(x_0) = 0$ et $\nabla f(x_0) = 0$ peuvent être validées (par exemple en utilisant le calcul algébrique [32]), il suffit de vérifier alors que $\nabla^2 f([x])$ est inclus dans S^{n+} pour conclure.

En pratique, ce dernier test est vérifié en utilisant le calcul par intervalles et des résultats sur les matrices symétriques intervalles : avec $\underline{A}$ et $\overline{A}$ deux matrices symétriques telles que $\underline{A} \leq \overline{A}$, une matrice symétrique intervalle [43] est un ensemble $[A]$ de matrices symétriques de la forme :

$$[A] = \{A \in \mathbb{R}^{n \times n}, \underline{A} \leq A \leq \overline{A}, A^T = A\} \tag{2.81}$$

Ici, la relation d'ordre $\leq$ est à comprendre termes à termes.

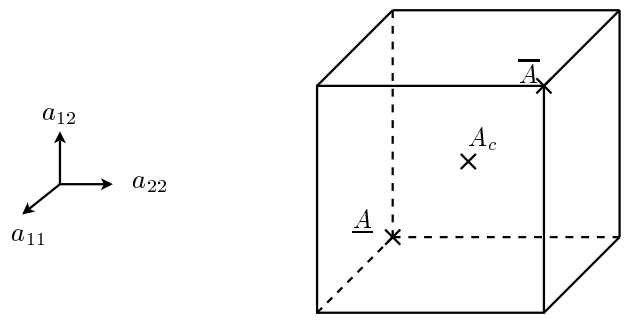


FIG. 2.20 – Avec $n = 2$, une matrice intervalle symétrique $[\underline{A}, \overline{A}\ ]$.

Définition 2.4.5

Une matrice intervalle symétrique $[A]$ est définie positive si $[A] \subset S^{n+}$.

Remarque 2.4.6

Soit $V([A])$ l'ensemble fini composé des sommets de $[A]$. Comme S^{n+} et $[A]$ sont des sous-ensembles convexes de l'ensemble des matrices symétriques , on a l'équivalence suivante [33] :

$$[A] \subset S^{n+} \Leftrightarrow V([A]) \subset S^{n+} \tag{2.82}$$

L'ensemble des matrices symétriques $n \times n$ est un espace vectoriel de dimension $\begin{pmatrix} n+1 \\ 2 \end{pmatrix} = \frac{n(n+1)}{2}$. Par conséquent, on a : $\#V([A]) = 2^{\frac{n(n+1)}{2}}$. En particulier, en testant la positivité de ces $2^{\frac{n(n+1)}{2}}$ matrices, on teste l'inclusion de $[A]$ dans S^{n+}. La figure 2.21 illustre dans le cas de la dimension 2 cette remarque.

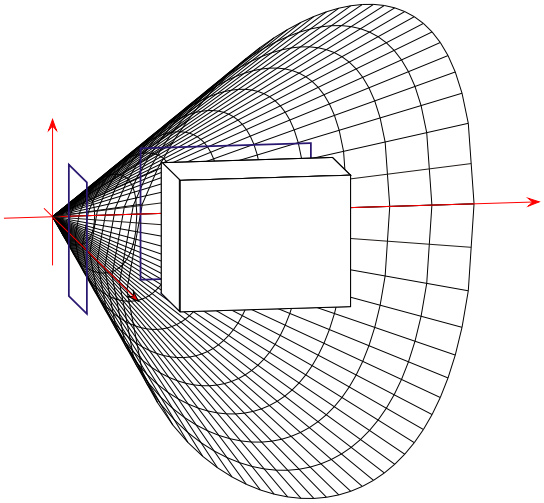


FIG. 2.21 – Une matrice intervalle symétrique $[A]$ est définie positive si elle est incluse dans S^{n+}.

Il est possible d'améliorer le nombre de matrices à tester. Si $[A]$ est une matrice intervalle symétrique, on peut créer deux matrices symétriques A_c et Δ telles que $[A] = \{A, A_c - \Delta \leq A \leq A_c + \Delta\}$ où

$$A_c = \frac{1}{2}(\underline{A} + \overline{A}) \tag{2.83}$$

$$\Delta = \frac{1}{2}(\overline{A} - \underline{A}) \tag{2.84}$$

On note par C l'ensemble fini :

$$C = \{x \in \mathbb{R}^n, |x_i| = 1, \forall i \in \{1, \ldots, n\}\}. \tag{2.85}$$

On a $\#C = 2^n$. Pour chaque z de C, on note par T_z la matrice diagonale définie par z, i.e. $T_z = Diag(z)$. On note par A_z la matrice $A_c - T_z \Delta T_z$. Chaque A_z, avec $z \in C$ est évidemment un élément de $[A]$, et comme $A_{-z} = A_z$, l'ensemble $\{A_z, z \in C\}$ est fini et de cardinal 2^{n-1}. Dans [33], Rohn montre que $[A] \subset S^{n+}$ est équivalent à la positivité des 2^{n-1} matrices de $\{A_z, z \in C\}$.

Exemple 2.4.7

Soit $f : \mathbb{R}^2 \to \mathbb{R}$ la fonction définie par : $f(x, y) = -\cos(x^2 + \sqrt{2}\sin^2 y) + x^2 + y^2 + 1$. Cette fonction vérifie $f(0, 0) = 0$ et $\nabla f(0, 0) = 0$ car

$$\nabla f(x, y) = \begin{pmatrix} 2x(\sin(x^2 + \sqrt{2}\sin^2 y) + 1) \\ 2\sqrt{2}\cos y \sin y \sin\left(\sqrt{2}\sin^2 y + x^2\right) + 2y \end{pmatrix}. \tag{2.86}$$

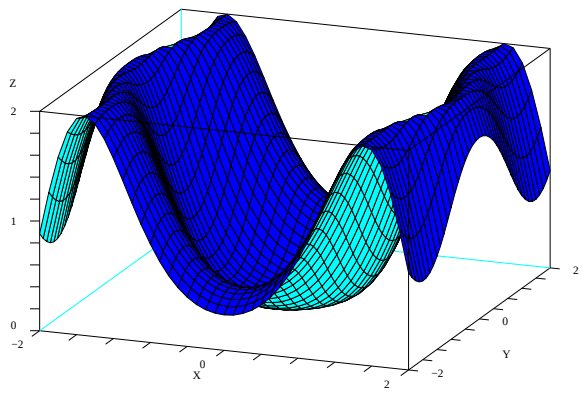


FIG. 2.22 – Graphe de f.

On a :

$$\nabla^2 f = \begin{pmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{pmatrix} \tag{2.87}$$

où les $a_{i,j}$ sont donnés par :

$$\begin{aligned} a_{1,1} \quad = \quad & 2\sin\left(\sqrt{2}\sin^2 y + x^2\right) \\ & +4x^2\cos\left(\sqrt{2}\sin^2 y + x^2\right) + 2. \end{aligned} \tag{2.88}$$

$$\begin{aligned} a_{2,2} \quad = \quad & -2\sqrt{2}\sin^2 y \sin(\sqrt{2}\sin^2 y + x^2) \\ + \quad & 2\sqrt{2}\cos^2 y \sin(\sqrt{2}\sin^2 y + x^2) \\ + \quad & 8\cos^2 y \sin^2 y \cos(\sqrt{2}\sin^2 y + x^2) + 2. \end{aligned} \tag{2.89}$$

$$a_{1,2} = a_{2,1} = 4\sqrt{2}\, x\, \cos y\, \sin y\, \cos\left(\sqrt{2}\, \sin^2 y + x^2\right). \tag{2.90}$$

Il est facile de construire des fonctions d'inclusion pour les fonctions $a_{i,j}$. On en déduit donc que pour tout x dans $[-1/2, 1/2]^2$, $\nabla^2 f(x)$ est un élément de $[A]$ avec :

$$[A] = \begin{pmatrix} [1.9, 4.1] & [-1.3, 1.4] \\ [-1.3, 1.4] & [1.9, 5.4] \end{pmatrix} \tag{2.91}$$

Finalement, la remarque 2.4.6 nous permet de dire que $f(x) \geq 0$ pour tout x dans $[-\frac{1}{2}, \frac{1}{2}]^2$ si les 2^{2-1} matrices :

$$A_1 = \begin{pmatrix} 1.9 & -1.3 \\ -1.3 & 1.9 \end{pmatrix} \text{ et } A_2 = \begin{pmatrix} 1.9 & 1.4 \\ 1.4 & 1.9 \end{pmatrix} \tag{2.92}$$

sont définies positives.

2.5 Conclusion

L'utilisation ingénieuse du calcul par intervalles a permis construire des algorithmes dont les résultats sont certifiés. En particulier, il est possible de faire de l'optimisation globale comme le montre l'annexe B. Cette optimisation globale vient en opposition aux méthodes classiques d'optimisation locale dont le résultat dépend cruellement de l'initialisation. Excepté dans le cas convexe, ce genre de méthodes ne converge en général que vers un minimum local. Dans la suite de cette thèse, on présente des méthodes garanties pour l'étude de propriétés topologiques d'ensembles décrits par des inégalités.

Chapitre 3

Invariants topologiques

On entend par invariant topologique toute propriété d'un ensemble qui reste invariant via bijection bi-continue (c'est à dire une bijection continue dont la réciproque est aussi continue). Pour un ensemble donné, il existe plusieurs manières de calculer plus ou moins efficacement des invariants topologiques. Un invariant qui est classiquement recherché est le nombre de composantes connexes. On verra dans la section 3.5 pourquoi cet invariant est si important pour les roboticiens.
On commencera ce chapitre par quelques rappels classiques de topologie. On présentera ensuite succinctement les différentes méthodes existantes. Puis, un algorithme capable de renseigner certaines propriétés topologiques d'ensembles décrits par des inégalités $\mathcal{C}^\infty$ est présenté dans chacune des deux sections 3.3 et 3.4. On terminera par quelques applications de ces algorithmes à la planification de trajectoires.

3.1 Rappels topologiques

Dans les algorithmes développés dans ce chapitre, on décompose les sous-ensembles de $\mathbb{R}^n$ en parties dont la topologie est simple. On commence donc cette section avec des rappels de topologie et quelques définitions concernant les ensembles étoilés, contractiles ...

Définition 3.1.1
Deux ensembles X et Y topologiques[1] sont homéomorphes s'il existe une fonction

[1] Un espace topologique X est un ensemble muni d'une famille $\mathcal{T}$ de parties de celui-ci qui vérifie
- $\emptyset \in \mathcal{T}$, et $X \in \mathcal{T}$,
- $O_1, O_2 \in \mathcal{T} \Rightarrow O_1 \cap O_2 \in \mathcal{T}$,

$f : X \to Y$ continue, bijective et dont la réciproque est aussi continue.

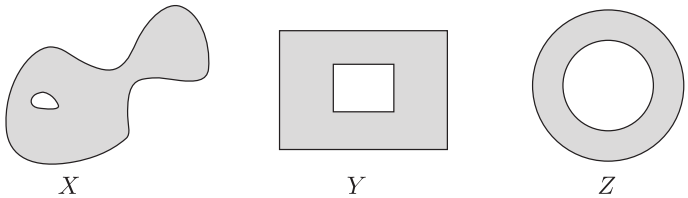


FIG. 3.1 – Les ensembles X, Y et Z sont homéomorphes.

Dans toutes la suite de cette thèse, tous les ensembles considérés sont muni d'une topologie.

Définition 3.1.2

Un ensemble X est *connexe par arcs* [17] si pour chaque paire $x, y \in X$, il existe une fonction continue f de $[0, 1]$ vers X telle que $f(0) = x$ et $f(1) = y$. L'ensemble représenté sur la gauche de la figure 3.2 est connexe par arcs alors que celui de droite ne l'est pas (il a 4 composantes connexes).

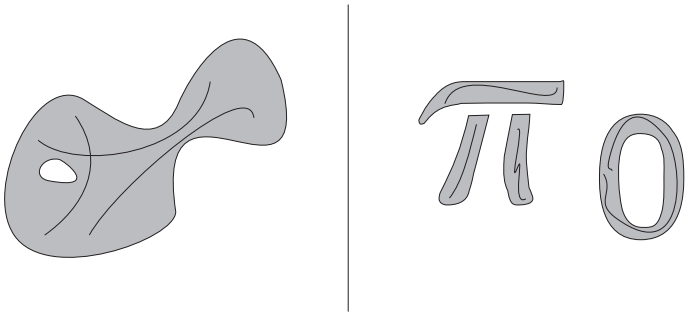


FIG. 3.2 – Exemple d'un ensemble connexe et d'un ensemble à 4 composantes connexes. Les ensembles connexes par arcs sont aussi appelés les ensembles *0-connected*. $\pi_0(S)$ est la notation classiquement employée par les topologues pour désigner le nombre de composantes connexes d'un ensemble S.

Définition 3.1.3

Un point v est une *étoile* pour un sous-ensemble X de $\mathbb{R}^n$ si X contient tous les segments reliant chacun de ses points à v.

– $\{O_i\}_{i \in I} \subset \mathcal{T} \Rightarrow \bigcup_{i \in I} O_i \in \mathcal{T}$.

Les éléments de la topologie $\mathcal{T}$ sont les ouverts. C'est à partir de cette définition que l'on définit la notion de continuité.

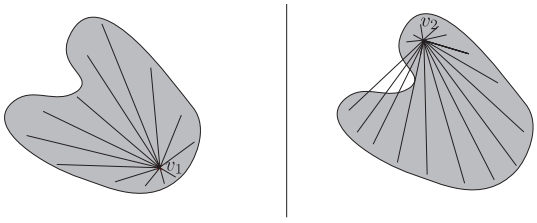


FIG. 3.3 – v_1 est une étoile pour ce sous-ensemble de $\mathbb{R}^2$ alors que v_2 ne l'est pas.

Définition 3.1.4
Si v est une étoile pour un sous-ensemble X d'un espace vectoriel, on dit que X est *étoilé*. Lorsque l'on veut préciser que v est une étoile, on dit que X est *v-étoilé*.

Proposition 3.1.5 *Si X et Y sont deux ensembles v-étoilés, alors $X \cap Y$ et $X \cup Y$ sont aussi des ensembles v-étoilés.*

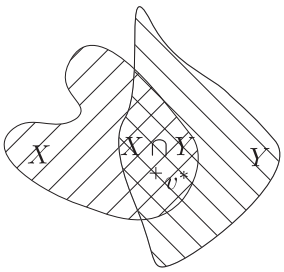


FIG. 3.4 – La famille des ensembles v-étoilés est stable par intersection et par union.

Définition 3.1.6 (Fonctions homotopes)
Deux fonctions continues $f, g : X \to Y$ sont *homotopes* (ou f est *homotope* à g) s'il existe une fonction continue $F : X \times [0,1] \to Y$, telle que : $F(x,0) = f(x)$ et $F(x,1) = g(x)$, $\forall x \in X$. On dit que la fonction F est une *homotopie*. "f est homotote à g" se note $f \simeq g$. La figure 3.5 illustre cette notion.

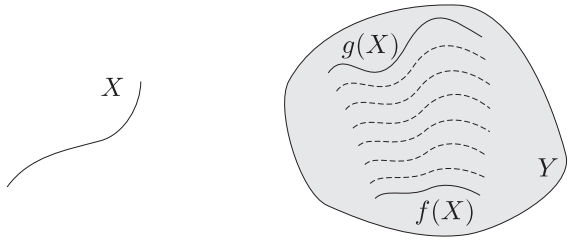


FIG. 3.5 – Les deux fonctions $f, g : X \to Y$ sont homotopes. ($f \simeq g$)

Remarque 3.1.7

Il est facile de vérifier que la relation $\simeq$ définie précédemment est une relation d'équivalence.

Définition 3.1.8 (Equivalence d'homotopie entre des ensembles)

On dit que deux ensembles X et Y sont *du même type d'homotopie* et on note $X \simeq Y$, s'il existe deux fonctions continues $f : X \to Y$ et $g : Y \to X$, telles que : $f \circ g \simeq 1_X$ et $g \circ f \simeq 1_Y$, où 1_X et 1_Y sont les fonctions identités de X et Y respectivement. Dans ce cas, f est dit une *équivalence d'homotopie* et g est *l'homotopie inverse* de f. La figure 3.6 illustre cette notion.

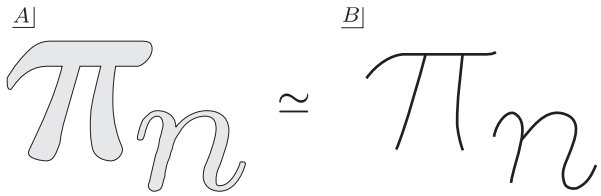


FIG. 3.6 – Les deux ensembles A et B sont du même type d'homotopie. ($A \simeq B$)

Définition 3.1.9

On dit qu'un ensemble X est *contractile* s'il est du même type d'homotopie qu'un point.

Remarque 3.1.10

Un ensemble étoilé est contractile (voir figure 3.7).

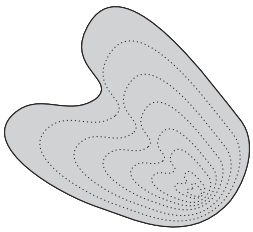


FIG. 3.7 – Un ensemble étoilé est contractile.

Dans la section 3.1.1, on présente une classe d'ensembles de $\mathbb{R}^n$ qui sont décrits par des points, segments, triangles et tétraèdres [48] ...Ces ensembles admettent donc une représentation finie, il suffit d'énumérer l'emplacement des sommets et l'existence ou non des segments, triangles ...qui les relient. Pour cette classe d'ensembles, le problème de connaître leur topologie est un problème de nature combinatoire. La topologie algébrique est quelquefois appelée topologie combinatoire.

3.1.1 Triangulation

Dans cette sous section, on présente une façon de construire des espaces à partir d'ensembles élémentaires appelés *simplexes*.

Un simplexe est une généralisation en dimension n d'un triangle ou d'un tétraèdre. Ces simplexes sont placés de telle manière que deux simplexes s'intersectent en une face commune.

Définition 3.1.11
Considérons $(n+1)$ points $a_0, \ldots, a_n$ de $\mathbb{R}^m$. Ils sont dits *indépendants* si les vecteurs $a_1 - a_0, a_2 - a_0, \ldots, a_n - a_0$ sont linéairement indépendants. Il est facile de vérifier que les points $a_0, \ldots, a_n$ de $\mathbb{R}^m$ sont indépendants si l'implication suivante est satisfaite :

$$\sum_{i=0}^{n} \lambda_i a_i = 0 \text{ et } \sum_{i=0}^{n} \lambda_i = 0 \Rightarrow \lambda_0 = \cdots = \lambda_n = 0 \tag{3.1}$$

Par exemple, trois points non alignés de $\mathbb{R}^2$ sont indépendants.

Définition 3.1.12
Soient $a_0, \ldots, a_n$ $n+1$ points indépendants de $\mathbb{R}^m$. Le n-simplexe géométrique σ_n

engendré par $a_0, \ldots, a_n$ est l'ensemble des points de la forme

$$\sum_{i=0}^{n} \lambda_i a_i = 0 \text{ où } \forall i, \lambda_i \geq 0 \text{ et } \sum_{i=0}^{n} \lambda_i = 1 \tag{3.2}$$

Le simplexe engendré par $a_0, \ldots, a_n$ points indépendants est le plus petit convexe contenant $a_0, \ldots, a_n$. On remarque aussi qu'un point A de σ_n est le barycentre de $\{(\lambda_0, a_0), \ldots, (\lambda_n, a_n)\}$.

Définition 3.1.13

Soit σ un simplexe de $\mathbb{R}^n$ engendré par $\{a_0, \ldots, a_m\}$. Tout simplexe engendré par une partie de $\{a_0, \ldots, a_m\}$ est dit *face* de σ.

Définition 3.1.14

Un *complexe simplicial* K est une famille finie de simplexes, tous contenus dans $\mathbb{R}^n$. De plus, les éléments de K vérifient les implications suivantes :

1. si σ_n est un simplexe K et τ_p une face de σ_n, alors τ_p est aussi dans K.
2. si σ_n et σ_p sont des simplexes de K, alors $\sigma_n \cap \sigma_p$ est soit vide, soit une face commune à σ_n et σ_p.

Un complexe simplicial est simplement une collection de simplexes. L'ensemble des points qui appartiennent à au moins l'un des simplexes de K, muni de la topologie induite de $\mathbb{R}^n$, est un espace topologique, appelé un *polyèdre* de K, noté $|K|$. Les complexes simpliciaux sont des collections de simplexes qui vivent dans un certain espace euclidien $\mathbb{R}^n$. Pour s'affranchir de cette restriction, on introduit la notion de *complexe simplicial abstrait*.

Définition 3.1.15

Soit $\mathcal{K}$ une collection finie de sous ensembles de $\{a^0, a^1, \ldots, a^n\}$. On dit que cette collection $\mathcal{K}$ est un *complexe simplicial abstrait* si

$$\sigma \in \mathcal{K}, \sigma' \subset \sigma \Rightarrow \sigma' \in \mathcal{K} \tag{3.3}$$

Les singletons $\{a^0\}, \{a^1\}, \ldots$ sont appelés *noeuds abstraits* et les sous ensembles de $\{a^{i_0}, a^{i_1}, \ldots, a^{i_n}\}, \ldots$ les *simplexes abstraits*. La *dimension* d'un complexe simplicial abstrait est le maximum des dimensions de ses simplexes.

Exemple 3.1.16

L'ensemble

$$\begin{aligned} \mathcal{K} \quad = \quad & \{\emptyset, \{a^0\}, \{a^1\}, \{a^2\}, \{a^3\}, \{a^4\}, \{a^0, a^1\}, \ldots \\ & \ldots \{a^1, a^2\}, \{a^0, a^2\}, \{a^3, a^4\}, \{a^0, a^1, a^2\}\}. \end{aligned} \tag{3.4}$$

est un complexe simplicial abstrait.
Il est de dimension 2.

L'énumération complète de tous les éléments d'un complexe simplicial abstrait $\mathcal{K}$ n'est pas nécessaire pour le caractériser. En effet, le fait que $\{a^0, a^1, a^2\} \in \mathcal{K}$ implique que $\emptyset, \{a^0\}$, $\{a^1\}$, $\{a^2\}$, $\{a^0, a^1\}$, $\{a^1, a^2\}$, $\{a^0, a^2\}$ sont aussi des éléments de $\mathcal{K}$.

Notation 3.1.17
Avec $\mathcal{V}$ une collection finie d'éléments (noeuds abstraits) $\mathcal{V} = \{a^0, a^1, \ldots, a^n\}$ et $2^{\mathcal{V}}$ l'ensemble des parties de $\mathcal{V}$, un complexe simplicial $\mathcal{K}$ est un sous-ensemble de $\mathcal{P}(\mathcal{V})$ vérifiant (3.3).
Soit $\{\sigma_1, \ldots, \sigma_m\}$ un ensemble inclus dans $2^{\mathcal{V}}$, (pas nécessairement un complexe simplicial abstrait), on note par $\sigma_1 + \cdots + \sigma_m$ le complexe simplicial abstrait suivant[2] :

$$\mathcal{K} = \bigcup_{i=1}^{i=m} \mathcal{P}(\sigma_i) \tag{3.5}$$

On dit que $\sigma_1 + \cdots + \sigma_m$ est une forme réduite de $\mathcal{K}$.

En utilisant cette notation, le complexe simplicial abstrait $\mathcal{K}$ défini à l'exemple 3.1.16 peut s'écrire : $\mathcal{K} = \{a^0, a^1, a^2\} + \{a^3, a^4\}$ ou plus concisement $\mathcal{K} = a^0a^1a^2 + a^3a^4$.

Définition 3.1.18
Soit K un complexe simplicial de $\mathbb{R}^n$, à chaque point α_i de $\mathbb{R}^n$ qui contribue à engendrer un simplexe de K, il est associé un unique symbole a_i. Une abstraction $\mathcal{K}$ de K est un complexe simplicial abstrait qui vérifie

$$\sigma \in K \text{ engendré par } \{\alpha_{i_0}, \ldots, \alpha_{i_m}\} \subset \mathbb{R}^n \Leftrightarrow \{a_{i_0}, \ldots, a_{i_m}\} \in \mathcal{K} \tag{3.6}$$

Définition 3.1.19
Une réalisation d'un complexe simplicial abstrait $\mathcal{K}$ est un complexe simplicial K qui a pour abstraction $\mathcal{K}$.

[2]On peut vérifier que $\sigma_1 + \cdots + \sigma_m$ est le plus petit, au sens de l'inclusion définie sur $2^{2^{\mathcal{V}}}$, complexe simplicial abstrait qui contient σ_1, ..., σ_m, comme simplexes.

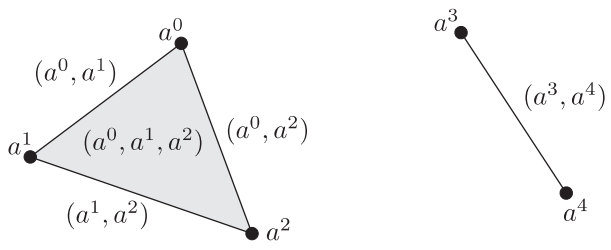


FIG. 3.8 – Une réalisation de $\mathcal{K} = a^0a^1a^2 + a^3a^4$.

Remarque 3.1.20

Si K_1 et K_2 sont deux réalisations d'un même complexe simplicial abstrait $\mathcal{K}$, alors $|K_1|$ et $|K_2|$ sont homéomorphes [48] ; par conséquent, les propriétés topologiques d'un complexe simplicial abstrait $\mathcal{K}$ sont bien définies.

Définition 3.1.21

Soit $\mathcal{K}$ un complexe simplicial abstrait, et $\{x\}$ un noeud abstrait. On note par $\mathcal{C}(x, \mathcal{K})$ l'ensemble suivant :

$$\mathcal{C}(x, \mathcal{K}) = \mathcal{K} \cup \bigcup_{s \in \mathcal{K}} \{\{x\} \cup s\}. \tag{3.7}$$

On peut vérifier que $\mathcal{C}(x, \mathcal{K})$ est complexe simplicial abstrait. Il est classiquement appelé le *cône* de sommet $\{x\}$ et de base $\mathcal{K}$. Avec la notation 3.1.17, un cône est le produit $*$ de x avec $K = \sigma_1 + \cdots + \sigma_m$, on a :

$$\mathcal{C}(x, \mathcal{K}) = x * (\sigma_1 + \cdots + \sigma_m) = x\sigma_1 + \cdots + x\sigma_m. \tag{3.8}$$

Remarque 3.1.22

On utilise la notation $x * \mathcal{K}$ pour le cône de sommet x et de base $\mathcal{K}$, comme un cas particulier de la notion de *join* [48] de deux triangulations abstraites.

Exemple 3.1.23

Pour le $\mathcal{K}$ défini dans l'exemple 3.1.16, $\mathcal{C}(x, \mathcal{K})$ est

$$\begin{aligned} \mathcal{C}(x, \mathcal{K}) \quad = \quad & \mathcal{K} \cup \big\{\{x\}, \{x, a^0\}, \{x, a^1\}, \{x, a^2\}, \{x, a^3\}, \{x, a^4\}, \ldots \\ & \ldots \{x, a^0, a^1\}, \{x, a^1, a^2\}, \{x, a^2, a^3\}, \{x, a^4, a^5\}, \{x, a^0, a^1, a^2\}\big\}. \end{aligned} \tag{3.9}$$

En utilisant la forme réduite, on a :

$$\begin{aligned} \mathcal{C}(x, \mathcal{K}) \quad &= \quad x * \{a^0a^1a^2 + a^3a^4\}, \\ &= \quad xa^0a^1a^2 + xa^3a^4. \end{aligned}$$

La figure 3.9 montre une réalisation de $\mathcal{C}(x, \mathcal{K})$.

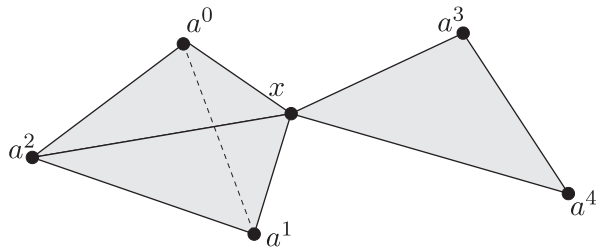


FIG. 3.9 – Une réalisation de $\mathcal{C}(x, \mathcal{K})$.

Proposition 3.1.24 *Soit $\mathcal{K}$ un complexe simplicial abstrait, et $\mathcal{C}(x, \mathcal{K})$ un cône de base $\mathcal{K}$, alors $\mathcal{C}(x, \mathcal{K})$ est homotope au singleton $\{x\}$.* i.e. $\mathcal{C}(x, \mathcal{K}) \simeq \{x\}$.

Preuve : Soit K une réalisation de $\mathcal{K}$, par construction $|K|$ est $|\{x\}|$-étoilé.

□

3.1.2 Quelques invariants ...

L'un des invariants topologiques les plus simples est le nombre de composantes connexes par arcs. La proposition suivante permet de montrer aisément que c'est un invariant topologique.

Proposition 3.1.25 *Soient X et Y deux espaces topologiques et f une fonction continue de X dans Y. L'image par f d'un ensemble connexe est connexe.*

Par conséquent, si X a n composantes connexes alors $f(X)$ a au plus n composantes connexes. Si de plus f est un homéomorphisme alors l'application f^{-1} vérifie la même relation.

Proposition 3.1.26 *Si X et Y sont homéomorphes alors ils ont le même nombre de composantes connexes. En notant par $\pi_0(X)$ le nombre de composantes connexes de l'ensemble X, on écrit succinctement :*

$$X \simeq Y \Rightarrow \pi_0(X) = \pi_0(Y). \tag{3.10}$$

La figure 3.10 illustre cette dernière proposition.

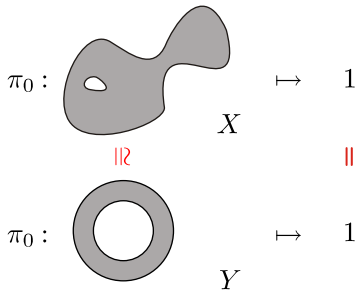


FIG. 3.10 – Deux ensembles homéomorphes ont le même nombre de composantes connexes .

La réciproque de cette implication n'est pas vraie. En effet, il existe des ensembles X et Y qui ne sont pas homéomorphes mais qui ont le même nombre de composantes connexes comme le montre la figure 3.11.

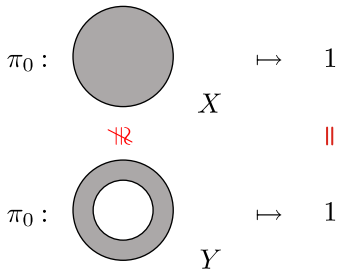


FIG. 3.11 – Deux ensembles non homéomorphes qui ont le même nombre de composantes connexes .

Le π_0 d'un ensemble n'est donc pas un invariant assez fort, puisqu'il ne permet pas à coup sûr de distinguer d'un point de vue topologique deux ensembles X et Y. La plupart des invariants créés jusqu'à aujourd'hui sont de nature algébrique.[3]. Un autre invariant topologique est le groupe fondamental, nous allons l'introduire dans la section suivante.

3.1.3 Le groupe fondamental

Dans cette section, nous allons nous intéresser à la construction de ce groupe. On note par S^1 l'image de $\mathbb{R}$ dans $\mathbb{R}^2$ via la fonction $t \mapsto (\cos t, \sin t)$.

[3]La discipline qui tente de classer les ensembles modulo les homéomorphismes s'appelle la *topologie algébrique*

Définition 3.1.27
Une fonction continue de S^1 à valeur dans Y est un *lacet*.
Soit y un élément de l'ensemble Y ; on dit que le lacet f est pointé en y si $f(0) = y$.

Définition 3.1.28
Soient y un élément de l'ensemble Y, et f_1 et f_2 deux lacets pointés de Y. On note par $f_1 \star f_2$ la fonction :

$$\begin{array}{rcll} f_1 \star f_2(t) & = & f_1(2t) & \text{si } t \in [0, \pi[\\ f_1 \star f_2(t) & = & f_2(2(t-\pi)) & \text{si } t \in [\pi, 2\pi[\end{array} \tag{3.11}$$

Par construction, la fonction $f_1 \star f_2$ est encore un lacet pointé. Ce lacet est la concaténation des deux lacets f_1 et f_2. Intuitivement, on "multiplie" par deux la vitesse de parcours des lacets f_1 et f_2 et on les "colle", de telle manière à construire une fonction de S^1 à valeur dans Y.

Proposition 3.1.29 *1. L'ensemble $\{f : S^1 \to Y, f$ continue $, f(0) = y\}$ est un groupe pour la loi de composition $\star$.*

2. La loi de composition $\star$ est compatible avec la relation d'équivalence "être homotopes" $:\simeq$.

De la dernière proposition, on peut déduire que $\{f : S^1 \to Y, f$ continue $, f(0) = y\}/\simeq$ est un groupe. On peut montrer que si Y est connexe par arcs alors ce groupe ne dépend pas de y. Il est appelé le groupe fondamental de Y et est noté $\pi_1(Y)$. C'est un invariant topologique, deux ensembles homéomorphes connexes par arcs ont des groupes fondamentaux qui sont isomorphes. Il existe bien d'autres invariants topologiques : on peut citer les groupes d'homologie et de cohomologie (qui ne sont pas que des groupes), les nombres de Betti *etc* ...

3.2 Etat de l'art et contributions

3.2.1 Etat de l'art.

Comme indiqué dans l'introduction, il existe de multiples façons de calculer plus ou moins efficacement des invariants topologiques. Ici, nous présentons ces méthodes en nous appuyant sur la manière dont sont décrits les ensembles.

Le cas polygonal de $\mathbb{R}^2$ On commence par une classe très fine d'ensembles. Dans ce paragraphe, on s'intéresse aux parties de $\mathbb{R}^2$ décrites à l'aide de segments.

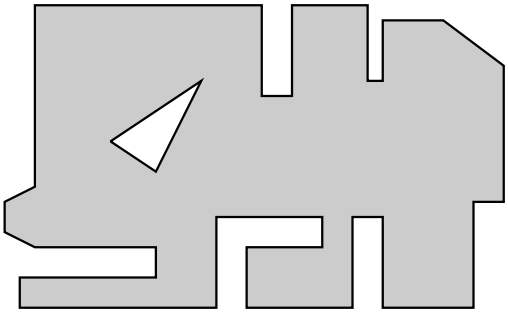


FIG. 3.12 – Une partie de $\mathbb{R}^2$ décrite à l'aide de segments.

La plupart des travaux essaie de décomposer l'ensemble en parties convexes comme le montre la figure 3.13. Chacune de ces parties est appelée cellules.

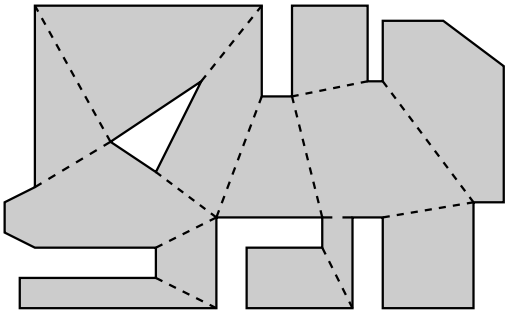


FIG. 3.13 – Décomposition à l'aide de cellules qui sont convexes.

Lorsque l'ensemble n'a pas de trou, *i.e.* avec un π_1 nul, il est possible de créer la composition qui minimise le nombre de cellules en un temps polynômial avec le nombre de sommets. De plus, cet algorithme produit un nombre de cellules proportionnel au nombre de sommets. Cependant, Lingas [42] a montré que la présence "de trous" dans l'ensemble rend le problème NP-difficile. Pour notre problème de topologie, connaître la décomposition qui minimise le nombre de cellules importe peu.

La méthode de Chazelle est sans doute la plus connue. Elle permet de construire une décomposition ayant un nombre linéaire de cellules avec le nombre de sommets n, et peut être obtenue en un temps proportionnel à $n \log n$. L'idée est de construire un segment par noeud qui a pour extrémités des éléments de la frontière de l'ensemble. Cette décomposition est appelée *décomposition verticale ou trapézoïdale* de Chazelle [28].

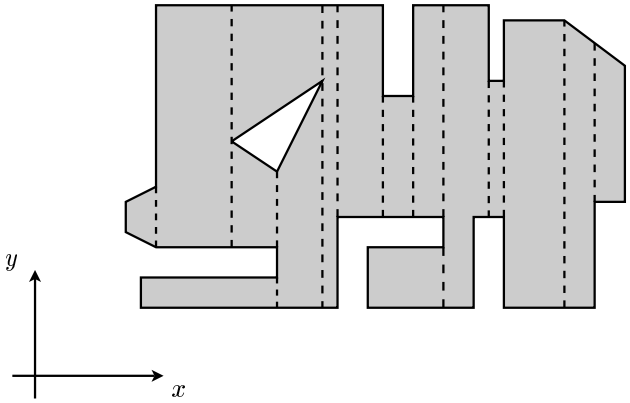


FIG. 3.14 – Décomposition de Chazelle.

Le cas polynomial Dans le cas polynômial, on s'intéresse à des ensembles décrits par des inégalités polynomiales. En définitive, on s'intéresse à des ensembles de la forme :

$$S = \bigcup_{i=1}^{s} \bigcap_{j=1}^{r_i} \{x \in \mathbb{R}^n; f_{i,j}(x) \diamond_{i,j} 0\} \text{ où } f_{i,j} \in \mathbb{R}[X_1, \ldots, X_n] \text{ et } \diamond_{i,j} \in \{=, \leq, <\} \tag{3.12}$$

Les ensembles de cette classe (3.12) forment les ensembles semi-algébriques. L'algorithme de Collins [29] peut renseigner sur leur topologie. L'objectif de cette méthode de créer une décomposition cylindrique adaptée à S. La notion de décomposition cylindrique est définie par récurrence finie sur n.

Définition 3.2.1

Une décomposition cylindrique de $\mathbb{R}^n$ est une partition finie de $\mathbb{R}^n$ en cellules $(C_i)_i$ vérifiant les propriétés suivantes :

- n=1 : une décomposition cylindrique de $\mathbb{R}$ est donnée par une subdivision de $a_1 < \cdots < a_l$ de $\mathbb{R}$. Les cellules sont les singletons $\{a_i\}$ avec $0 < i \leq l$, et les intervalles $]a_i, a_{i+1}[$ avec $0 \leq i \leq l$, où $a_0 = -\infty$ et $a_{n+1} = +\infty$.
- $n > 1$: une décomposition cylindrique de $\mathbb{R}^n$ est donnée par
  - une décomposition cylindrique de $\mathbb{R}^{n-1}$,
  - pour chaque cellule D de la décomposition de $\mathbb{R}^{n-1}$, sont données $l(D)$ fonctions semi-algébriques[4] :

$$\zeta_{D,1} < \cdots < \zeta_{D,l(D)} : D \to \mathbb{R} \tag{3.13}$$

[4]Avec $X \subset \mathbb{R}^n$ et $Y \subset \mathbb{R}^m$ deux ensembles semi-algébriques, une fonction $f : X \to Y$ est dite semi-algébrique si son graphe $\{(x, y) \in X \times Y; y = f(x)\}$ est semi-algébrique.

Les cellules de la décomposition cylindrique de $\mathbb{R}^n$ sont les graphes des applications $\zeta_{D,i}$:

$$\{(x, \zeta_{D,i}(x)) = 0; x \in D\},\ 0 < i \leq l(D) \tag{3.14}$$

et les bandes

$$]\zeta_{D,i}, \zeta_{D,i+1}[= \{(x, y) \in D \times \mathbb{R}; \zeta_{D,i}(x) < y < \zeta_{D,i+1}(x)\},\ 0 < i \leq l(D) \tag{3.15}$$

avec $\zeta_{D,0} = -\infty$ et $\zeta_{D,l(D)+1} = +\infty$.

Exemple 3.2.2

La figure 3.15 montre une décomposition cylindrique de $\mathbb{R}^2$. La décomposition de $\mathbb{R}^1$ dans cette figure est donnée par les cellules :

$$D_1 = \{]-\infty, a_1[, \{a_1\},]a_1, a_2[, \{a_2\},]a_2, a_3[, \{a_3\},]a_3, a_4[, \{a_4\},]a_4, a_5[, \{a_5\},]a_5, a_6[, \{a_6\},]a_6, +\infty[\} \tag{3.16}$$

Deux fonctions $\zeta_{]a_5,a_6[,1}$ et $\zeta_{]a_5,a_6[,2}$ sont définies sur la cellule $]a_5, a_6[$. Au dessus de cette cellule $]a_5, a_6[$, il y a cinq cellules de la décomposition de $\mathbb{R}^2$ qui sont les trois bandes et les deux graphes suivants :

- $\{(x, y); y < \zeta_{]a_5,a_6[,1}(x)\}$
- $\{(x, y); y = \zeta_{]a_5,a_6[,1}(x)\}$
- $\{(x, y); \zeta_{]a_5,a_6[,2}(x) \leq y < \zeta_{]a_5,a_6[,2}(x)\}$
- $\{(x, y); y = \zeta_{]a_5,a_6[,2}(x)\}$
- $\{(x, y); \zeta_{]a_5,a_6[,2}(x) < y\}$

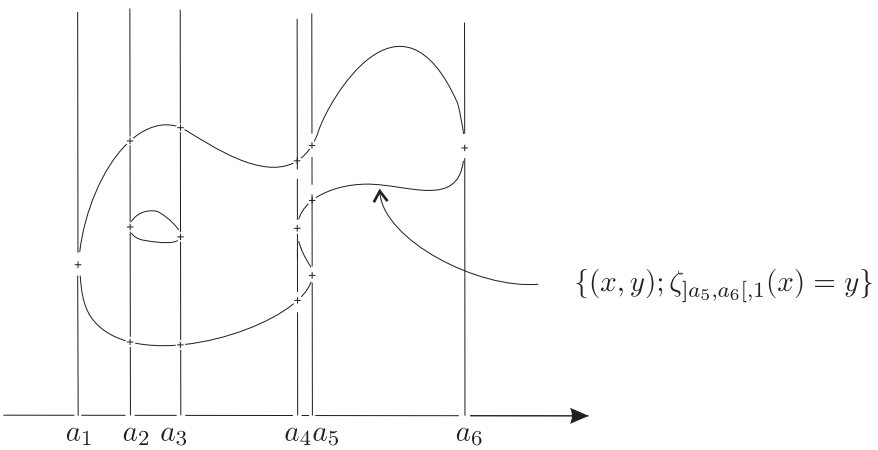


FIG. 3.15 – Une décomposition cylindrique de $\mathbb{R}^2$.

Etant donnée une famille finie $\mathcal{P}_n$ de polynômes de $\mathbb{R}[X_1, \ldots, X_n]$, l'algorithme de Collins construit une décomposition cylindrique algébrique de $\mathbb{R}^n$ telle que tout polynôme de $\mathcal{P}_n$ soit de signe constant sur chacune des cellules, une telle décomposition est dite adaptée à $\mathcal{P}_n$.

C'est un algorithme récursif ; l'idée de base est, à partir d'une famille $\mathcal{P}_n$ de polynômes de $\mathbb{R}[X_1, \ldots, X_n]$, de créer une famille $\mathcal{P}_{n-1}$ de polynômes de $\mathbb{R}[X_1, \ldots, X_{n-1}]$ telle qu'une décomposition adaptée à $\mathcal{P}_{n-1}$ puisse servir de base à une décomposition adaptée à $\mathcal{P}_n$. L'algorithme 3 présente globalement cette méthode.

Alg. 3 `Collins` - *Décomposition cylindrique de Collins*

Entrée: une famille finie $\mathcal{P}_n$ de polynômes de $\mathbb{R}[X_1, \ldots, X_n]$.

Sortie: Une décomposition algébrique cylindrique adaptée à $\mathcal{P}_n$.

1: **si** n=1 **alors**
2: Retourner une décomposition cylindrique D_1 de la famille $\mathcal{P}_1$.
3: **sinon**
4: Construire une famille de polynômes $\mathcal{P}_{n-1}$ de $\mathbb{R}[X_1, \ldots, X_{n-1}]$ telle qu'une décomposition cylindrique adaptée à $\mathcal{P}_{n-1}$ puisse servir de base à une décomposition cylindrique adaptée à $\mathcal{P}_n$.
5: $D_{n-1} \leftarrow$ `Collins`$(\mathcal{P}_{n-1})$.
6: Pour chaque cellule C de la décomposition D_{n-1}, calculer la décomposition du cylindre au dessus de C induite par les polynômes de $\mathcal{P}_n$.
7: Ajouter toutes ces cellules à D_n, et retourner D_n.
8: **fin si**

L'étape 4 est sans aucun doute la plus compliquée. L'étape 6 n'est pas difficile, si P_{n-1} a bien été choisi, chaque élément de $\mathcal{P}_n$ (vu comme un polynôme en x_n) possède un nombre constant de racines réelles sur chaque cellule C, et les surfaces définies par ces racines ne se coupent pas. La difficulté réside donc en la création de la famille $\mathcal{P}_{n-1}$. De ce point de vue, on peut dire que l'algorithme de Collins est une machine à créer des polynômes. Dans cette partie, nous ne présentons qu'une version limitée aux ensembles semi-algébriques de $\mathbb{R}^2$. Le cas général est un peu plus compliqué mais les grands principes restent les mêmes.

Etant donnée une famille $\mathcal{P}_2$ de polynômes à deux variables, on doit trouver une famille $\mathcal{P}_1$ telle qu'une décomposition D_1 de $\mathbb{R}$ adaptée à $\mathcal{P}_1$ puisse servir de base à une décomposition adaptée à $\mathcal{P}_2$. Les contraintes que doivent satisfaire les cellules

de D_1 sont les suivantes :

1. Pour chaque cellule C de D_1, tous les polynômes P de $\mathcal{P}_2$, vus comme des polynômes de la variable x_2, possèdent un nombre constant de racines.
2. Pour chaque cellule ouverte C de D_1, les courbes définies par les racines des éléments de $\mathcal{P}_2$ ne se coupent pas.

On reprend ces deux contraintes en énumérant les conditions que doit satisfaire D_1 :

1. Pour la première contrainte, il existe deux possibilités pour que le nombre de racines réelles d'un polynôme $P(x_1, x_2)$ vu comme un polynôme en la variable[5] x_2 puisse changer :

   (a) *une racine réelle devient complexe*, comme les racines complexes d'un polynôme à coefficients réels vont toujours par paires, une racine réelle seule ne peut disparaître. Deux racines d'un polynôme $P \in \mathbb{R}[X_1][X_2]$ de la variable X_2 peuvent disparaître seulement s'il existe un x_1 particulier, disons x_{1_0}, tel que $P(x_{1_0}, \cdot)$ ait une racine double. Par conséquent, les singletons $\{x_1\}$ de $\mathbb{R}$ qui vérifient :

   $$\exists x_2 \in \mathbb{R}, P(x_1, x_2) = 0 = \frac{\partial P}{\partial x_2}(x_1, x_2) \tag{3.17}$$

   doivent appartenir à D_1. La figure 3.16 illustre ce phénomène.

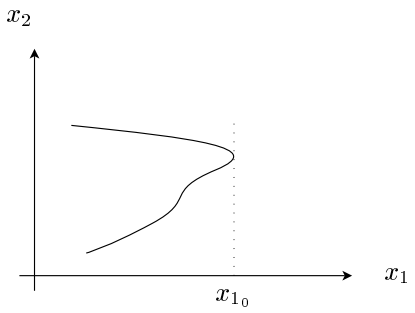


FIG. 3.16 – Lorsque $x_1 < x_{1_0}$, le polynôme $P(x_1, x_2)$, vu comme un polynôme en x_2, a deux racines distinctes. Lorsque $x_1 = x_{1_0}$, il a une racine double.

   (b) *une racine x_2 réelle devient infinie* ; l'équation $P(x_1, x_2) = 0$ s'écrit $a_n(x_1)x_2^n + \cdots + a_0(x_1) = 0$, avec a_i des éléments de $\mathbb{R}[X_1]$, et donc

[5] On considère donc ici le polynôme P comme un élément de $\mathbb{R}[X_1][X_2]$, c'est-à-dire un polynôme de la variable x_2 à coefficient dans $\mathbb{R}[X_1]$.

toujours à valeur fini, il est évident que $a_n(x_1)$ doit s'annuler pour fournir une racine de $P(x_1, x_2) = 0$ qui tende vers l'infini. Par conséquent, la décomposition D_1 doit contenir les singletons $\{x_1\}$ tels que x_1 soit solution de $a_n(x_1) = 0$. La figure 3.17 illustre cette situation.

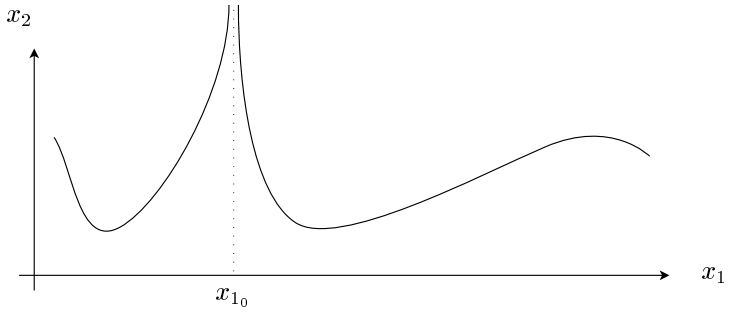


FIG. 3.17 – Une racine réelle devient infinie.

2. Il y deux possibilités pour que deux courbes se coupent :

   (a) *les deux courbes sont les racines d'un même polynôme,* dans ce cas, cela signifie que $P(x_1, x_2)$ a une racine multiple. La proposition 3.17 doit donc aussi être vérifiée. La figure 3.18 illustre cette situation.

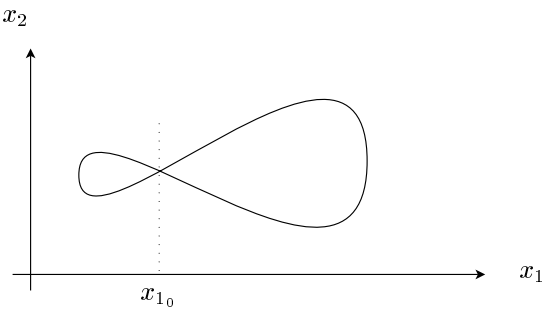


FIG. 3.18 – Le polynôme $P(x_1, x_2)$ a une racine multiple x_{2_0} lorsque $x_1 = x_{1_0}$.

   (b) *les deux courbes sont les racines de deux polynômes P et Q,* par conséquent les singletons $\{x_1\}$ de $\mathbb{R}$ qui vérifient

$$\exists x_2 \in \mathbb{R}, P(x_1, x_2) = 0 = Q(x_1, x_2) \tag{3.18}$$

   doivent être des éléments de D_1. La figure 3.19 illustre cette situation.

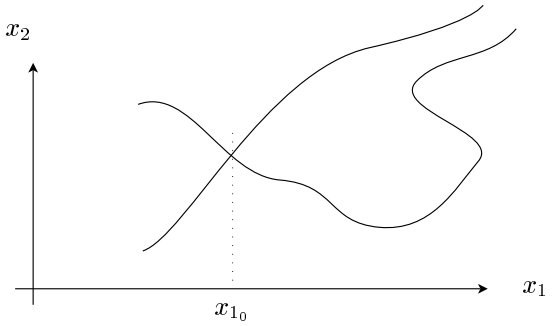


FIG. 3.19 – Les deux polynômes $P(x_1, x_2)$ et $Q(x_1, x_2)$ ont une racine en commun lorsque $x_1 = x_{1_0}$.

Après cette analyse géométrique, il manque un ingrédient pour être consistant. En effet, comment trouver des polynômes univariés (qui formeront $\mathcal{P}_1$) de la variable x_1 tels que leurs zéros contiennent les éléments de D_1 énumérés précédemment. Pour la situation 1(b), il suffit d'ajouter $a_n(x_1)$ à $\mathcal{P}_1$, pour les autres situations, il faut que $\mathcal{P}_1$ contienne des polynômes qui s'annulent lorsque (3.17) ou (3.18) est vérifiée.

Globalement, les assertions (3.17) ou (3.18) reviennent à décider si deux polynômes univariés (de la variable x_2) ont une racine en commun. C'est à mon sens ici que la géométrie algébrique réelle prend naissance. En effet, bien que Galois ait démontré qu'il n'existe pas de méthode par radicaux pour trouver les racines de n'importe quel polynôme donné, il existe une méthode algébrique permettant de décider si deux polynômes ont une racine en commun. Elle est basée sur le raisonnement suivant : deux polynômes univariés P et Q (de degré p et q) ont une racine en commun si et seulement s'il existe deux polynômes $\tilde{P}$ et $\tilde{Q}$ (de degré $p-1$ et $q-1$) tels que $P = (X-\alpha)\tilde{P}$ et $Q = (X-\alpha)\tilde{Q}$. De façon équivalente, on peut dire que P et Q ne sont pas premiers entre eux. Comme $p+q = \deg(P \vee Q)\deg(P \wedge Q)$, P et Q ont une racine en commun si et seulement si $\deg(P \vee Q) < p+q$. De plus comme pour tous polynômes P et Q il existe des polynômes U et V tels que $PU = QV = P \vee Q$, P et Q ont une racine en commun s'il existe U et V *non nuls* de degré strictement inférieur respectivement à q et à p tel que $PU = VQ$. Si on note par $\mathbb{R}_n[X]$ l'espace vectoriel de dimension n des polynômes de degré inférieur à $n-1$, de façon équivalente, l'application linéaire

$$\begin{array}{rcccl} Sylvester_{P,Q} & : & \mathbb{R}_p[X] \times \mathbb{R}_q[X] & \rightarrow & \mathbb{R}_{p+q}[X] \\ & & (U,V) & \mapsto & PU - QV \end{array} \tag{3.19}$$

a un déterminant nul.

Revenons maintenant à l'algorithme de Collins, plus particulièrement, à la condition 2 (b) de l'analyse géométrique. Si l'on considère les polynômes P et Q comme des polynômes de la variable x_2 paramétrés par x_1 (*i.e.* des éléments de $\mathbb{R}[X_1][X2]$), on peut donc parler de l'application linéaire $Sylvester_{P(x_1,\ .\ ),Q(x_1,\ .\ )}$ paramétrée par x_1. On remarque que le déterminant d'une matrice est une application polynomiale en ses coefficients, et que les coefficients de la matrice de $Sylvester_{P(x_1,\ .\ ),Q(x_1,\ .\ )}$ dans la base canonique, sont des polynômes en x_1. En définitive, il suffit d'ajouter le polynôme en x_1 : $\det Sylvester_{P(x_1,\ .\ ),Q(x_1,\ .\ )}$ à $\mathcal{P}_1$ pour que la contrainte 2 (b) soit vérifiée.

On peut remarquer une grande croissance de la taille des données qui interviennent. Si la famille $\mathcal{P}_2$ contient m polynômes de degré inférieur à d, alors $\mathcal{P}_1$ contient $O(m^2)$ polynômes et de degré borné par $O(d^2)$.

Dans le cas général, avec une famille $\mathcal{P}_n$ de m polynômes à n variables dont le degré est borné par d, le temps de calcul de l'algorithme de Collins est donné par :

$$(2d)^{2^{2n+8}} m^{2^{2n+6}} \tag{3.20}$$

L'implémentation de cet algorithme est quasiment inutile à cause de la complexité doublement exponentielle du temps de calcul avec la dimension. De plus, il n'est possible de l'implémenter que dans le cas où l'on peut représenter exactement et facilement manipuler les coefficients des polynômes. A ma connaissance, cet algorithme n'a été implémenté que pour des polynômes de $\mathbb{Q}[X_1, \ldots, X_n]$.

Géométrie o-minimale La précédente méthode de Collins peut être étendue à des classes d'ensembles plus larges que l'ensemble des semi-algébriques. Si une famille $\mathcal{S} = (\mathcal{S}^n)_{n\in\mathbb{N}}$ satisfait les conditions suivantes :

1. $\mathcal{S}^n$ est une famille de sous-ensembles de $\mathbb{R}^n$.
2. la famille des semi-algébriques de $\mathbb{R}^n$ est incluse dans $\mathcal{S}^n$ pour tout $n \in \mathbb{N}$.
3. pour tout $n \in \mathbb{N}$, $\mathcal{S}^n$ est stable par intersection, union et complémentaire.
4. si $A \in \mathcal{S}^n$ et $B \in \mathcal{S}^m$ alors $A \times B \in \mathcal{S}^{m+n}$.
5. On note par $p : \mathbb{R}^n \to \mathbb{R}^{n-1}$ la projection $(x_1, \ldots, x_n) \mapsto (x_1, \ldots, x_{n-1})$, si $A \in \mathcal{S}^{n+1}$ alors $p(A) \in \mathcal{S}^n$
6. Les éléments de S^1 sont des unions finies de singletons et d'intervalles ouverts,

elle est qualifiée de *structure o-minimale*. Par exemple, les ensembles semi-algébriques forment une structure o-minimale[6]. Les ensembles d'une structure sont appelés

[6]Dans le cas des semi algébriques, on a de la chance (Tarski-Seidenberg), ils sont stables par projection. Cette propriété s'appelle élimination des quantificateurs. Elle s'exprime aussi en

définissables.

Les conditions qui définissent ces structures sont intéressantes, puisqu'à partir de celles-ci, il est possible de montrer :

1. pour tout ensemble d'une structure o-minimale, il existe une décomposition cylindrique (composée d'éléments de cette structure) adaptée à cet ensemble.
2. tous les ensembles définissables ont un nombre fini de composantes connexes.
3. les définitions de connexité et de connexité par arcs coïncident.

Ces structures sont en quelque sorte rassurantes puisqu'elles excluent les ensembles qui ont des géométries délicates comme celui présenté dans l'annexe A.

La famille d'ensembles

$$\left\{ S = \bigcup_{i=1}^{s} \bigcap_{j=1}^{r_i} \{x \in \mathbb{R}^n ; f_{i,j}(x) \ \diamond_{i,j} \ 0\} \text{ où les } f_{i,j} \right. \tag{3.21}$$

$$\text{sont analytiques et à support compact } \diamond_{i,j} \in \{=, \leq, <\}\}$$

est appelée la famille des *semi-analytiques* et ne forment pas une structure o-minimale. Elle n'est pas stable par projection. Par contre, si on considère la famille des semi-analytiques et leurs projections, on crée la famille des *sous-analytiques*, et cette classe est une structure o-minimale[7].

De même, la famille d'ensembles

$$\left\{ S = \bigcup_{i=1}^{s} \bigcap_{j=1}^{r_i} \{x \in \mathbb{R}^n ; f_{i,j}(x) \ \diamond_{i,j} \ 0\} \text{ où les } f_{i,j} \subset \mathcal{C}^\infty, \diamond_{i,j} \in \{=, \leq, <\} \right\} \tag{3.22}$$

ne forment pas une structure o-minimale. Par contre, A. Grigoriev a montré que l'ensemble des n fonctions $\mathcal{C}^\infty$ qui engendre une structure o-minimale est une partie résiduelle de $\mathcal{C}^{\infty n}$. Autrement dit de façon moins rigoureuse, si on prend au hasard n fonctions C^∞ sur un compact, on a toutes les chances pour que la structure engendrée soit o-minimale.

Ce dernier paragraphe nous montre que les hypothèses sur les fonctions qui définissent nos ensembles ne sont pas à prendre à la légère. Tout au long de cette section, nous nous sommes écartés de l'implémentatibilité. En particulier, même si tout ensemble définissable admet une décomposition cellulaire, l'implémentation

disant que n'importe quel énoncé du premier ordre est équivalent à un énoncé du premier ordre sans quantificateur. Autrement dit, résoudre un problème général est équivalent à faire un calcul algébrique, et vérifier des signes.

[7] On doit principalement ce résultat à Gabrielov qui a montré que les sous-analytiques sont stables par complémentaires.

de l'algorithme de Collins semble très difficile. Il manque par exemple les outils algébriques pour décider si deux fonctions ont un zéros en communs. En définitive, le fait qu'une famille forme une structure o-minimale nous permet de dire que ses ensembles "ne sont pas trop compliqués", mais ne nous donne pas d'algorithme implémentable sur machine permettant de connaître leurs topologies.

Position de nos contributions Avec le calcul par intervalles, on l'a vu dans le second chapitre, il est possible de traiter des problèmes qui font intervenir des inégalités (et des égalités) $\mathcal{C}^\infty$. Dans la fin de ce chapitre, on propose deux algorithmes qui permettent de renseigner sur la topologie d'ensembles décrits par des inégalités $\mathcal{C}^\infty$. Ces méthodes sont voisines et sont basées sur une décomposition (créée avec un recouvrement) en éléments dont on connaît bien la topologie. Les algorithmes qui en découlent sont souvent capables de renseigner sur la topologie des ensembles de la famille suivante :

$$\left\{ S = \bigcup_{i=1}^{s} \bigcap_{j=1}^{r_i} \{x \in \mathbb{R}^n ; f_{i,j}(x) \diamond_{i,j} \ 0\} \text{ où les } f_{i,j} \subset \mathcal{C}^\infty, \diamond_{i,j} \in \{=, \leq, <\} \right\} \quad (3.23)$$

Bien que les algorithmes qui vont être explicités ne terminent pas pour tout ensemble de cette famille, ils ont le mérite d'être implémentés.

3.3 Connexité - Algorithme C.I.A.

3.3.1 Introduction

Dans cette section, on présente une méthode qui a pour objectif de compter le nombre de composantes connexes d'un ensemble décrit par

$$\left\{ S = \bigcup_{i=1}^{s} \bigcap_{j=1}^{r_i} \{x \in \mathbb{R}^n ; f_{i,j}(x) \diamond_{i,j} \ 0\} \text{ où les } f_{i,j} \subset \mathcal{C}^\infty, \diamond_{i,j} \in \{=, \leq, <\} \right\} \quad (3.24)$$

L'idée maîtresse est de décomposer celui-ci avec des ensembles étoilés en usant des résultats de la section 3.3.2 puis de recoller les morceaux avec un graphe (voir 3.3.3). Globalement, on ramène le problème de compter le nombre de composantes connexes par arc d'un ensemble S à celui de compter le nombre de composantes connexes d'un graphe. D'une certaine manière, on a discrétisé notre problème.

3.3.2 Condition suffisante pour qu'un ensemble soit étoilé.

Cette section montre que, lorsqu'un ensemble S est défini par des inégalités ($S \subset \mathbb{R}^n$), montrer que S est v^*-étoilé revient à prouver qu'un système d'inéquations est inconsistant. Ce résultat est motivant car démontrer l'inconsistance d'un système d'inéquations peut être vérifié grâce aux méthodes présentées dans le chapitre 2.

Théorème 3.3.1 *Soit $S = \{x \in D \subset \mathbb{R}^n | f(x) \leq 0\}$ où f est une fonction C^1 de D vers $\mathbb{R}$, et E un sous-ensemble convexe. Soit v^* un élément de S. Si*

$$f(x) = 0,\ Df(x) \cdot (x - v^*) \leq 0,\ x \in E \tag{3.25}$$

est inconsistant, alors v^* est une étoile pour S.

Preuve :

La preuve est basée sur un raisonnement par l'absurde. Supposons que v^* ne soit pas une étoile pour S, alors il existe $x_0 \in S$ tel que le segment $[v^*, x_0] \not\subset S$. Comme E est convexe, il existe donc $x_1 \in [v^*, x_0]$ tel que $f(x_1) > 0$. Soit g la fonction : $g : [0,1] \to \mathbb{R},\ \ t \mapsto g(t) = f((1-t)v^* + tx_0)$. Comme la fonction réelle f est $\mathcal{C}^1$, g est différentiable. De plus, elle satisfait l'inégalité suivante : $g(0) \leq 0,\ g(1) \leq 0,\ g(t_1) > 0$ où t_1 est tel que $x_1 = (1-t_1)v^* + t_1 x_0$.
Comme g est continue, le théorème des valeurs intermédiaires garantit qu'il existe $t_2 \in [t_1, 1]$ tel que $g(t_2) = 0$. Dans le cas où plusieurs réels de $[t_1, 1]$ vérifient $g(t) = 0$, on note t_2 le plus petit d'entre eux. Par conséquent, on a $g(t_2) = 0$ and $\forall t \in]t_1, t_2[,\ \ g(t) > 0$. Comme g est différentiable sur l'intervalle ouvert $]0,1[$, on a :

$$g'(t_2) = \lim_{h \to 0} \frac{g(t_2 + h) - g(t_2)}{h} = \lim_{h \to 0^-} \frac{g(t_2 + h)}{h}. \tag{3.26}$$

Il existe $\epsilon > 0$ tel que $\forall h < 0, |h| < \epsilon \Rightarrow t_2 + h \in [t_1, t_2]$ (prendre $\epsilon = (t_1 - t_2)/2$). D'où :

$$\forall h < 0, |h| < \epsilon, \frac{g(t_2 + h)}{h} < 0. \tag{3.27}$$

On en déduit $g'(t_2) \leq 0$. Finalement, en prenant $x_2 = (1-t_2)v^* + t_2 x_0$, $x_2 \in E$, il est tel que : $f(x_2) = 0, Df(x_2) \cdot (x_2 - v^*) \leq 0$.

□

Une interprétation géométrique de ce théorème est qu'un ensemble est v^*-étoilé si les rayons lumineux émis par v^* traversent le bord ∂S de S au plus une fois (de l'intérieur vers l'extérieur).

Exemple 3.3.2

On souhaite montrer que le point v_1 de coordonnées $(0, 0.7)$ est une étoile pour le sous ensemble S de $\mathbb{R}^2$ défini par $f(x_1, x_2) \leq 0$ où f est une fonction C^1 de $\mathbb{R}^2$ vers $\mathbb{R}$ dont l'expression est : $f(x_1, x_2) = -e^{-(2x_1)^2} - e^{-(2x_1-2.8)^2} + 0.1 + x_2^2$.

En utilisant le théorème précédent 3.3.1, v_1 est une étoile pour S si

$$\begin{cases} \partial_1 f(x_1, x_2).(x_1 - 0) + \partial_2 f(x_1, x_2).(x_2 - 0.7) \leq 0 \\ f(x_1, x_2) = 0 \end{cases} \tag{3.28}$$

est inconsistant. Le gradient $\nabla f(x)$ et les rayons lumineux $x - v_1$ sont représentés sur la frontière S ($\{f(x) = 0\}$) sur la figure 3.20.

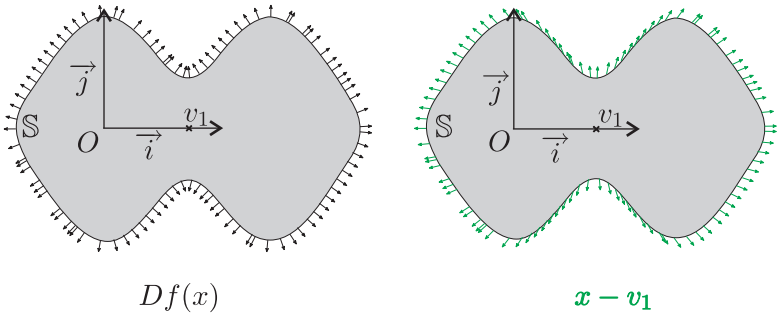


FIG. 3.20 – Champ de vecteurs qui représentent respectivement $\nabla f(x)$ et $x - v_1$ sur la frontière de S.

La figure 3.21 illustre que pour tout x vérifiant $f(x) = 0$, on a $Df(x)\cdot(x - v_1) > 0$, i.e. l'angle formé par les deux vecteurs est aigu. Autrement dit, tous les rayons lumineux traversent le bord de l'intérieur vers l'extérieur.

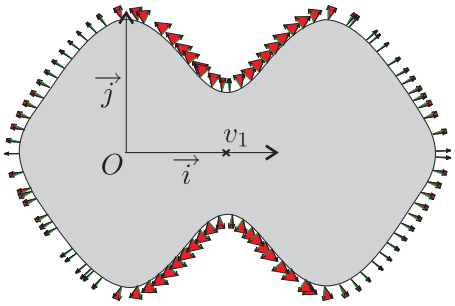


FIG. 3.21 – Les rayons lumineux traversent le bord de l'intérieur vers l'extérieur.

La figure 3.22 montre la contraposée du théorème précédent : v_2 n'est pas une étoile pour S et il existe $x \in \mathbb{R}^2$ tel que $f(x) = 0$ et $Df(x) \cdot (x - v_2) \leq 0$.

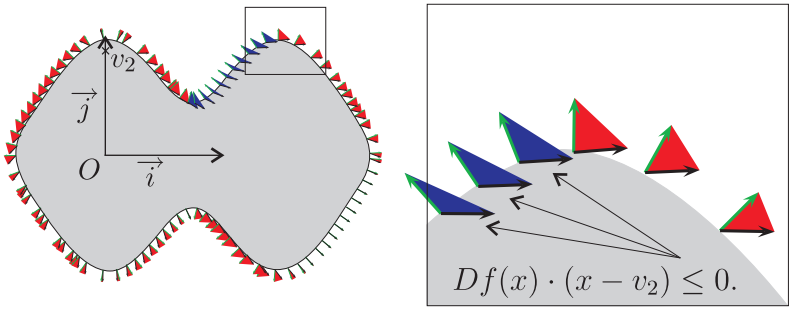


FIG. 3.22 – v_2 n'est pas une étoile.

Remarque 3.3.3

Pour vérifier qu'un point v^* est une étoile pour S, il suffit donc de montrer que l'ensemble des x de E qui vérifient $f(x) = 0$ et $Df(x) \cdot (x - v^*) \leq 0$ est vide. Ce qui est équivalent à montrer que $\forall x \in D, f(x) = 0 \Rightarrow Df(x) \cdot (x - v^*) > 0$.

L'assertion précédente peut être montrée en utilisant le calcul par intervalles. De plus la proposition 2.4.1 nous indique que l'algorithme qui vérifie cette implication termine si et seulement si cette assertion est vraie.

La condition du théorème 3.3.1 est une condition suffisante pour que le point v^* soit une étoile de S. La figure 3.23 montre que cette condition n'est pas nécessaire.

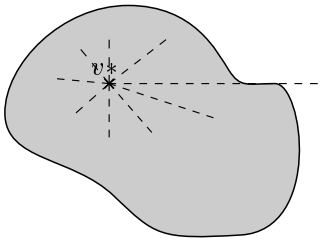


FIG. 3.23 – La condition du théorème 3.3.1 est nécessaire mais pas suffisante.

Autrement dit, l'algorithme employé pour vérifier que v^* est une étoile pour S peut ne pas terminer alors que v^* est une étoile. Le résultat suivant montre que cette situation est exceptionnelle. On note par U_1 l'ensemble des points v^* pour lesquels l'algorithme de bissection montre que S est v^*-étoilé et par V^* l'ensemble des étoiles de S.

Corollaire 3.3.4 *Sous les hypothèses du théorème 3.3.1, si* $\{0\}$ *est une valeur régulière de* f *et* $int(V^*) \neq \emptyset$ *alors l'ensemble* U_1 *est une* partie résiduelle[8] *de l'ensemble* V^*.

Preuve : On commence par montrer, par l'absurde, que si v est une étoile pour l'ensemble S alors l'assertion suivante est satisfaite :

$$\forall x \in S, f(x) = 0 \Rightarrow Df(x) \cdot (x - v) \geq 0 \tag{3.29}$$

Supposons qu'il existe $x_0 \in D$, tel que $f(x_0) = 0$ et $Df(x_0) \cdot (x_0 - v) < 0$. Alors sur le segment $[v, x_0]$, il existe x_1 dans un voisinage de x_0 tel que $f(x_1) < 0$. Par conséquent, v n'est pas une étoile.

Comme 0 est une valeur régulière de 0, $f^{-1}(\{0\})$ est une variété M de dimension $n-1$. On note par T_xM le plan tangent à M en x, et $T_x^< M$ l'ensemble (demi-espace) défini par

$$T_x^< M = \{v \in D \mid Df(x) \cdot (x - v) < 0\} \tag{3.30}$$

En notant par $\overline{A}$ l'adhérence de A, on a

$$\left(\underbrace{\bigcap_{x \in D, f(x)=0} T_x^< M}_{U_1} \right) \subset V^* \subset \left(\underbrace{\bigcap_{x \in D, f(x)=0} \overline{T_x^< M}}_{U_2} \right) \tag{3.31}$$

Comme la fonction f est continue, l'ensemble U_2 est un convexe fermé comme intersection de convexes fermés. On en déduit que U_1 contient son intérieur (comme intersection des intérieurs des fermés qui composent U_2). Par conséquent, on a montré que les convexes U_1 et U_2 vérifient

$$\emptyset \neq \overset{\circ}{V^*} \subset \overset{\circ}{U_2} \subset U_1 \subset V^* \subset U_2 \tag{3.32}$$

Ainsi, U_2 est un convexe de $\mathbb{R}^n$ qui a un point intérieur, l'ensemble de tous ces points intérieurs est donc dense dans U_2 (*i.e.* $\overline{\overset{\circ}{U_2}} = U_2$). En conclusion, U_1 contient un ouvert ($\overset{\circ}{U_2}$) de V^* qui est dense dans V^*. D'où U_1 est une partie résiduelle de V^*.

□

[8] Une partie X de E est dite résiduelle si X contient une intersection dénombrable d'ouverts denses dans E.

Remarque 3.3.5
La preuve précédemment proposée utilise la convexité des ensembles U_1 et U_2 ; on aurait pu conclure en utilisant le *théorème de transversalité faible* [12] pour étudier les droites affines qui passent par v^* et qui ne sont pas transverses à T_xM.

3.3.3 Discrétisation

Comme les ensembles étoilés sont aussi connexes par arcs, le théorème 3.3.1 nous donne une condition suffisante pour montrer qu'un ensemble est connexe par arcs. Mais la plupart des ensembles connexes ne sont pas étoilés comme le montre la figure 3.24 : il est impossible de trouver un point v^* qui éclaire tout l'ensemble ($V^* = \emptyset$).

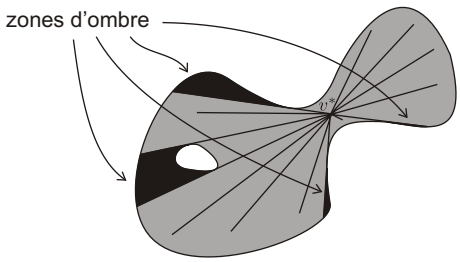


FIG. 3.24 – Exemple d'un ensemble connexe par arcs qui n'est pas étoilé.

L'idée de notre approche, pour montrer qu'un ensemble S est connexe par arcs, est de diviser notre ensemble à l'aide d'un pavage[9] [40] $\mathcal{P}$ tel que, sur chaque morceau p de $\mathcal{P}$, $S \cap p$ est étoilé (voir Figure 3.25).

[9]Un pavage est une collection finie d'intervalles de $\mathbb{R}^n$, le choix de ce type de découpage peut paraître restrictif. Notre objectif est de développer des algorithmes effectifs et un pavage est facilement représentable sur machine.

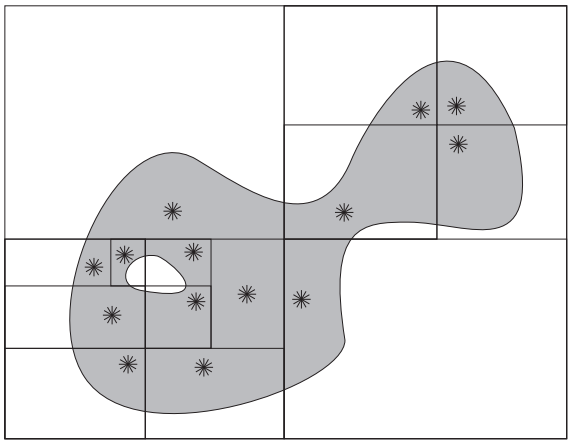


FIG. 3.25 – Exemple d'un pavage $\mathcal{P}$ vérifiant $\forall p \in \mathcal{P}, S \cap p$ est étoilé.

Pour "recoller les morceaux", nous introduisons la notion de graphe *parsemé d'étoiles*.

Définition 3.3.6

Un graphe *parsemé d'étoiles* d'un ensemble S, noté $\mathcal{G}_S$, est une relation $\mathcal{R}$ sur un pavage $\mathcal{P}$ où :

- $\mathcal{P}$ est un pavage, i.e. une collection de pavés (produit cartésien de n intervalles), $\mathcal{P} = (p_i)_{i \in I}$. De plus, pour chaque p de $\mathcal{P}$, $S \cap p$ est étoilé.
- $\mathcal{R}$ est la relation sur $\mathcal{P}$ définie par
  $p \, \mathcal{R} \, q \Leftrightarrow S \cap p \cap q \neq \emptyset$.[10]
- $S \subset \bigcup_{i \in I} p_i$

Par exemple, un graphe parsemé d'étoiles de S est donné par la figure 3.26. Sur cette figure, les noeuds du graphes représentent les éléments de $\mathcal{P}$.

[10] Ce sont les couples de pavés dont l'intersection intersecte S.

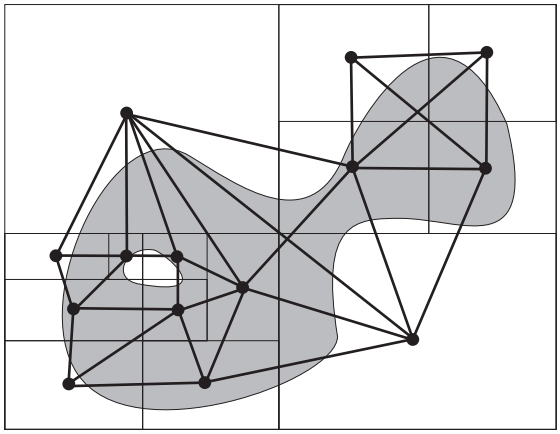


FIG. 3.26 – Un graphe parsemé d'étoiles $\mathcal{G}_S$.

Remarque 3.3.7

Il est facile de montrer que la relation $\mathcal{R}$ est une relation d'équivalence sur $\mathcal{P}$.

Définition 3.3.8

Le *support* d'un graphe parsemé d'étoiles $\mathcal{G}_S$ est le sous-ensemble P de $\mathbb{R}^n$ défini par : $P = \cup_{i \in I} p_i$.

Proposition 3.3.9 *Soit $\mathcal{G}_S$ un graphe parsemé d'étoiles d'un ensemble S.*
S est connexe par arcs $\Leftrightarrow \mathcal{G}_S$ est connexe.

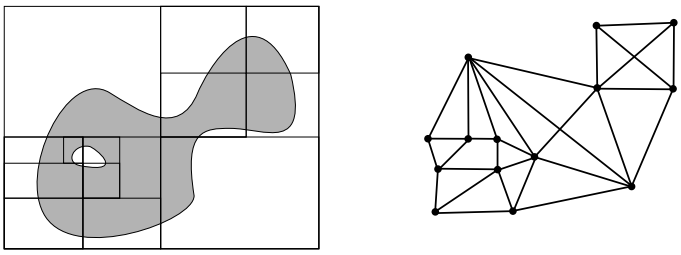


FIG. 3.27 – Si le graphe $\mathcal{G}_S$ est connexe alors S est connexe par arcs.

Preuve : Si $\mathcal{G}_S$ est connexe, alors il existe un chemin qui relie n'importe quelle paire de noeuds du graphe.

Soit n le nombre de noeuds , et $\mathcal{N} = (\alpha_i)_{i \in \{1,\ldots,n\}}$ les noeuds. Comme $\mathcal{G}_S$ est connexe, pour chaque i de $\{1, \ldots, n-1\}$, il existe un chemin reliant α_i à α_{i+1}, i.e. il

existe une suite finie $\{\alpha_{i_1}, \alpha_{i_2}, .., \alpha_{i_k}\} \in \mathcal{N}^k$ avec $(\alpha_{i_1}, \alpha_{i_2}), (\alpha_{i_2}, \alpha_{i_3}), \ldots, (\alpha_{i_{k-1}}, \alpha_{i_k})$ des arêtes de $\mathcal{G}_S$ (avec $\alpha_{i_1} = \alpha_i$, et $\alpha_{i_k} = \alpha_{i+1}$). Notons $p(\alpha_i, \alpha_{i+1})$ ce chemin.
Soient $path_1$ et $path_2$ deux chemins de $\mathcal{G}_S$.
Si l'une des extrémités de $path_1$ est aussi une extrémité de $path_2$, alors il est possible de créer un nouveau chemin à partir $path_1$ et $path_2$, noté $path_1 + path_2$, qui est la concaténation de $path_1$ et $path_2$.
Soit p_{all} le chemin défini par cette dernière opération associative :

$$p_{all} = p(\alpha_1, \alpha_2) + p(\alpha_2, \alpha_3) + \cdots + p(\alpha_{n-1}, \alpha_n). \tag{3.33}$$

Donc p_{all} est un chemin de $\mathcal{G}_S$ qui passe par chaque noeud du graphe au moins une fois. Soit $(\beta_i)_{i\in\{1,\ldots,m\}}$ la suite des noeuds visités par p_{all} avec $\beta_1 = \alpha_1$ et $\beta_m = \alpha_n$. La suite de pavés $(p_i)_{i\in\{1,\ldots,m\}}$, où p_i est le pavé associé au noeud β_i, vérifie :

$$\begin{cases} \forall i \in \{1, \ldots, m\}, \quad p_i \cap S \text{ est connexe par arcs } (p_i \cap S \text{ est étoilé}) \\ \forall i \in \{2, \ldots, m\}, \quad S \cap p_{i-1} \cap p_i \neq \emptyset. \end{cases} \tag{3.34}$$

En utilisant le résultat : "Pour toute famille dénombrable $(A_i)_{i\in I}$ d'ensembles connexes par arcs telle que [17] : $\forall i \in I\backslash\{0\}, A_{i-1} \cap A_i \neq \emptyset$ l'ensemble $\bigcup_{i\in I} A_i$ est connexe par arcs", on en déduit que

$$\bigcup_{i\in I}(S \cap p_i) = S \cap \bigcup_{i\in I} p_i = S \text{ est connexe par arcs.} \tag{3.35}$$

□

Corollaire 3.3.10 *Soit $\mathcal{G}_S$ un graphe parsemé d'étoiles d'un ensemble S.*
$\mathcal{G}_S$ a le même nombre de composantes connexes que S : $\pi_0(S) = \pi_0(\mathcal{G}_S)$.

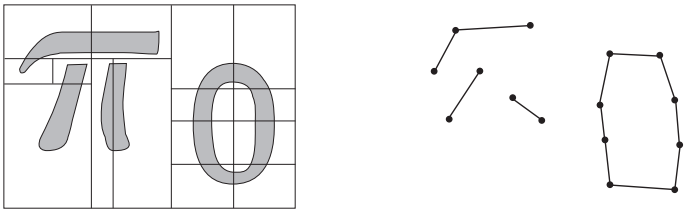


FIG. 3.28 – Le nombre de composantes connexes de $\mathcal{G}_S$ et de S coïncident.

Preuve : L'idée principale est de s'intéresser à chaque composante connexe de $(\mathcal{G}_i)_{1\leq i\leq n}$ de $\mathcal{G}_S$. (n est le nombre de composantes connexes $\mathcal{G}_S$.)

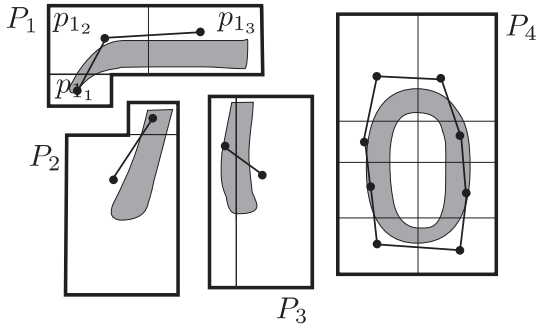


FIG. 3.29 – Sectionner le graphe parsemé d'étoiles $\mathcal{G}_S$ en suivant les composantes connexes.

Soient P_i le support de $\mathcal{G}_i$, et $\mathcal{P}_i = \{p_{i_j}\}_{1 \leq j \leq n_j}$. Pour chaque graphe parsemé d'étoiles $\mathcal{G}_i$, on peut appliquer le théorème 3.3.9, et affirmer que $S \cap P_i$ est connexe par arcs. Par conséquent, l'ensemble S a au plus n composantes connexes.
Par l'absurde, supposons que S a un nombre strictement inférieur à n de composantes connexes : il existe α, β dans $1, .., n$ tels que : $\alpha \neq \beta$ et $P_\alpha \cap P_\beta \cap S \neq \emptyset$ i.e. il existe $\alpha_0 \in 1, .., n_\alpha$ et $\beta_0 \in 1, .., n_\beta$ tels que : $p_{\alpha_0} \cap p_{\beta_0} \cap S \neq \emptyset$, i.e. $p_{\alpha_0} \mathcal{R} p_{\beta_0}$.
$p_{\alpha_0} \in \mathcal{P}_\alpha$, $p_{\alpha_0} \in \mathcal{P}_\beta$, $\mathcal{G}_\alpha$ et $\mathcal{G}_\beta$ sont deux composantes connexes de $\mathcal{G}_S$, donc $p_{\alpha_0} \not\mathcal{R} p_{\beta_0}$.

□

Remarque 3.3.11
Dans [51], Tarjan propose un algorithme qui compte le nombre de composantes connexes, en $O(n)$ unités de temps, où n est le nombre de sommets d'un graphe simple non-orienté.

Dans la section suivante, nous présentons un algorithme qui essaie de créer un graphe parsemé d'étoiles.

3.3.4 Algorithme - Nombre de composantes connexes par arcs.

Cette section présente un nouvel algorithme appelé : `CIA` (path-**C**onnected using **I**nterval **A**nalysis). Cet algorithme essaie de générer un graphe parsemé d'étoiles $\mathcal{G}_S$ (Corollaire 3.3.10). L'idée principale est de tester un pavage candidat $\mathcal{P}$: dans le cas où il ne vérifie pas la condition : $\forall p \in \mathcal{P},\ p \cap S$ est étoilé, l'objectif est de l'améliorer en bissectant tous les pavés responsables de cet échec.

Pour un pavage $\mathcal{P}$, on doit tester pour chaque pavé p de $\mathcal{P}$ si $S \cap p$ est étoilé ou non, et construire le graphe associé via la relation $\mathcal{R}$. Ces deux taches sont effectuées respectivement par les algorithmes Alg. 5 et Alg. 6.

Dans l'algorithme baptisé `CIA` (Alg. 4), $\mathcal{P}_*$, $\mathcal{P}_{out}$, $\mathcal{P}_\Delta$ sont trois pavages tels que $\mathcal{P}_* \cup \mathcal{P}_{out} \cup \mathcal{P}_\Delta = \mathcal{P}$, avec $\mathcal{P}$ un pavage dont le support est un pavé initial X_0 (qui contient S) :

- le pavage $\mathcal{P}_*$ contient des pavés p tel que $S \cap p$ est étoilé.
- le pavage $\mathcal{P}_{out}$ contient des pavés p tels que $S \cap p$ est vide.
- pour le pavage $\mathcal{P}_\Delta$, rien n'est connu concernant ses éléments.

Alg. 4 `CIA` - path connected components using Interval Analysis

Entrée: S un sous-ensemble de $\mathbb{R}^n$, X_0 un pavé de $\mathbb{R}^n$.

1: Initialisation : $\mathcal{P}_* := \emptyset$, $\mathcal{P}_\Delta := \{X_0\}$, $\mathcal{P}_{out} := \emptyset$
2: **tant que** $\mathcal{P}_\Delta \neq \emptyset$ **faire**
3: Dépiler le dernier élément de $\mathcal{P}_\Delta$ dans le pavé p
4: **si** "$S \cap p$ est montré vide" **alors**
5: Empiler $\{p\}$ dans $\mathcal{P}_{out}$, **Aller à l'étape 2.**
6: **fin si**
7: **si** "$S \cap p$ est montré étoilé" et `Build_Graph_Interval`$(S, \mathcal{P}_* \cup \{p\})$ est ponctuel **alors**
8: Empiler $\{p\}$ dans $\mathcal{P}_*$, Va à l'étape 2.
9: **fin si**
10: $Bisect(p)$ et Empiler le deux pavés résultants dans $\mathcal{P}_\Delta$
11: **fin tant que**
12: $n \leftarrow$ Nombre de composantes connexes de $\underline{g}$
13: Afficher "S a n composantes connexes par arcs."

Pour montrer que "$S \cap p$ est étoilé", il suffit de vérifier qu'un des sommets v_p de p est une étoile pour $S \cap p$. L' algorithme suivant appelé `Star-shaped` montre comment cette vérification peut être implémentée.

Alg. 5 `Star-shaped`(p, f)

Entrée: f une fonction C^1 de $\mathbb{R}^n$ vers $\mathbb{R}$

Entrée: p un pavé de $\mathbb{R}^n$

1: **si** $f(p)$ peut être montré élément de $\mathbb{R}^{+*}$ **alors**
2: Affiche "$S \cap p$ est vide"
3: **sinon**
4: **pour tout** sommet v_p de p **faire**
5: **si** $\{x \in p, f(x) = 0, Df(x) \cdot (x - v_p) \leq 0\}$ est montré inconsistant **alors**
6: Retourner "$S \cap p$ est étoilé"
7: **fin si**
8: **fin pour**
9: Retourner "Echec"
10: **fin si**

Remarque 3.3.12

L'algorithme 5 teste si un ensemble $S = f^{-1}(\mathbb{R}^-)$ est étoilé. Il est possible d'étendre cet algorithme à des ensembles décrits par

$$S = \bigcup_{i=1}^{s} \bigcap_{j=1}^{r_i} \{x \in \mathbb{R}^n; f_{i,j}(x) \leq 0\} \text{ où } f_{i,j} \in \mathcal{C}^1(\mathbb{R}^n, \mathbb{R}) \tag{3.36}$$

En effet, comme la propriété *être v^*-étoilé* est une propriété stable par union et intersection (Proposition 3.1.5), il suffit de vérifier que chaque $\{x \in \mathbb{R}^n; f_{i,j}(x) \leq 0\}$ est v^*-étoilé.

Pour construire le graphe associé à un pavage $\mathcal{P}$, on doit tester pour chaque paire $\{p_i, p_j\}$ du pavage $\mathcal{P}$ si $S \cap p_i \cap p_j$ est vide ou non. Lorsque l'on ne peut décider si $S \cap p_i \cap p_j$ est vide ou non, on crée un graphe intervalle qui contient le graphe recherché. L'algorithme suivant nommé `Build_Graph_Interval` propose une construction de ce graphe.

Remarque 3.3.13

Lorsque l'ensemble S est défini par des inégalités, la condition à l'étape 4 est vérifiée en utilisant le calcul par intervalles. La proposition 2.4.1 montre que cet outil permet aussi de prouver que $S \cap p_i \cap p_j = \emptyset$ (Etape 3).

Alg. 6 `Build_Graph_Interval`$(S, \mathcal{P})$

Entrée: S un sous-ensemble de $\mathbb{R}^n$, $\mathcal{P}$ un pavage.

Sortie: Un graphe intervalle $[\underline{g}, \overline{g}]$ associé au pavage $\mathcal{P}$.

1: Initialisation : $\overline{g} := \emptyset$, $\underline{g} := \emptyset$
2: **pour tout** (p_i,p_j) in $\mathcal{P} \times \mathcal{P}$ **faire**
3: **if** $S \cap p_i \cap p_j = \emptyset$ **then next**
4: **si** pour un des sommets v de $p_i \cap p_j$, $v \in S$ **alors**
5: ajouter (p_i, p_j) à $\underline{g}$ et à $\overline{g}$
6: **sinon**
7: ajouter (p_i, p_j) à $\overline{g}$ // *i.e.*(p_i,p_j) *est arc indéterminé de* $[\underline{g}, \overline{g}]$
8: **fin si**
9: **fin pour**

Exemple 3.3.14

La figure 3.30 montre le pavage généré pour

$$S = \left\{ (x,y) \in \mathbb{R}^2, \left\{ \begin{array}{lcl} f_1(x,y) = x^2 + 4y^2 - 16 & \leq & 0 \\ f_2(x,y) = 2\sin(x) - \cos(y) + y^2 - 1.5 & \leq & 0 \\ f_3(x,y) = -(x+2.5)^2 - 4(y-0.4)^2 + 0.3 & \leq & 0 \end{array} \right. \right\} \quad (3.37)$$

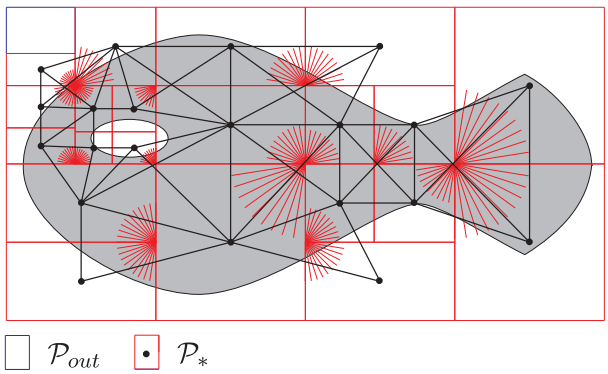


FIG. 3.30 – Exemple de graphe parsemé d'étoiles généré par `CIA`.

Exemple 3.3.15

La figure 3.31 montre le pavage généré pour $S = \bigcup_{i=1}^{i=4} S_i$ où

$$\begin{array}{lll} D = [-5,5] \times [-4.6, 4.6] & & \\ S_1 = \{(x,y) \in D, \ \ f_1(x,y) = -x^2 - y^2 + 9 & \leq & 0\} \\ S_2 = \{(x,y) \in D, \ \ f_2(x,y) = (x-1)^2 + (y-1.5)^2 - 0.5 & \leq & 0\} \\ S_3 = \{(x,y) \in D, \ \ f_3(x,y) = (x+1)^2 + (y-1.5)^2 - 0.5 & \leq & 0\} \\ S_4 = \{(x,y) \in D, \ \ f_4(x,y) = \cos^2(x+1.5) + 4(y+2)^2 - 0.5 & \leq & 0\} \end{array} \tag{3.38}$$

Sur cette figure, l'emplacement des étoiles est illustré.

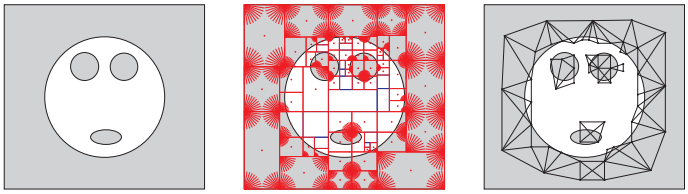


FIG. 3.31 – Graphe parsemé d'étoiles généré par `CIA`. S et $\mathcal{G}_S$ ont 4 composantes connexes.

Lorsque le solveur prouve qu'un sommet du pavé p est une étoile pour $S \cap p$, il utilise la même représentation que celle utilisée sur la figure 3.3.

3.3.5 Limites de cette méthode

Bien évidemment, l'algorithme 4 ne termine pas toujours. L'annexe A montre qu'il est possible de créer des ensembles particulièrement *tordus* avec les fonctions C^∞.

Sans aller jusqu'à ces extrêmes, si S est d'intérieur vide ($f_1 \leq 0 \wedge -f_1 \leq 0$) ou s'il existe un x qui vérifie $f(x) = 0$ et $Df(x) = 0$, il ne termine pas. Il semble naturel qu'une condition nécessaire pour qu'il termine est $\overline{\overset{\circ}{S}} = S$. La section suivante présente une amélioration de cet algorithme capable de renseigner plus de propriétés topologiques. Les limites de ces deux algorithmes sont quasiment les mêmes et seront discutées dans la section 3.4.4. Cet algorithme a été implémenté et une version exécutable peut être téléchargée via ma page personnelle.

3.4 Type d'homotopie - Algorithme H.I.A.

3.4.1 Introduction

Dans cette section, on présente une méthode effective (souvent) capable de construire une triangulation du même type d'homotopie qu'un ensemble S décrit par

$$S = \bigcup_{i=1}^{s} \bigcap_{j=1}^{r_i} \{x \in \mathbb{R}^n; f_{i,j}(x) \leq 0\} \text{ où } f_{i,j} \in \mathcal{C}^1(\mathbb{R}^n, \mathbb{R}) \tag{3.39}$$

Cette méthode utilise de façon un peu plus fine la condition suffisante des ensembles étoilés. Globalement, la technique se rapproche de l'homologie de Čech au sens où l'on affine le recouvrement de manière à créer un *bon recouvrement*[11] [10]. Tous ces travaux ont été développés en ignorant le théorème de nervure (traduction de *nerve theorem*). [9]

3.4.2 Discrétisation

Pour créer un complexe simplicial qui est du même type d'homotopie qu'un ensemble S, l'idée principale de notre approche est de générer un recouvrement $\{S_i\}_{i \in I}$ de S, puis dans une seconde étape de créer un complexe simplicial abstrait.

Chaque élément du recouvrement doit être *contractile* et de plus, l'intersection de n'importe quelle collection $\{S_i\}_{i \in J \subset I}$ doit être contractile (ou vide). Pour simplifier notre présentation, on introduit la notation 3.4.1.

Notation 3.4.1
Soit I une famille finie et J une sous famille de I ; on note par S_J l'ensemble $\bigcap_{j \in J} S_j$. Par exemple, l'ensemble $S_3 \cap S_4 \cap S_9$ est noté $S_{\{3,4,9\}}$.

Définition 3.4.2
Soit S un espace topologique de $\mathbb{R}^n$, $\{S_i\}_{i \in I}$ est un *recouvrement contractile* de S si :

- I est fini.
- $\forall J \subset I, S_J$ est contractile ou vide.

La figure 3.32 montre un exemple de recouvrement contractile.

[11] Dans le cas de la cohomologie de Čech, un recouvrement d'ouverts d'une variété différentiable M est un *bon recouvrement* si tous les ouverts et toutes les intersections finies d'ouverts sont contractiles.

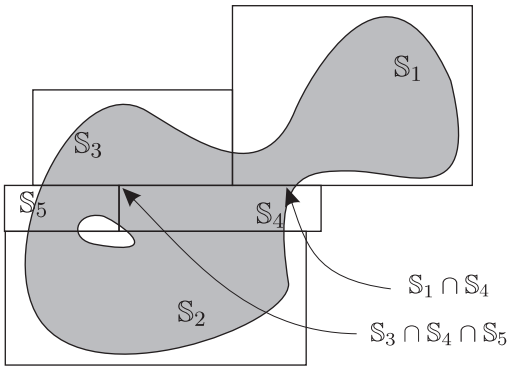


FIG. 3.32 – Un recouvrement contractile $\{S_i\}_{i\in I}$ of S.

Définition 3.4.3

Soit $\{S_i\}_{i\in I}$ un recouvrement contractile d'un ensemble S. On note par $\mathcal{J}$ l'ensemble des S_J non vides. Un complexe simplicial abstrait $\mathcal{K}(S)$ est dit *adapté* à $\{S_i\}_{i\in I}$ s'il est le plus petit complexe simplicial vérifiant :

- $\forall J \in \mathcal{J}$, un sommet abstrait (a^J) est dans $\mathcal{K}(S)$.
- $\forall J \in \mathcal{J}$, un complexe simplicial $\mathcal{K}_J$ défini par

$$\mathcal{K}_J = a^J * \left(\sum_{J' \in \mathcal{J} | S_{J'} \subset S_J} \mathcal{K}_{J'} \right). \tag{3.40}$$

  est un sous-complexe de $\mathcal{K}(S)$. Dans la littérature, ce complexe simplicial est classiquement appelé la *nervure* du recouvrement $\{S_i\}_{i\in I}$.

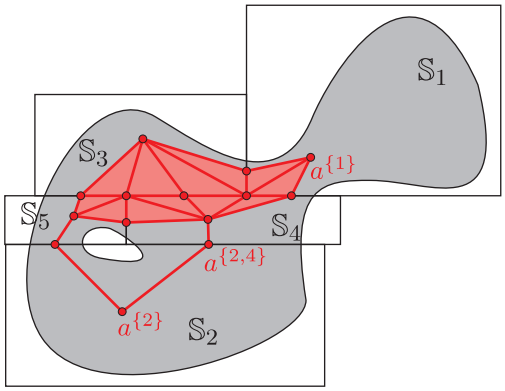


FIG. 3.33 – Un complexe simplicial abstrait adapté à $\{S_i\}_{i\in I}$.

Théorème 3.4.4 (Nerve theorem) *Si $\{S_i\}_{i\in I}$ est un recouvrement de $S \subset \mathbb{R}^n$*

et $\mathcal{K}(S)$ un complexe simplicial abstrait adapté à $\{S_i\}_{i\in I}$, alors $\mathcal{K}(S)$ et S sont du même type d'homotopie.

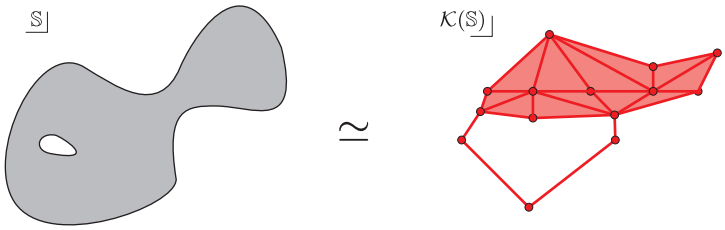


FIG. 3.34 – S et $\mathcal{K}(S)$ sont du même type d'homotopie.

Preuve : voir [9].

□

3.4.3 Un algorithme

Ici, on présente un algorithme, appelé `HIA` (**H**omotopy via **I**nterval **A**nalysis). Cet algorithme est souvent capable de créer un recouvrement contractile d'un ensemble défini par des inégalités. Dans un second temps, il crée un complexe simplicial adapté à ce recouvrement. Ce recouvrement $\{S_i\}_{i\in I}$ est créé grâce à un pavage. Avant d'énoncer l'algorithme, quelques définitions sont introduites.

Notation 3.4.5
Soit S une partie de $\mathbb{R}^n$ et $\{p_i\}_{i\in I}$ un pavage tel que $S \subset \bigcup_{i\in I} p_i$. On dénote par $\{S_i\}_{i\in I}$ le recouvrement de S où $S_i = S \cap p_i, i \in I$. La bissection de S_i est donc définie via la bissection du pavé p_i.

L'idée principale de cet algorithme est de subdiviser. A partir d'un recouvrement $\{S_i\}_{i\in I}$ créé grâce à un pavage $\{p_i\}_{i\in I}$ (voir notation 3.4.5), un sous-algorithme vérifie si :

$$\forall J \subset I, \bigcap_{j\in J} S_j \text{ est contractile ou vide. } (S_j := S \cap p_j) \tag{3.41}$$

Si cette condition n'est pas satisfaite, alors chaque S_i responsable de cet échec est bissecté. Cet algorithme utilise :

- le pavage $\mathcal{P}_*$, il est tel que : $\forall \{p_j\}_{j\in J} \subset \mathcal{P}_*$, $\bigcap_{j\in J} S_j$ est contractile ou vide.
- le pavage $\mathcal{P}_\Delta$, rien n'est connu concernant ses éléments.

Dans un second temps, grâce à $\mathcal{P}_* = \{p_i\}_{i\in I}$, le sous-algorithme `Nerve` crée un complexe simplicial abstrait $\mathcal{K}(S)$ adapté au recouvrement $\{S_i := S \cap p_i\}_{i\in I}$.

Alg. 7 `HIA` - Homotopy type via Interval Analysis

Entrée: S un sous-ensemble de $\mathbb{R}^n$, X_0 un pavé de $\mathbb{R}^n$ qui contient S.

Sortie: Une triangulation $\mathcal{K}(S)$ du même type d'homotopie que S.

1: Initialisation : $\mathcal{P}_* := \emptyset$, $\mathcal{P}_\Delta := \{X_0\}$
2: **tant que** $\mathcal{P}_\Delta \neq \emptyset$ **faire**
3: Dépiler le dernier élément de $\mathcal{P}_\Delta$ dans le pavé p
4: **si**

$$\forall \{p_j\}_{j\in J} \subset \mathcal{P}_* \cup \{p\}, \quad \bigcap_{j\in J} S_j \text{ est prouvé contractile ou vide} \tag{3.42}$$

alors
5: $\mathcal{P}_* \leftarrow \mathcal{P}_* \cup \{p\}$;
6: **sinon**
7: $Bisect(p)$ en deux intervalles p_1 et p_2 ;
8: $P_\Delta \leftarrow \mathcal{P}_* \cup \{p_1\} \cup \{p_2\}$;
9: **fin si**
10: **fin tant que**
11: $\mathcal{K}(S) \leftarrow$ `Nerve` à partir du recouvrement : $\{S_i\}_{i\in I}$,
(où $S_i := S \cap p_i, p_i \in \mathcal{P}_*$).

Remarque 3.4.6

La condition de l'étape 4 est vérifiée en utilisant la proposition 3.3.1.

`Nerve` est un algorithme (Alg. 8) qui produit une triangulation adaptée $\mathcal{K}(S)$ au recouvrement $\{S_i\}_{i\in I}$ de S. L'idée est d'ajouter un cône à $\mathcal{K}(S)$ pour chaque J dans $\mathcal{J}$ tel que deux cônes sont "attachés" via un cône créé précédemment.

Alg. 8 `Nerve`

Entrée: Un ensemble S et un recouvrement $\{S_i\}_{i\in I}$ de S.

Sortie: Une triangulation $\mathcal{K}(S)$ adaptée à $\{S_i\}_{i\in I}$.

{*Initialisation :*}

1: $\mathcal{K}(\mathcal{S}) \leftarrow \emptyset$, $\mathcal{J} \leftarrow \emptyset$.

{*Retire les indices inutiles, i.e. Créer $\mathcal{J}$:*}

2: **pour tout** $J \subset I$ **faire : si** S_J est contractile **then** $\mathcal{J} \leftarrow \mathcal{J} \cup \{J\}$ **fin pour**

{*Créer la triangulation :*}

3: $\mathcal{K}(S) \leftarrow \sum_{i\in I}\texttt{Cone}(\{i\})$

Où $\texttt{Cone}(J), J \in \mathcal{J}$ est défini récursivement par :

$$\texttt{Cone}(J) = a^J * \left(\sum_{J'\in\mathcal{J}|S_{J'}\subset S_J} \texttt{Cone}(J') \right) \tag{3.43}$$

avec comme convention que $\texttt{Cone}(\emptyset) = \emptyset$ et $a^J * \emptyset = a^J$.

Illustration : Pour illustrer l'alg. 8, considérons le recouvrement donné par la figure 3.32. Dans ce cas, l'ensemble : $\{J|J \subset I\}$ a 2^5 éléments car $\#I = 5$. Mais seulement un certain nombre d'entre eux sont tels que $\bigcap_{j\in J} S_j$ est contractile. Après l'étape 2, on a :

$$\begin{aligned}\mathcal{J} \quad = \quad & \{\{1\},\{1,4\},\{1,3\},\{1,3,4\},\{2\},\{2,4\},\{2,5\},\{3\},\{3,4\},\{3,5\},\\ & \{3,4,5\},\{4\},\{4,5\},\{5\}\}\end{aligned}$$

Pour expliquer ce qui est effectué à l'étape 3, par exemple, les éléments nécessaires au calcul de $\texttt{Cone}(\{1\})$. Pour l'effectuer, la collection des $J' \in \mathcal{J}$ tels que $S_{J'} \subset S_{\{1\}}$ doit être connue. Dans ce cas, cette collection est composée de $S_{\{1\}}, S_{\{1,4\}}, S_{\{1,3\}}$ et $S_{\{1,3,4\}}$. Plus généralement, les éléments de $\{S_J, J \in \mathcal{J}\}$ peuvent être partiellement ordonnés par l'inclusion. La figure 3.35 montre comment ces éléments sont ordonnés où $S_{J'} \subset S_J$ est représenté par $J' \leftarrow J$.

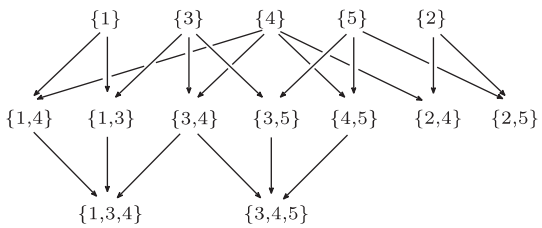


FIG. 3.35 – Relation d'ordre sur $\{S_J, J \in \mathcal{J}\}$

Pour alléger notre explication, on renomme $a^{\{1\}}$ par a, $a^{\{2\}}$ par b, $a^{\{3\}}$ par c. La figure 3.36 montre comment les a^J sont renommés.

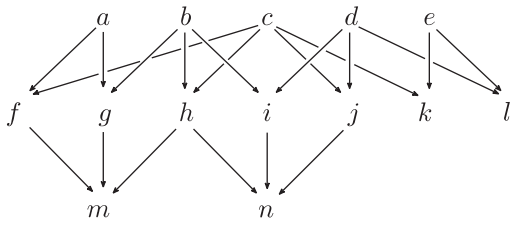


FIG. 3.36 – Relation d'ordre sur $\{S_J, J \in \mathcal{J}\}$

A l'étape 3, on obtient ce complexe simplicial abstrait :

$$\begin{aligned}
\mathcal{K}(S) &= a*(f*m+g*m)+b*(g*m+h*(m+n)+i*n)+ \\
&\quad c*(f*m+h*(m+n)+j*n+k)+ \\
&\quad d*(i*n+j*n+l)+e*(k+l).
\end{aligned}$$

$$\begin{aligned}
\mathcal{K}(S) &= afm+agm+bgm+bhm+bhm+bin+ \\
&\quad cfm+chm+chn+cjn+ck+din+djn+dl+ek+el.
\end{aligned}$$

Une illustration géométrique est donnée par la figure 3.37. On peut y voir une réalisation du complexe simplicial abstrait $\mathcal{K}$.

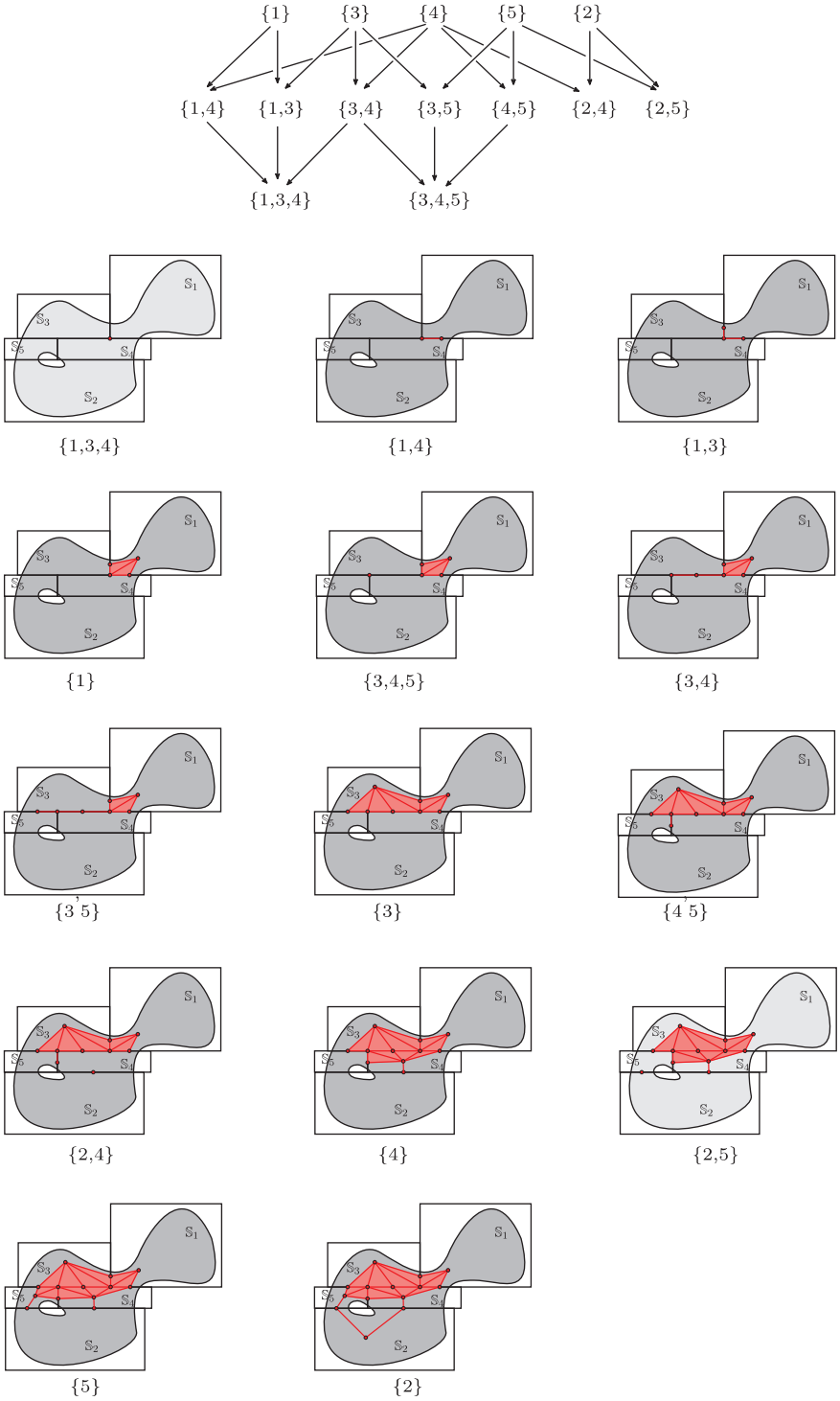


FIG. 3.37 – `Nerve` algorithme pas à pas.

Exemple 3.4.7

Les figures 3.38 et Figure 3.39 sont respectivement les exemples des réalisations de complexes simpliciaux générés par `HIA` pour les ensembles :

$$S_1 = \left\{(x,y) \in \mathbb{R}^2 \mid x^2 + y^2 + xy - 2 \leq 0 \text{ et } -x^2 - y^2 - xy + 1 \leq 0\right\} \tag{3.44}$$

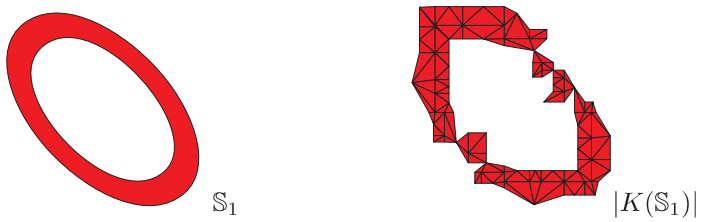


FIG. 3.38 – Complexe simplicial généré par `HIA`.

et

$$S_2 = \left\{(x,y) \in \mathbb{R}^2 \mid x^2 + y^2 - 6 \leq 0 \text{ et } 0.2\cos(x-y) - \sin(yx) - 0.6 \leq 0\right\} \tag{3.45}$$

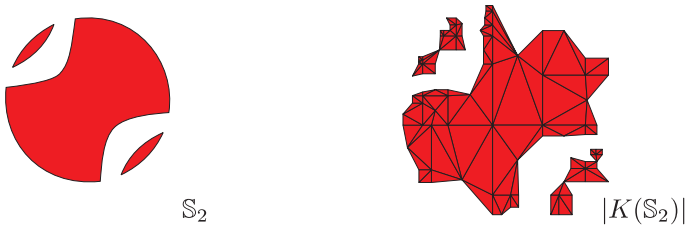


FIG. 3.39 – Complexe simplicial généré par `HIA`.

3.4.4 Limites de cette méthode

Cet algorithme a été implémenté et une version exécutable peut être téléchargée via ma page personnelle. Comme indiqué dans la section 3.3.5 page 80, il n'est pas difficile de créer un exemple pour lequel l'algorithme 7 ne termine pas. L'entrée de l'algorithme

$$\begin{aligned} S &= \left\{(x,y) \in \mathbb{R}^2 \mid (x^2 + y^2 \leq 1) \wedge (-x^2 - y^2 \leq -1)\right\} && (3.46) \\ &= \left\{(x,y) \in \mathbb{R}^2 \mid x^2 + y^2 = 1\right\} && (3.47) \\ & && (3.48) \end{aligned}$$

fournit un exemple. Ici S est le cercle de centre 0 et de rayon 1. Pour cet ensemble, il est impossible de trouver un nombre fini d'étoiles qui éclairent tous S. Dans ce cas, l'algorithme 7 tourne indéfiniment. Une piste éventuelle de recherche consisterait à "gonfler" S sans en changer le type d'homotopie en considérant par exemple l'ensemble

$$S_\epsilon = \left\{(x,y) \in \mathbb{R}^2 \mid (x^2+y^2-1)^2 \leq \epsilon\right\}. \tag{3.49}$$

La principale difficulté dans cette situation serait de caler ϵ tel manière que $S \simeq S_\epsilon$. On pourrait éventuellement faire une analyse sur les points critiques ...Mais pour ce genre d'ensemble, *i.e.* décrit par une égalité, il semble plus judicieux de combiner la calcul par intervalles à la théorie de Morse [3] comme l'on fait Stander et Hart [2].

Ce n'est pas la seule situation où l'algorithme 7 échoue. Si on considère l'ensemble

$$\begin{aligned} S &= \left\{(x,y) \in \mathbb{R}^2 \mid ((x-1)^2+y^2 \leq 1) \vee ((x+1)^2+y^2 \leq 1)\right\} && (3.50)\\ &= \left\{(x,y) \in \mathbb{R}^2 \mid x^2+y^2 = 1\right\} && (3.51)\\ & && (3.52) \end{aligned}$$

qui correspond géométriquement à la réunion de deux disques dont l'intersection est la point O de coordonnées $(0,0)$. Dans cette situation, il faut nécessairement qu'une étoile soit placée au point O. Et la condition 3.25 ne peut être satisfaite par un algorithme de bissection puis que O se trouve sur la frontière de l'ensemble :

$$\left\{(x,y) \in \mathbb{R}^2 \mid ((x-1)^2+y^2 \leq 1)\right\}$$

Cette situation est exceptionnelle et on conjecture

Conjecture 3.4.8 *L'ensemble des m fonctions qui engendrent une partie de $\mathbb{R}^n$ pour lesquelles l'algorithme 7 termine forme une partie résiduelle des fonctions $\mathcal{C}^\infty$.*

3.5 Applications

Un des problèmes de la robotique est celui de la *planification de trajectoire*, c'est à dire de décider comment un robot peut aller d'un endroit à un autre sans toucher les murs. Supposons que tous les obstacles soient fixes et sont donnés par des équations ou inéquations. Si on considère un robot rigide, la donnée de trois

points qui ne sont pas alignés permet de le fixer sans ambiguïté. Les obstacles définissent plusieurs inégalités qui doivent être satisfaites pour que le robots soit dans une configuration admissible. En principe, le problème de planification est assez simple : est ce que la position de départ et la position d'arrivée sont connexes dans l'ensemble des configurations admissibles. Autrement dit, existe-il un chemin qui relie la position de départ à la position désirée ? Dans le cas où les inégalités sont polynomiales, ces idées ont été étudiés par Schwartz et Sharir [26] qui ont expliqué comment trouver un chemin. Le livre de Steven M. LaValle [11] propose un état de l'art des méthodes de planification de trajectoires.

3.5.1 Exemple de robotique

Considérons le robot à deux degrés de liberté donné par la figure 3.40. Deux obstacles sont représentés en gris sur la figure 3.40 ; la distance entre les deux murs est y_0. Un robot avec deux degrés de liberté α et β est placé dans cet environnement. (Voir Figure 3.40).

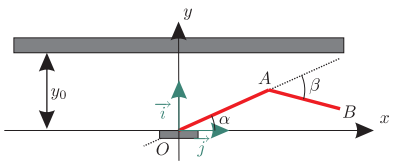


FIG. 3.40 – Un robot avec deux bras, avec comme distance $OA = 2$ et $AB = 1.5$.

Les coordonnées de A et B sont données par les expressions suivantes :

$$\begin{cases} x_A = 2\cos(\alpha) \\ y_A = 2\sin(\alpha) \end{cases} \qquad \begin{cases} x_B = 2\cos(\alpha) + 1.5\cos(\alpha+\beta) \\ y_B = 2\sin(\alpha) + 1.5\sin(\alpha+\beta) \end{cases} \tag{3.53}$$

3.5.2 Espace des configurations

Chaque coordonnée d'un point de l'*espace des configurations* représente un degré de liberté (Voir Figure 3.41). Le nombre de paramètres nécessaires pour caractériser la configuration d'un objet correspond à la dimension de l'espace des configurations. Cet ensemble de configuration n'est pas nécessairement $\mathbb{R}^n$ avec n le nombre de degré de liberté. Dans notre exemple, seulement α et β sont nécessaires pour déterminer la configuration du robot, l'espace des configurations est donc un espace de dimension 2.

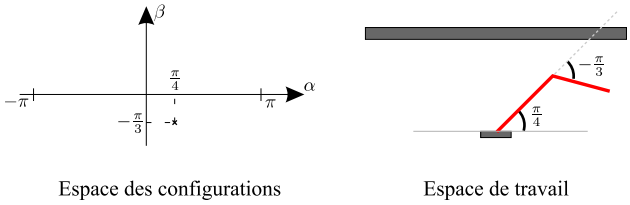


FIG. 3.41 – Un point dans l'espace des configurations et la configuration du robot qui lui correspond.

Comme le robot ne peut traverser les murs, nous avons les contraintes $y_A \in [0, y_0]$ et $y_B \in]-\infty, y_0]$, de plus nous ajoutons les contraintes (dites articulaires par les roboticiens) $\alpha \in [-\pi, \pi]$ and $\beta \in [-\pi, \pi]$. Quand ces contraintes sont satisfaites, le robot est dit dans une *configuration admissible*. L'ensemble des configurations admissibles S est, par conséquent, défini par :

$$S = \left\{ (\alpha, \beta) \in [-\pi, \pi]^2 / \left\{ \begin{array}{lcl} 2\sin(\alpha) & \in & [0, y_0] \\ 2\sin(\alpha) + 1.5\sin(\alpha + \beta) & \in &]-\infty, y_0] \end{array} \right. \right\} \tag{3.54}$$

Les figures 3.42, 3.43 et 3.44 montrent dans trois cas l'influence de y_0 sur l'espace des configurations admissibles :

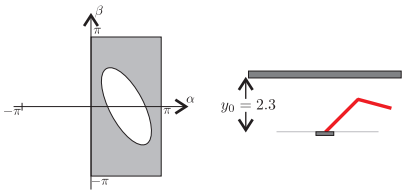


FIG. 3.42 – Espace des configurations admissibles lorsque $y_0 = 2.3$. Le robot peut se déplacer de n'importe quel état *initial* admissible vers n'importe quel état *final* admissible. Dans ce cas, S a une composante connexe.

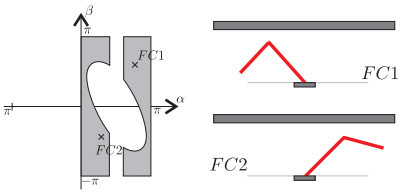


FIG. 3.43 – Espace des configurations admissibles lorsque $y_0 = 1.9$. L'espace des configurations a deux composantes connexes . Il lui est impossible de se déplacer de la configuration $FC1$ à la configuration $FC2$ sans violer les contraintes.

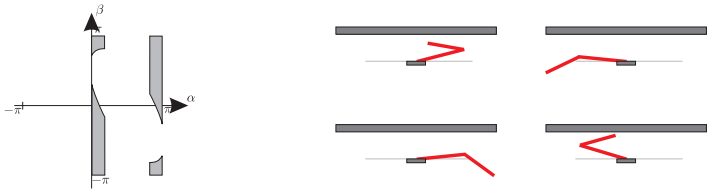


FIG. 3.44 – Espace des configurations admissibles lorsque $y_0 = 1.1$. Le robot peut être piégé dans 4 régions différentes. L'ensemble S a 4 composantes connexes par arcs.

3.5.3 Topologie de l'espace des configurations

Considérons l'exemple présenté dans la section précédente. L'espace des configurations admissibles S est :

$$S = \left\{ (\alpha, \beta) \in [-\pi, \pi]^2 / \left\{ \begin{array}{lcl} -2\sin(\alpha) & \leq & 0 \\ 2\sin(\alpha) - y_0 & \leq & 0 \\ 2\sin(\alpha) + 1.5\sin(\alpha+\beta) - y_0 & \leq & 0 \end{array} \right. \right\} \tag{3.55}$$

Lorsque y_0 vaut 2.3, 1.9 et 1.1, l'algorithme `CIA` génère les graphes parsemés d'étoiles présentés respectivement sur les Figures 3.45,3.46 et 3.47.

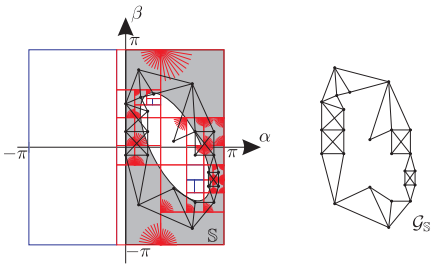


FIG. 3.45 – L'espace des configurations admissibles et un graphe parsemé d'étoiles généré par `CIA` quand $y_0 = 2.3$. Le graphe parsemé d'étoiles $\mathcal{G}_S$ est connexe . En utilisant la proposition 3.3.10, on en déduit que pour chaque couple de points, il est possible de créer un chemin qui les relie. (La section 3.5.4 montre comment un tel chemin peut être trouvé.)

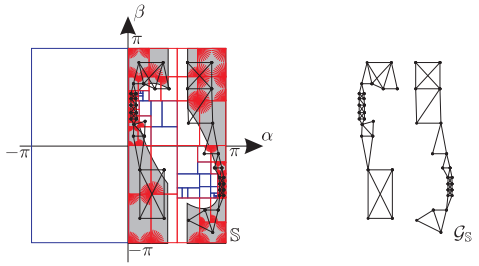


FIG. 3.46 – L'espace des configurations admissibles et un graphe parsemés d'étoiles généré par `CIA` quand $y_0 = 1.9$. Comme $\mathcal{G}_S$ a deux composantes connexes, l'ensemble S a aussi deux composantes connexes par arcs.

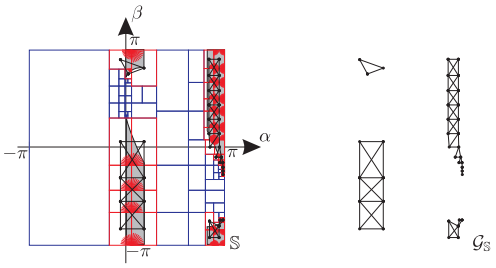


FIG. 3.47 – L'espace des configurations admissibles et un graphe parsemés d'étoiles généré par `CIA` quand $y_0 = 1.1$. $\mathcal{G}_S$ et S ont 4 composantes connexes.

3.5.4 Planification de trajectoires

Comme nous allons le montrer, un graphe parsemé d'étoiles est un objet suffisamment riche pour créer un chemin entre deux points. Notre objectif est de trouver un chemin d'une configuration initiale x vers une configuration finale y (voir figure 3.48).

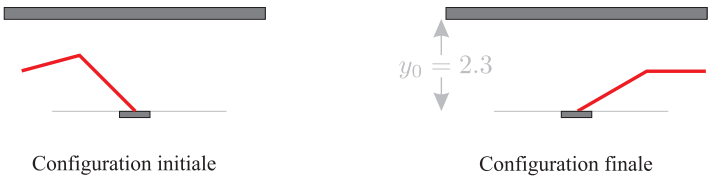


FIG. 3.48 – Configuration initiale $x = (\frac{3\pi}{4}, \frac{\pi}{3})$ et configuration finale, $y = (\frac{\pi}{6}, -\frac{\pi}{6})$

Il suffit de trouver, un chemin qui relie x à y dans l'espace des configurations admissibles. L'algorithme `Path-planning with CIA`, grâce à un graphe parsemé

d'étoiles, construit un chemin γ vérifiant $\gamma([0,1]) \subset S$. Cet algorithme utilise l'algorithme de `Dijkstra` [19] qui trouve le chemin le plus court entre deux points dans un graphe. Comme $\mathcal{G}_S$ est un graphe parsemé d'étoiles, chaque p de $\mathcal{P}$ est nécessairement étoilé et on note par v_p une de ces étoiles.

Alg. 9 `Path-planning with CIA`

Entrée: Un ensemble S, $x, y \in S$, $\mathcal{G}_S$ un graphe parsemé d'étoiles de S (*i.e.* une relation $\mathcal{R}$ sur le pavage $\mathcal{P}$).

Sortie: $\gamma \subset S$ un chemin qui a pour extrémités x et y.

1: Initialisation : $\lambda \leftarrow \emptyset$
2: **pour tout** $p \in \mathcal{P}$ **faire**
3: **si** $x \in p$ **alors** $p_x \leftarrow p$;
4: **si** $y \in p$ **alors** $p_y \leftarrow p$
5: **fin pour**
6: **si** `Dijkstra`$(\mathcal{G}_S, p_x, p_y) =$ "Erreur" **alors**
7: Retourne "x et y sont dans 2 composantes connexes par arcs différentes"
8: **sinon**
9: $(p_k)_{1 \leq k \leq n} = (p_x, \ldots, p_y) \leftarrow$ `Dijkstra`$(\mathcal{G}_S, p_x, p_y)$
10: **fin si**
11: $\gamma \leftarrow [x, v_{p_x}]$
12: **pour** $k \leftarrow 2$ à $n-1$ **faire**
13: $w_{k-1,k} \leftarrow$ un point de $p_{k-1} \cap p_k \cap S$;
14: $w_{k,k+1} \leftarrow$ un point de $p_k \cap p_{k+1} \cap S$
15: $\gamma \leftarrow \gamma \cup [w_{(k-1,k)}, v_{p_k}] \cup [v_{p_k}, w_{(k,k+1)}]$
16: **fin pour**
17: $\gamma \leftarrow \gamma \cup [v_{p_y}, y]$

La figure 3.49 montre le chemin γ produit par `Path-planning with CIA`.

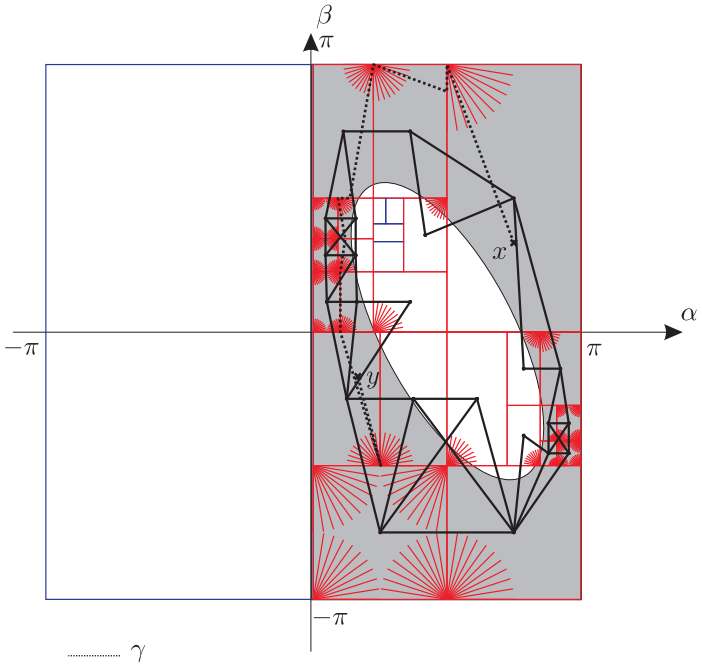


FIG. 3.49 – Chemin γ généré par `Path-planning with CIA` de x à y lorsque $y_0 = 2.3$.

Les différentes étapes des configurations correspondant au chemin γ sont illustrées via la figure 3.50.

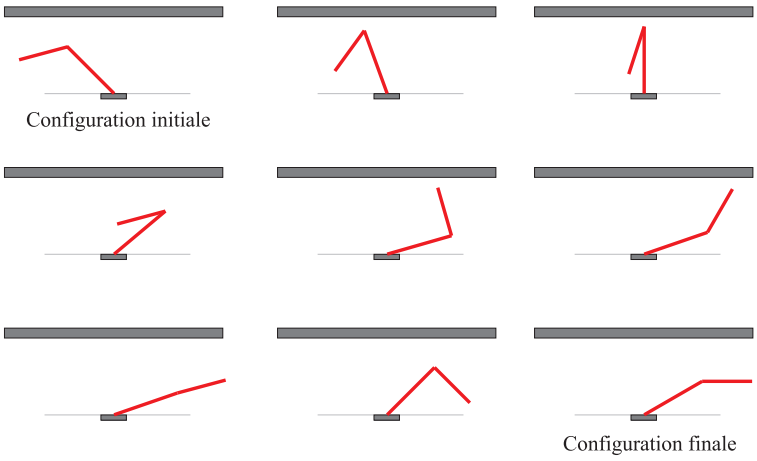


FIG. 3.50 – Mouvement du robot correspondant au chemin γ.

3.5.5 Applications de HIA

Exemple 3.5.1

Cet exemple est encore issu de la robotique. L'algorithme présenté dans la section 3.4 est utilisé pour calculer une triangulation du même type d'homotopie que l'ensemble des configurations admissibles. Les intérêts de cet exemple sont multiples. Contrairement à l'exemple présenté dans la section 3.5.1, il nous semble impossible de réécrire l'ensemble des configurations admissibles avec un ensemble semi-algébrique. De plus, l'ensemble des configurations admissibles est décrit par des inégalités dont certains paramètres sont solutions d'autres équations transcendantes. A notre connaissance, aucun autre algorithme, actuellement implémenté, n'est capable de traiter ce problème.

On considère une corde de longueur L suspendue entre deux points A et B, qui sont les extrémités de deux bras (Figure 3.51).

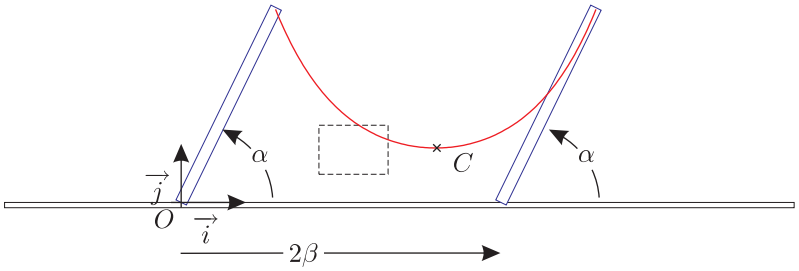


FIG. 3.51 – Une corde suspendue entre deux points points A and B.

Les paramètres indépendants nécessaires et suffisants pour caractériser une configuration sont les nombres réels α et β. On impose que le robot vérifie quelques contraintes. La corde ne doit pas toucher le sol, et le point le plus bas de la corde, appelé C ne doit pas se trouver dans le domaine délimité par la ligne pointillée (Figure 3.51). On veut donc étudier les propriétés topologiques de l'ensemble des configurations admissibles S :

$$S = \left\{(\alpha, \beta) \in [0, \pi] \times [\beta^-, \beta^+] \text{ tel que } \; y_C \geq 0 \text{ et } C \notin \boxed{}\right\}. \tag{3.56}$$

Pour appliquer notre algorithme, premièrement, nous devons pouvoir décrire l'ensemble S par une formule sans quantificateur dont les atomes sont $f \leq 0$, $f \in \mathcal{F}$ où $\mathcal{F}$ est un sous-ensemble fini de $\mathcal{C}^1(\mathbb{R}^n, \mathbb{R})$. Deuxièmement, pour montrer qu'un sous-ensemble S_J de S est contractile, on a aussi besoin d'être capable d'évaluer Df sur un intervalle pour toute fonction f de $\mathcal{F}$. Les paragraphes suivant montrent

comment cela est possible.

En appliquant le principe d'Hamilton, il est possible de montrer que les coordonnées cartésiennes de C sont données par :

$$\begin{array}{rcl} x_C & = & g_1(\alpha,\beta) = 2\cos\alpha + \beta \\ y_C & = & g_2(\alpha,\beta) = 2\sin(\alpha) - a\cosh\left(\frac{\beta}{a}\right) + a \end{array} \tag{3.57}$$

où a est le réel positif solution de l'équation (3.58) dépendante du paramètre réel positif β :

$$a\sinh\left(\frac{\beta}{a}\right) - L = 0 \tag{3.58}$$

Soit $z = \frac{\beta}{a}$, par conséquent l'équation (3.58) est équivalente à l'équation (3.59) d'inconnue z.

$$f(\beta, z) = \sinh(z) - \frac{L}{\beta}z = 0 \tag{3.59}$$

On a

$$Df(\beta, z) = \left(\frac{Lz}{\beta^2}, \cosh(z) - \frac{L}{\beta}\right) \tag{3.60}$$

Avec $\beta = \beta_0$, le tableau 3.52 présente les variations de la fonction :

$$(\mathbb{R}^{+*} \ni z \mapsto f(\beta_0, z) \in \mathbb{R}). \tag{3.61}$$

z	0		z_0		z^*		∞
$D_2f(\beta_0,z)$		$-$	0		$+$		
$f(\beta_0, .)$	0	$\searrow$	$f(\beta_0,z_0) < 0$	$\nearrow$	0	$\nearrow$	∞

FIG. 3.52 – Variations de la fonction $z \mapsto f(\beta_0, z)$ où $z_0 = \arccos(\frac{L}{\beta})$.

Grâce au théorème des valeurs intermédiaires, et comme $D_2f(\beta_0, z) > 0$ si $z \in]z_0; +\infty[$, il existe un unique z^* in $]z_0; +\infty[$ tel que $f(\beta_0, z^*) = 0$. Par conséquent, la fonction notée ϕ, qui associe à un nombre réel positif β_0 le réel positif z^* vérifiant $f(\beta_0, z^*) = 0$ est bien définie.

Les coordonnées cartésiennes de C sont données par :

$$\begin{array}{rcl} x_C & = & g_1(\alpha,\beta) = 2\cos\alpha + \beta \\ y_C & = & g_2(\alpha,\beta) = 2\sin(\alpha) - \frac{\beta}{\phi(\beta)}\cosh\phi(\beta) + \frac{\beta}{\phi(\beta)} \end{array} \tag{3.62}$$

On en déduit que l'ensemble des configurations admissibles peut être décrit par l'expression booléenne sans quantificateur d'atomes $f \leq 0$, $f \in \mathcal{F}$ où $\mathcal{F}$ est une famille finie de $\mathcal{C}^1(\mathbb{R}^n, \mathbb{R})$:

$$S = \{(\alpha, \beta) \in [0, \pi] \times [\beta_-, \beta_+] \text{ tel que } c_1 \wedge (c_2 \vee c_3 \vee c_4 \vee c_5)\} \tag{3.63}$$

où

$$\begin{array}{rcrcrcl} c_1 & \Leftrightarrow & (f_1(\alpha,\beta) & = & -g_2(\alpha,\beta) & \leq & 0) \\ c_2 & \Leftrightarrow & (f_2(\alpha,\beta) & = & g_1(\alpha,\beta) - l & \leq & 0) \\ c_3 & \Leftrightarrow & (f_3(\alpha,\beta) & = & -g_1(\alpha,\beta) + r & \leq & 0) \\ c_4 & \Leftrightarrow & (f_4(\alpha,\beta) & = & g_2(\alpha,\beta) - b & \leq & 0) \\ c_5 & \Leftrightarrow & (f_5(\alpha,\beta) & = & -g_2(\alpha,\beta) + t & \leq & 0) \end{array} \tag{3.64}$$

et $\square =]l, r[\times]b, t[$.

Il reste seulement à vérifier que pour tout i de $\{1, \ldots, 5\}$, f_i est une fonction $\mathcal{C}^1$, et comment Df_i peut être calculé. Les fonctions f_2 et f_3 sont clairement différentiables comme g_1 est $\mathcal{C}^1$. On peut obtenir Df_2 et Df_3 à partir de Dg_1 :

$$\begin{aligned} D_1 g_1(\alpha, \beta) &= -2\sin(\alpha) \\ D_2 g_1(\alpha, \beta) &= 1 \end{aligned}$$

Les fonctions f_1, f_4 et f_5 sont différentiables si g_2 l'est, et g_2 est différentiable si ϕ l'est. Montrons donc que ϕ est différentiable. Comme $D_2 f(\beta, z^*) > 0$, $D_2 f(\beta, z^*)$ est un isomorphisme de $\mathbb{R}$, à partir du théorème des fonctions implicites, on en déduit que ϕ est une fonction $\mathcal{C}^1$ et de plus :

$$\phi'(\beta) = -D_2^{-1} f(\beta, \phi(\beta)) \circ D_1 f(\beta, \phi(\beta)) \tag{3.65}$$

$$\phi'(\beta) = -\frac{\frac{L\phi(\beta)}{\beta^2}}{\cosh(\phi(\beta)) - \frac{L}{\beta}} = \frac{-L\phi(\beta)}{\beta^2 \cosh(\phi(\beta)) - \beta L} \tag{3.66}$$

On obtient :

$$\begin{aligned} D_1 g_2(\alpha, \beta) &= 2\cos(\alpha) \\ D_2 g_2(\alpha, \beta) &= \frac{L}{\beta \cosh(\phi(\beta)) - L} \left(\frac{1 - \cosh \phi(\beta)}{\phi(\beta)} + \sinh \phi(\beta) \right) + \frac{1 - \cosh \phi(\beta)}{\phi(\beta)} \end{aligned}$$

La figure 3.53 montre l'ensemble des configurations admissibles S avec $\square =]0.8, 1.2[\times]0.5, 0.7[$, $L = 4$ et $(\alpha, \beta) \in [0, \pi] \times [1/2, 7/4]$. Elle donne aussi une réalisation du complexe simplicial généré par l'algorithme `HIA`. Ce complexe simplicial $\mathcal{K}(S)$ peut être collapsé, et on peut affirmer que S est du même type d'homotopie qu'un cercle.

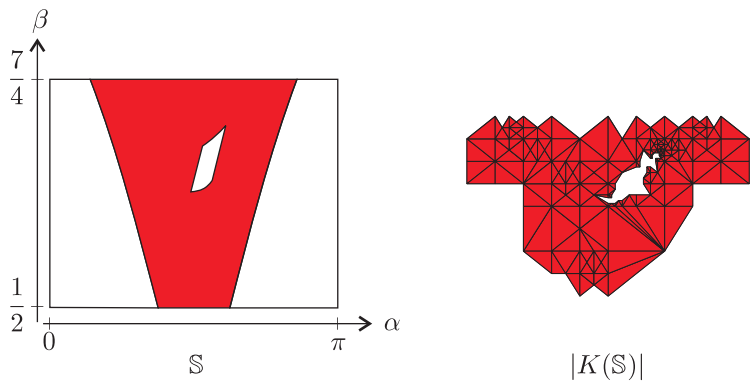


FIG. 3.53 – L'ensemble des configurations admissibles S et la triangulation générée par `HIA`.

Les applications pour la planification de trajectoires sont multiples. De la même manière qu'avec l'algorithme de la section 3.3, on peut créer un chemin qui relie deux configurations qui sont dans la même composante connexe. Mais on peut aussi affirmer, sachant que $\pi_1(S) = \mathbb{Z}$, qu'il existe deux grandes classes de chemins qui relient deux points, ceux qui passent à "gauche" du trou et ceux qui passent à "droite". Supposons que nous ayons à notre disposition une méthode itérative qui améliore un chemin proposé reliant A et B. Sachant que $\pi_1(S) = \mathbb{Z}$, on sait qu'il existe deux points fixes à cette méthode.

3.5.6 Calcul par intervalles sur les variétés

Les algorithmes précédemment présentés sont principalement basés sur le calcul par intervalles dans $\mathbb{R}^n$. Comme à l'habitude en géométrie, grâce à un atlas, (une certaine famille de fonctions $\{\varphi_i : \mathcal{O}_i \to \mathbb{R}^n\}$), il est possible, théoriquement, de faire du calcul par intervalles sur une variété M. Dans cette section, on n'essaie pas de donner les conditions nécessaires pour combiner le calcul par intervalles et les variétés. On se contente de donner quelques pistes basées sur des exemples.

Définition 3.5.2

On dira qu'un espace topologique M est une variété ($\mathcal{C}^\infty$) de dimension n s'il existe un atlas de M, i.e. un recouvrement de M par des ouverts U_i sur lesquels sont définies des cartes locales permettant d'identifier $\mathcal{O}_i$ à un ouvert de $\mathbb{R}^n$ avec une compatibilité ($\mathcal{C}^\infty$) entre les cartes locales, c'est-à-dire qu'il existe des applications φ_i telles que :

1. $M \subset \cup_{i \in I}$.
2. $\varphi_i : \mathcal{O} \to U_i$ est un homéomorphisme de $\mathcal{O}_i$ sur un ouvert U_i de $\mathbb{R}^n$.
3. si $\mathcal{O}_i \cap \mathcal{O}_j \neq \emptyset$, alors $(\varphi_i \circ \varphi_j^{-1})_{|\varphi_j(\mathcal{O}_i \cap \mathcal{O}_j)}$ est un difféomorphime de classe $(\mathcal{C}^\infty)$ sur $\varphi_j(\mathcal{O}_i \cap \mathcal{O}_j)$.

Exemple 3.5.3

1. $\mathbb{R}^n$ est une variété de dimension n, avec comme atlas l'application identité définie sur l'ouvert $\mathbb{R}^n$
2. le cercle est une variété de dimension 1, avec comme atlas $\mathcal{O}_0 = \{\theta \neq 0[2\pi]\}$ et $\mathcal{O}_\pi = \{\theta \neq \pi[2\pi]\}$, les applications φ_i consistant à prendre la détermination de l'angle respectivement dans les intervalles $]0, 2\pi[$ et $]-\pi, \pi[$. Le changement de carte est une application affine, donc $\mathcal{C}^\infty$
3. ce n'est pas le seul atlas que l'on peut considérer sur le cercle, la figure suivante donne un atlas avec 4 cartes.

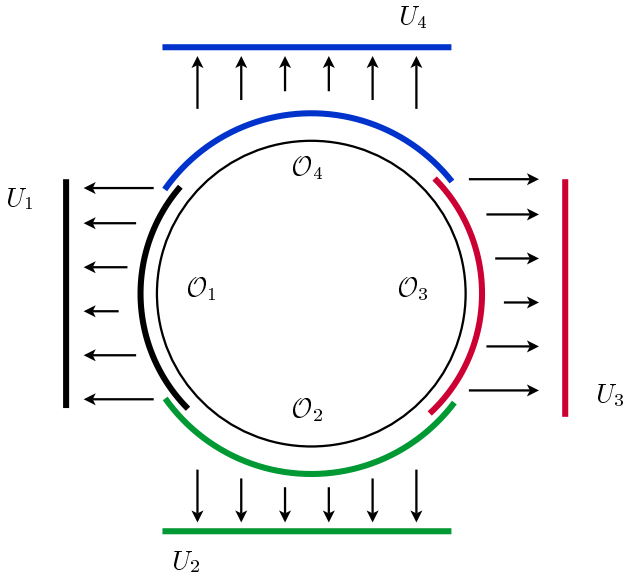


FIG. 3.54 – Un atlas du cercle à 4 cartes

4. Le tore est une variété de dimension 2, avec par exemple comme atlas :

$$\begin{aligned}
\mathcal{O}_{--} &= \{(x, y) \mid x \neq 0[1], y \neq 0[1]\}, \\
\mathcal{O}_{-+} &= \{(x, y) \mid x \neq 0[1], y \neq 0.5[1]\}, \\
\mathcal{O}_{+-} &= \{(x, y) \mid x \neq 0.5[1], y \neq 0[1]\}, \\
\mathcal{O}_{++} &= \{(x, y) \mid x \neq 0.5[1], y \neq 0.5[1]\}
\end{aligned} \tag{3.67}$$

les applications φ consistant à prendre le représentant de la classe d'équivalence de (x, y) respectivement dans les intervalles

$$]0,1[\times]0,1[,]0,1[\times]0.5,1.5[,]0.5,1.5[\times]0,1[,]0.5,1.5[\times]0.5,1.5[\tag{3.68}$$

Il existe une autre façon de créer des variétés de façon beaucoup moins rigide. On obtient une variété en recollant certains morceaux d'une première variété. Par exemple, on peut considérer, au moins topologiquement, le cercle comme un segment où l'on a identifié ses extrémités. La figure 3.55 montre comment on peut imaginer ceci.

FIG. 3.55 – Le cercle est homéomorphe à un segment dont on a identifié les extrémités.

De façon plus rigoureuse, on crée sur le segment $[0, 1]$ une relation d'équivalence $\mathcal{R}$. Cette relation $\mathcal{R}$ est définie de la façon suivante :

$$x \mathcal{R} y \Leftrightarrow ((x = 0 \wedge y = 1) \vee (x = 1 \wedge y = 0) \vee (x = y)) . \tag{3.69}$$

Avec la topologie quotient, on peut donc écrire $[0, 1]/\mathcal{R} \simeq S^1$ où S^1 est le cercle. De la façon équivalence, le tore peut être obtenu en recollant les faces du carré comme indiqué sur la figure suivante :

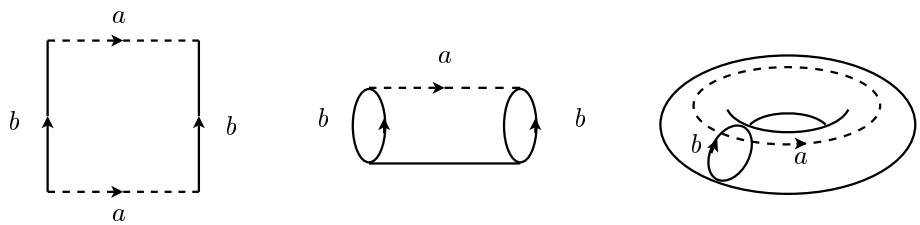


FIG. 3.56 – Le tore homéomorphe a un carré dont on a identifié les cotés comme indiqué sur cette figure.

Cette variété peut servir d'espace des configurations du robot de la section 3.5.1. Les angles α et β vivent respectivement sur le cercle S^1. En effet, lorsque α vaut $-\pi$ ou π, le robot est dans la même configuration. Donc l'espace des configurations du robot 3.5.1 est le produit cartésien de deux cercles *i.e.* un tore. Pour appliquer les algorithmes 9 et 8, on considère les cartes suivantes :

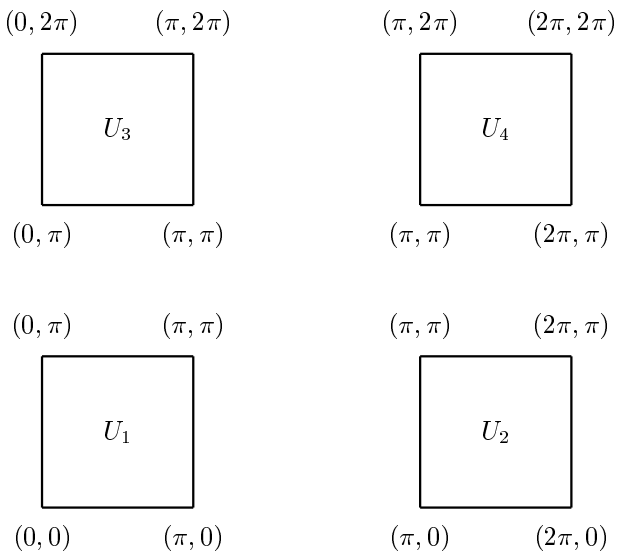


FIG. 3.57 – Le tore est homéomorphe à quatre carrés collés comme le montre cette figure.

Et finalement, le résultat obtenu par l'algorithme 4 est le suivant :

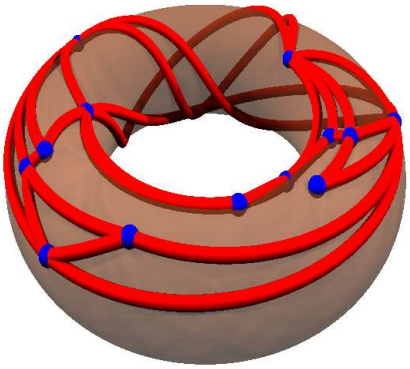


FIG. 3.58 – Graphe parsemé d'étoiles sur le tore obtenu pour l'espace des configurations admissibles du robot 3.5.1.

3.6 Conclusion

Dans ce chapitre, on a présenté deux algorithmes basés sur le calcul par intervalles pour étudier la topologie d'ensembles décrits par des inégalités. L'avantage de ces algorithmes est d'être garanties, implémentables (et implémentés). Par contre,

à ce jour, aucune étude sérieuse n'a été accomplie sur leur terminaison. Ils semblent néanmoins très prometteurs au vue de la conjecture 3.4.8. L'autre contribution de ce travail est la possibilité de prendre en compte des incertitudes sur les modèles. En effet dans l'exemple du robot à deux bras de ce chapitre, on aurait pu supposer que les longueurs des bras étaient seulement comprises dans des intervalles de longueurs. La méthode aurait, modulo un changement de topologie, terminé.

Chapitre 4

Systèmes dynamiques

4.1 Introduction

On se donne un système qui puisse être dans 9 états notés $x_1, \ldots, x_9$ dont la dynamique est donnée par la figure suivante :

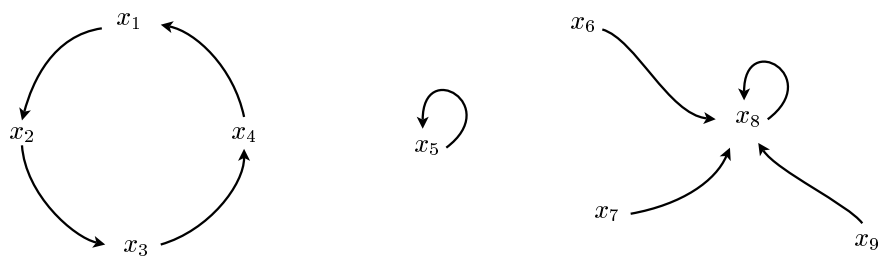


FIG. 4.1 – Exemple de système dynamique.

Les flèches symbolisent les changements d'états. Par exemple, la flèche $x_1 \to x_2$ signifie que si le système est dans l'état x_1 à l'instant t alors il est dans l'état x_2 à l'instant $t+1$. On dit que x_5 est un état d'équilibre puisque si on prend comme condition initiale x_5 alors le système demeure dans cet état dans le futur. L'état x_8 est aussi un point d'équilibre. L'ensemble des états qui mènent à cette situation d'équilibre forme ce qui s'appelle le bassin d'attraction de x_8. Le bassin d'attraction de x_8 est donc l'ensemble $\{x_6, x_7, x_8, x_9\}$. De même, le bassin d'attraction de x_5 est $\{x_5\}$.

Il existe deux grands modèles pour les systèmes dynamiques : les systèmes à temps continus et les systèmes à temps discrets. On peut de la même manière qu'introduite précédemment définir la notion de bassin d'attraction d'un point

d'équilibre dans le cas des systèmes continus. Jusqu'à présent, pour un système à temps continu donné, il n'existait pas de méthode permettant de trouver un ensemble dont on a la garantie qu'il est dans le bassin d'attraction d'un point. L'article de Roberto Genesio [4] donne un excellent état de l'art des méthodes numériques non garanties capables de construire un bassin d'attraction. On présente dans ce chapitre une démarche pour créer une suite d'ensembles qui converge vers le bassin d'attraction d'un point. Ce processus est décomposé en deux étapes.

- La première étape, basée sur la théorie de Lyapunov (des rappels sont donnés dans la section 4.3.1) et sur le calcul par intervalles, permet de montrer la stabilité d'un point d'équilibre. La méthode proposée permet en outre de construire un voisinage qui est contenu dans le bassin d'attraction de ce point. Ceci est détaillé dans la section 4.3.
- La seconde étape affine le voisinage obtenu à la première étape en combinant la théorie des graphes et les méthodes d'inclusion différentielles (rappelées dans la section 4.4.1). Cette méthode est développée dans la section 4.4.

Avant de détailler cette démarche, on donne quelques définitions concernant les systèmes dynamiques et la notion de stabilité.

4.2 Rappels - Systèmes dynamiques

Les systèmes dynamiques sont des systèmes qui évoluent au cours de temps. De tels systèmes sont causaux, c'est-à-dire que leur avenir ne dépend que de phénomènes du passé ou du présent. Ils sont aussi déterministes : à une " condition initiale" donnée à l'instant "présent" va correspondre un seul état "futur" possible à chaque instant ultérieur.

L'évolution déterministe du système dynamique admet en général l'une de ces deux modélisations :

- une évolution continue dans le temps, représentée par une équation différentielle ordinaire. C'est a priori la plus naturelle physiquement, puisque le paramètre temps nous semble continu. Ce genre de comportement dynamique est représenté par une équation de la forme :

$$\dot{x} = f(x) \tag{4.1}$$

  où $x : \mathbb{R} \to M$ est différentiable à valeur dans une variété différentielle M et f une fonction qui à un point x de M associe un vecteur du plan tangent à

M en x noté T_xM.

– une évolution discontinue dans le temps. Ce genre de comportement dynamique est représenté par une équation de la forme :

$$x_{n+1} = \varphi(x_n) \tag{4.2}$$

où x_0 est un élément d'un ensemble M et φ est une fonction $M \to M$. Dans le cas où l'ensemble M est fini, on peut représenter graphiquement la dynamique de φ à l'aide d'un graphe.

On donne maintenant un exemple pour chacune des modélisations.

Exemple 4.2.1

L'espace des configurations est $M = \mathbb{R}^2$. Ici comme $M = \mathbb{R}^2$, l'espace tangent à M au point x peut être vu comme $\mathbb{R}^2$. On note par TM l'ensemble $\cup_{x \in M} T_xM$ et par f la fonction $M \to TM$ définie par l'expression suivante :

$$\begin{pmatrix} \dot{x_1} \\ \dot{x_2} \end{pmatrix} = \begin{pmatrix} -x_2 \\ x_1 - (1 - x_1^2)x_2 \end{pmatrix} \tag{4.3}$$

On peut illustrer graphiquement cette dynamique via la figure suivante. Le champ de vecteurs est normalisé pour une meilleure lisibilité.

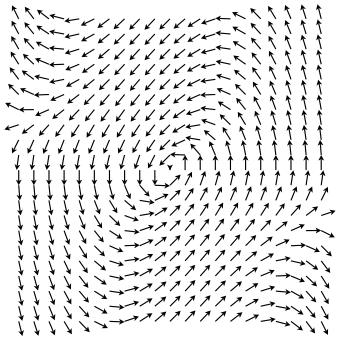


FIG. 4.2 – Exemple de système dynamique à évolution continue.

Exemple 4.2.2

L'espace des configurations est $M = \{x_1, \ldots, x_5\}$ et la fonction $\varphi : M \to M$ est la fonction suivante :

$$\begin{array}{rccc} \varphi : & M & \to & M \\ & x_1 & \mapsto & x_2 \\ & x_2 & \mapsto & x_3 \\ & x_3 & \mapsto & x_4 \\ & x_4 & \mapsto & x_1 \\ & x_5 & \mapsto & x_5 \end{array} \tag{4.4}$$

On peut représenter cette dynamique grâce au graphe suivant :

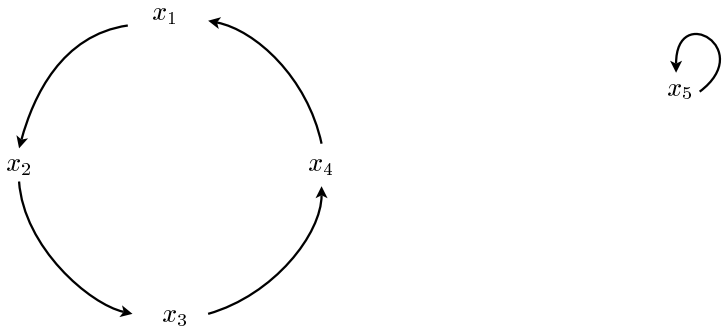


FIG. 4.3 – Exemple de système dynamique à évolution discontinue.

Dans la suite de ce chapitre, on s'intéresse aux systèmes dynamiques de la forme

$$\dot{x} = f(x) \tag{4.5}$$

où $f : M \to TM$ est une fonction différentiable. On note par $\{\varphi^t\}_{t\in\mathbb{R}}$ la famille de fonctions $M \to M$ indexées par $t \in \mathbb{R}$ telles que

$$\left.\frac{d}{dt}\right|_{t=0} \varphi^t x = f(x). \tag{4.6}$$

Cette famille de fonctions est appelée le flot associé au champ de vecteurs $x \mapsto f(x)$. Lorsqu'on peut définir le flot pour tout $t \in \mathbb{R}$, on dit que le champ de vecteurs est *complet*. Dans ce cas cette famille est un groupe indexé par les réels, en effet on a $\varphi^0 = Id$ et $\varphi^t \circ \varphi^{t'} = \varphi^{t+t'}$. Pour toute la suite, on supposera les champs de vecteurs complets. Cette hypothèse de complétude n'est pas démesurée. En effet, en pratique, pour pouvoir appliquer les méthodes proposées, les espaces M sont compacts et la compacité implique que tout champ de vecteurs différentiable est complet. On a ainsi l'existence du flot global et le problème de savoir si l'équation 4.5 admet une unique solution ne se posera pas.

Dans toute la suite, M sera une partie compacte de $\mathbb{R}^n$. On rappelle maintenant quelques définitions de stabilité.

Définition 4.2.3

Un sous-ensemble D de $\mathbb{R}^n$ est stable si $\varphi^{\mathbb{R}^+}(D) \subset D$, où $\varphi^{\mathbb{R}^+}(D) = \{\varphi^t(x), x \in D, t \in \mathbb{R}^+\}$

Définition 4.2.4

Soit D et D' deux sous-ensembles de $\mathbb{R}^n$ tels que $D \subset D'$.

Un point d'équilibre x_∞ est asymptotiquement (D, D')-stable si $\varphi^{\mathbb{R}^+}(D) \subset D'$ et

$\varphi^\infty(D) = \{x_\infty\}$, où $\varphi^\infty(D)$ dénote l'ensemble $\{x_\infty \in \mathbb{R}^n \mid x_\infty = \lim_{t\to\infty} \varphi^t(x), x \in D\}$. Cette notion est illustrée sur la Figure 4.4.

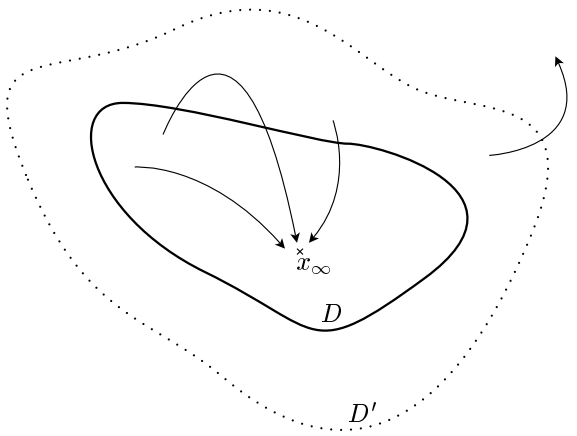


FIG. 4.4 – Le point x_∞ est asymptotiquement (D, D')-stable.

4.3 Preuve de la stabilité

Les points d'équilibre sont caractérisés par la condition $f(x) = 0$, par conséquent, on peut utiliser la méthode de Newton par intervalles pour montrer l'existence et l'unicité d'un point d'équilibre sur un intervalle donné. Considérant cette unicité et existence prouvées, la section suivante propose une méthode pour vérifier que ce point d'équilibre est asymptotiquement stable. On combine le calcul par intervalles et la théorie de Lyapunov. L'algorithme qui découle de ces résultats permet en outre de construire un ensemble $[x]$ qui est contenu dans le bassin d'attraction de ce point asymptotiquement stable.

4.3.1 Théorie de Lyapunov

Pour montrer la stabilité d'un système dynamique, la plupart des méthodes sont basées sur la théorie de Lyapunov. Elle consiste à créer une fonction réelle L qui ressemble à une sorte de *fonction énergie*. Pour affirmer l'asymptotique stabilité d'un point, il suffit de vérifier

1. qu'en ce point l'énergie est nulle,
2. qu'ailleurs elle est positive,
3. et que de plus, toute particule qui subit l'action du flot voit son énergie strictement décroître.

Définition 4.3.1
Soit D' un sous ensemble de $\mathbb{R}^n$ et x_∞ un point à l'intérieur de D'. Une fonction différentiable réelle L est une fonction de Lyapunov pour $\dot{x} = f(x)$ si :

1. $L(x) = 0 \Leftrightarrow x = x_\infty$.
2. $x \in D' - \{x_\infty\} \Rightarrow L(x) > 0$.
3. $\langle \nabla L(x), f(x) \rangle < 0,\ \forall x \in D' - \{x_\infty\}$.

où $\langle \cdot, \cdot \rangle$ est le produit scalaire sur $\mathbb{R}^n$. Cette définition est motivée par le résultat suivant qui donne une condition suffisante de stabilité.

Théorème 4.3.2 *Si $L : D' \to \mathbb{R}$ est une fonction de Lyapunov pour le système dynamique (4.5) alors il existe un sous-ensemble D de D' tel que le point $x_\infty \in D$ (vérifiant $L(x_\infty) = 0$) soit asymptotiquement (D, D')-stable.*

La preuve de ce théorème peut être trouvée dans le livre de Slotine [39]. Par ailleurs, ce livre constitue une excellente introduction à la théorie de Lyapunov. Le lecteur intéressé par cette théorie pourra aussi consulter [50]. Concrètement, pour démontrer l'asymptotique stabilité d'un point pour un système donné, il suffit de :

1. trouver un candidat pour la fonction de Lyapunov,
2. vérifier que ce candidat est effectivement de Lyapunov.

A priori, la fonction L appartient au moins à $C^1(\mathbb{R}^n, \mathbb{R})$ qui est un espace vectoriel de dimension infinie. On pourrait rechercher un candidat dans cet espace fonctionnel, mais en pratique il est préférable de chercher L dans un espace de dimension finie.
Chercher L dans le dual de $\mathbb{R}^n$ est vain car il n'existe pas dans $\mathbb{R}^{n^*} - \{0\}$ de fonction qui vérifie $L(x) > 0$ pour x dans un voisinage de 0. On arrive naturellement à chercher L parmi les formes quadratiques.
Ainsi L est une fonction de la forme $L(x) = x^t W x$ où W est une matrice symétrique carré. Il est bien connu [39], dans le cas linéaire ($\dot{x} = Ax, A \in \mathbb{R}^{n \times n}$), que l'origine 0 est asymptotiquement stable si et seulement s'il existe deux matrices définies positives S^{n+} telles que

$$A^T W + W A = -I. \tag{4.7}$$

[1]

[1] Cette équation provient d'une réécriture de la dernière condition de Lyapunov $\langle \nabla L(x), f(x) \rangle = \frac{d}{dt}\Big|_{t=0} L(x(t)) = \frac{d}{dt}\Big|_{t=0} (x^T(t) W x(t)) = \dot{x}^T(0) W x(0) + \dot{x}^T(0) W x(0) = x^T (A^T W + W A) x$.

Résoudre cette équation d'inconnue W revient à résoudre un système linéaire. Quand la matrice W est définie positive, toutes les conditions du théorème 4.3.2 sont réunies, et par conséquent, 0 est asymptotiquement stable. En d'autres termes, dans le cas linéaire, une méthode effective, basée sur la théorie de Lyapunov, pour montrer l'asymptotique stabilité d'un point existe[2]. L'algorithme présenté plus loin est partiellement basé sur cette approche.

4.3.2 Vérification de l'existence d'une fonction de Lyapunov.

Il existe de nombreux travaux, utilisant le calcul par intervalles, dédiés au problème de stabilité d'un point d'équilibre d'un système dynamique non linéaire [44] [45] [46] [47]. Dans cette section, on présente un théorème que l'on appliquera pour montrer l'asymptotique stabilité d'un point d'équilibre.

Définition 4.3.3
Avec $[x]$ un intervalle de $\mathbb{R}^n$, on note par $B(r,[x])$ l'ensemble $\{x \in \mathbb{R}^n, \min_{a\in[x]} \|a-x\| < r\}$. Soit d la fonction définie sur $\mathbb{IR}^n \times \mathbb{IR}^n$ par

$$d : ([x],[y]) \mapsto \sup\{r \in \mathbb{R} \mid B(r,[x]) \subset [y]\}. \tag{4.8}$$

Théorème 4.3.4 *Considérons le système dynamique (4.5) et une matrice $W \in S^{n+}$ dont les valeurs propres maximum et minimum sont respectivement λ_{max} et λ_{min}. Définissons la fonction g_a paramétrée par a via $g_a(x) = -\langle W(x-a), f(x)\rangle$. Si $[x_\infty]$ est un pavé inclus dans le pavé $[x_0]$ et $[x]$ est un pavé qui a pour centre $[x_\infty]$, rayon $\sqrt{n\frac{\lambda_{min}}{\lambda_{max}}}d([x_\infty],[x_0])$ alors on a l'implication suivante :*

1. *il existe un unique $x_\infty \in [x]$ à l'intérieur de $[x_\infty]$, tel que $f(x_\infty) = 0$.*
2. *$\nabla^2 g_{[x_\infty]}([x_0]) \subset S^{n+}$.*

implique que x_∞ est asymptotiquement $([x_0],[x])$-stable.

Preuve : Soit L_{x_∞} la forme quadratique définie par

$$\begin{array}{rccl} L_{x_\infty} : & D & \to & \mathbb{R} \\ & x & \mapsto & (x-x_\infty)^T W (x-x_\infty) \end{array} \tag{4.9}$$

Comme $W \in S^{n+}$, on a :

1. $L_{x_\infty}(x) = 0 \Leftrightarrow x = x_\infty$

[2]Ce n'est pas la seule façon de la montrer, en effet il suffit de vérifier que les valeurs propres de A sont à partie réelle négative.

2. $x \in D - \{x_\infty\} \Rightarrow L_{x_\infty}(x) > 0$

Avec $h(x) = -\langle \nabla L_{x_\infty}(x), f(x) \rangle$, pour montrer que L_{x_∞} est de Lyapunov, il suffit de vérifier que : $h([x_0]) \leq 0$. Par construction, on a :

1. $h(x_\infty) = 0$ et $\nabla h(x_\infty) = 0$.

2. $\nabla^2 h([x_0]) \subset S^{n+}$ car $\nabla^2 h([x_0]) \subset 2\nabla^2 g_{[x_\infty]}([x_0])$.

En appliquant le théorème 2.4.4 à h, on en conclut que L_{x_∞} est de Lyapunov pour le système dynamique (4.5). Par conséquent, il existe un sous-ensemble $[x]$ de $[x_0]$ et $x_\infty \in [x]$ tels que :

$$\begin{cases} \varphi^{+\infty}([x]) = \{x_\infty\}. \\ \varphi^t([x]) \subset [x_0], \forall t \in \mathbb{R}^+. \end{cases} \tag{4.10}$$

Soit $\mathcal{E}$ une ellipsoïde W, dont le centre x_∞, et dont le grand axe a pour longueur $\sqrt{\lambda_{min}} d([x_\infty], [x_0])$. Évidemment, l'ensemble $\mathcal{E}$ est inclus dans $[x_0]$ et est stable. Par conséquent, un intervalle $[x]$ qui a pour centre un point de $[x_\infty]$ et pour rayon $\sqrt{n \frac{\lambda_{min}}{\lambda_{max}}} d([x_\infty], [x_0])$ est, par construction, inclus dans l'ellipsoïde $\mathcal{E}$. Par conséquent, x_∞ est asymptotiquement $([x_0], [x])$-stable.

□

Pour un système dynamique donné $\dot{x} = f(x)$ et un ensemble $D' = [x_0]$, l'algorithme qui va être présenté montre qu'il existe un unique point d'équilibre x_∞, dans un ensemble calculé $D = [x]$. Il montre aussi que x_∞ est asymptotiquement (D, D')-stable. L'ensemble $D = [x]$ est par conséquent inclus dans le bassin d'attraction de x_∞.

4.3.3 Algorithme

L'idée principale de cet algorithme est premièrement de linéariser le système dynamique autour d'un point voisin du point d'équilibre. Dans une seconde étape, on vérifie que la fonction de Lyapunov créée pour le système linéarisé est aussi une fonction de Lyapunov pour le système de départ en usant des résultats donnés dans la partie 2.4.3. Ceci peut être résumé dans l'algorithme 10.

Alg. 10

Entrée: Un intervalle $[x_0]$ de $\mathbb{R}^n$ et un système dynamique

$$\dot{x} = f(x) \tag{4.11}$$

où $f \in \mathcal{C}^\infty(D, \mathbb{R}^n)$.

Sortie: un intervalle $[x]$ et une preuve qu'il existe $x_\infty \in [x]$ qui est asymptotyquement $([x], [x_0])$-stable.

1: $[x_\infty] \leftarrow$ Algorithme de Newton par intervalles $f(x) = 0$, $x \in [x_0]$.

2: $\tilde{x}_\infty \leftarrow$ un élément de $[x_\infty]$.

3:

$$A \leftarrow \left.\frac{df}{dx}\right|_{x=\tilde{x}_\infty} \tag{4.12}$$

4: Résoudre l'équation $A^T W + WA = -I$ d'inconnue W.

5: **si** $W \in S^{n+}$ et $\nabla^2 g_{[x_\infty]}([x_0]) \subset S^{n+}$ **alors**

6: Retourne "L'intervalle de centre $\tilde{x}_\infty$ et de taille $\sqrt{n}\sqrt{\frac{\lambda_{min}}{\lambda_{max}}} d([x_0], [x_\infty])$ vérifie :

$$\varphi^{\mathbb{R}^+}([x]) \subset [x_0] \text{ et } \varphi^{+\infty}([x]) = x_\infty \in [x_\infty] \tag{4.13}$$

7: **fin si**

A l'étape 1, l'intervalle $[x_\infty]$ peut être calculé en utilisant la méthode Newton par intervalles présentée dans la section 2.3.3 page 30. Si l'algorithme de Newton ne retourne pas de $[x_\infty]$ inclus dans $[x_0]$ alors on retourne "failure".

A l'étape 4, le problème se ramène à la résolution d'un système linéaire. Comme précédemment, si la résolution est impossible alors on retourne "failure".

Enfin, à l'étape 5, on utilise les résultats présentés dans la section 2.4.3 pour montrer que :

$$\nabla^2 g_{[x_\infty]}([x_0]) \subset S^{n+} \tag{4.14}$$

4.3.4 Exemple illustratif

Dans cette section, la méthode précédemment proposée est discutée via l'exemple :

$$\begin{pmatrix} \dot{x_1} \\ \dot{x_2} \end{pmatrix} = \begin{pmatrix} -x_2 \\ x_1 - (1 - x_1^2)x_2 \end{pmatrix} \tag{4.15}$$

où $[x_0] = [-0.6, 0.6]^2$.

Premièrement, la méthode de Newton par intervalles est utilisée pour montrer que l'intervalle $[x_0]$ contient un unique point d'équilibre x_∞. De plus, on montre

que ce point fixe du flot est un élément de l'intervalle $[x_\infty] = [-0.02, 0.02]^2$. Puis, on "linéarise" le système dynamique autour du point $\tilde{x}_\infty = (0.01, 0.01)$.
Le champ de vecteurs associé à cette dynamique est représenté sur la figure 4.5. Cette figure montre aussi le champ de vecteurs linéarisé autour du point $\tilde{x}_\infty$. Dans ce cas, la fonction de Lyapunov créée à l'étape 4 de l'algorithme 10 est :

$$L_{x_\infty}(x) = (x - x_\infty)^T \begin{pmatrix} -1,51 & 0,49 \\ 0,49 & -1,01 \end{pmatrix} (x - x_\infty) \tag{4.16}$$

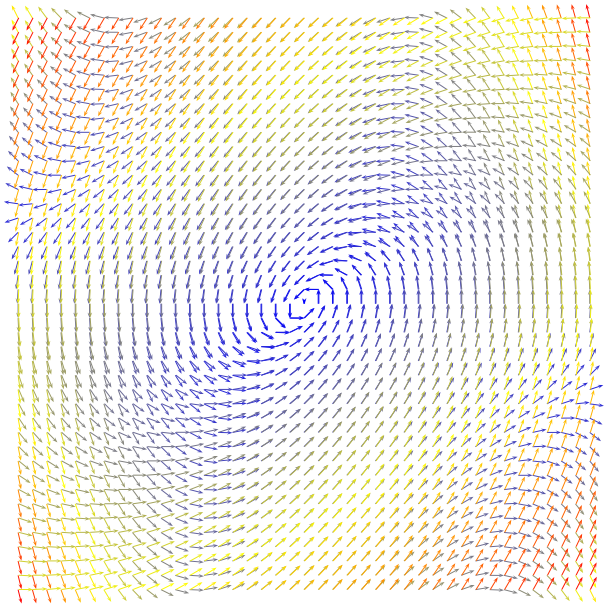


FIG. 4.5 – Champ de vecteurs "normalisé" et sa linéarisation au voisinage de $\tilde{x}_\infty$. Le champ de vecteurs linéaire est représenté par les lignes en pointillés.

Quelques lignes de niveau de L_{x_∞} sont représentées sur la figure 4.6. Dans un voisinage $[x_\infty]$, la fonction L_{x_∞} semble être une fonction de Lyapunov car les vecteurs $f(x)$ traversent les lignes de niveau de l'extérieur vers l'intérieur. Comme L_{x_∞} est une fonction de Lyapunov pour le système linéaire, la dernière interprétation géométrique est équivalente à $g_{x_\infty}(x) > 0, \forall x \in [x_0] - x_\infty$. Cette dernière est vraie car :

- $g_{x_\infty}(x_\infty) = 0$
- $\nabla g_{x_\infty}(x_\infty) = 0$
- $\nabla^2 g_{[x_\infty]}([x_0]) \subset S^{n+}$ comme $\nabla^2 g_{[x_\infty]}([x_0]) \subset [A]$

où $[A] = \begin{pmatrix} [-1.78, 5.78] & [-4.14, 4.15] \\ [-4.14, 4.15] & [0.56, 3.45] \end{pmatrix}$ est définie positive.

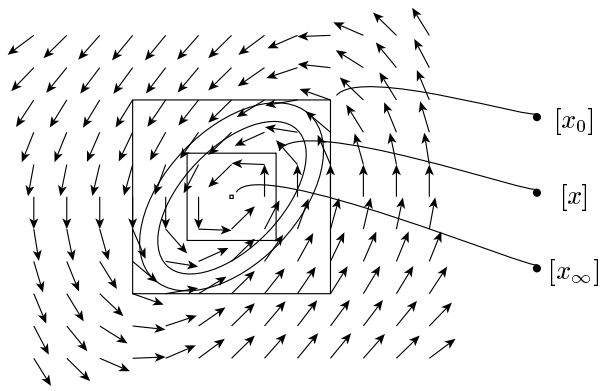


FIG. 4.6 – Lignes de niveau de la fonction de Lyapunov et un intervalle $[x_\infty]$ qui contient un unique point d'équilibre.

4.4 Bassin d'attraction

Il existe de nombreuses méthodes qui approchent le bassin d'attraction d'un point. L'article de Genesio [4] donne un aperçu global des méthodes et techniques employées. Aucune de ces méthodes n'est garantie.

Dans cette section, on présente une méthode capable d'approcher le bassin d'attraction pour un point d'équilibre asymptotiquement stable. L'approche proposée améliore l'ensemble $[x]$ retourné par l'algorithme 10 en combinant les méthodes d'inclusion de la section 4.4.1 et la théorie des graphes. Pour pouvoir mettre en relation des intervalles, on utilise des méthodes de résolutions numériques mais garanties d'équations différentielles ordinaires. Ces méthodes sont rappelées dans la section suivante.

4.4.1 Fonction d'inclusion du flot

Liouville a montré l'impossibilité de résoudre explicitement certaines équations différentielles même d'ordre peu élevé. En dehors des équations linéaires à coefficients constants (dont les solutions s'expriment à l'aide de l'exponentielle de matrice), il existe très peu d'équations différentielles dont on connaît les solutions. Par conséquent pour une équation différentielle ordinaire donnée, on ne peut pas, en général, trouver une expression pour le flot.

Néanmoins, il existe plusieurs méthodes numériques pour calculer, une fonction d'inclusion du flot. Etant donnée une équation différentielle ordinaire, un réel t et un intervalle $[x_0]$ de conditions initiales, ces méthodes construisent, sans résoudre l'équation différentielle, un intervalle $[x_t]$ tel que :

$$\varphi^t([x]) \subset [x_t]$$

Toutes ces méthodes sont basées sur le même schéma. Dans cette section, nous n'allons pas étudier les particularités de chacune de ces méthodes mais uniquement présenter leur principe.

La construction de cet intervalle $[x_t]$ est en deux étapes. La première étape consiste en la création de ce que Demailly [7] appelle un cylindre de sécurité, on calcule un intervalle $[\tilde{x}_t]$ qui vérifie :

$$\forall \tau \in [0, t], \varphi^\tau([x_0]) \subset [\tilde{x}_t] \tag{4.17}$$

On a donc $\varphi^t([x_0]) \subset [\tilde{x}_t]$, mais généralement la distance[3] de Hausdorff qui sépare $[\tilde{x}_t]$ de $\varphi^t([x_0])$ est grande. Dans un second temps, on calcule un intervalle $[x_t]$ à partir de $[\tilde{x}_t]$ tel que :

$$\varphi^t([x_0]) \subset [x_t] \subset [\tilde{x}_t] \tag{4.18}$$

Ce qui constitue l'étape 2. Dans la suite de cette section, on détaille chacune de ces étapes.

Etape 1

C'est en partie pour résoudre l'équation différentielle

$$\dot{x} = f(x) \tag{4.19}$$

que Newton a introduit la notion d'intégrale. En effet, une fonction $t \mapsto x(t)$ est solution de (4.19) si et seulement si x vérifie la relation :

$$x(t) = x(0) + \int_0^t f(x(\tau))d\tau. \tag{4.20}$$

La preuve du théorème de Cauchy-Lipschitz qui garantit l'existence et l'unicité d'une solution à l'équation (4.19) est basée sur l'équation (4.20). En effet, en considérant l'opérateur

$$P : x \mapsto \left(t \mapsto x(0) + \int_0^t x(\tau)d\tau \right), \tag{4.21}$$

[3]Cette distance peut bien être calculée car $[x_0]$ et $\varphi^t([x_0])$ sont bien des compacts de $\mathbb{R}^n$.

une fonction est solution de (4.19) si et seulement si elle est point fixe de l'opérateur[4] P. On exploite cette idée pour construire une fonction d'inclusion au flot. Etant donné un intervalle $[x_0]$ et t un réel, on cherche un intervalle $[x_t]$ tel que $\varphi^t([x_0]) \subset [x_t]$. On rappelle d'abord le théorème du point fixe de Banach.

Théorème 4.4.1 *Soit $\Phi : N \to N$ une fonction d'un espace métrique complet N dans lui même muni d'une métrique d. S'il existe $0 \leq \gamma < 1$ tel que*

$$\forall x, y \in M,\ d(\Phi(x), \Phi(y)) \leq \gamma d(x, y) \tag{4.22}$$

alors Φ admet un unique point fixe.

Définition 4.4.2
Soit $\alpha \in \mathbb{R}^+$ et $x : [0, t] \to \mathbb{R}^n$ une fonction continue, on note par $\|\cdot\|_\alpha$

$$\|x\|_\alpha = \max_{\tau \in [0,t]} e^{-\alpha\tau} \|x(\tau)\| \tag{4.23}$$

la norme exponentielle.

On peut montrer que si f est continuement différentiable alors il existe un $\alpha > 0$ tel que l'opérateur de Picard-Lindelöf soit contractant pour la métrique induite par $\|\cdot\|_\alpha$. Par conséquent, en supposant N métrique complet avec la distance $\|\cdot\|_\alpha$, si $PN \subset N$, alors P admet un unique point fixe dans N.

En pratique, on essaie de construire un espace N qui vérifie l'inclusion $PN \subset N$. Soient $[x_0]$ et $[\tilde{x}_t]$ deux intervalles de $\mathbb{IR}$, on note par $N_{[\tilde{x}_t]}$ l'ensemble des fonctions :

$$N_{[\tilde{x}_t]} = \{x : [0, t] \to [\tilde{x}_t] \text{ différentiable telle que } x(0) \in [x_0]\}. \tag{4.24}$$

Soit x un élément de $N_{[\tilde{x}_t]}$, l'opérateur de Banach-Lindelöf appliqué à x donne :

$$(Px)(t) = x(0) + \int_0^t f(x(\tau))d\tau \tag{4.25}$$

$$\subset x(0) + [0, t]f([\tilde{x}_t]) \tag{4.26}$$

Par conséquent, si on montre que $[\tilde{x}_t]$ vérifie l'inclusion

$$[x_0] + [0, t]f([\tilde{x}_t]) \subset [\tilde{x}_t], \tag{4.27}$$

alors on aura prouvé que $PN_{[\tilde{x}_t]} \subset N_{[\tilde{x}_t]}$. On pourra alors en déduire que P admet un unique point fixe dans $N_{[\tilde{x}_t]}$. Ce qui est équivalent à dire que la solution au problème

$$\dot{x} = f(x), x(0) \in [x_0] \tag{4.28}$$

[4]Cet opérateur est habituellement appelé opérateur de Picard-Lindelöf.

restreinte à l'intervalle $[0,t]$ est un élément de $N_{[\tilde{x}_t]}$. On aura $\forall \tau \in [0,t]$, $\varphi^\tau([x_0]) \subset [\tilde{x}_t]$. Les différentes méthodes existantes tentent de trouver $[\tilde{x}_t]$ qui vérifie l'inclusion 4.27. Ce sont des méthodes récursives. Dans tous les cas, on se donne un réel t positif et un intervalle $[\tilde{x}_t]$. Les deux méthodes jouent sur chacun des paramètres.

1. Méthode 1 : on fixe t et on agrandit $[\tilde{x}_t]$ ($[\tilde{x}_t] \leftarrow [\tilde{x}_t] + [x]$, avec $[x] \in \mathbb{IR}$) jusqu'à ce que (4.27) soit satisfaite.
2. Méthode 2 : on fixe $[\tilde{x}_t]$ et on diminue t ($t \leftarrow \alpha t$, $\alpha \in]0,1[$) jusqu'à ce que 4.27 soit satisfaite.

Dans chacune des deux méthodes, il existe des variantes. En particulier pour la méthode 1, on peut imaginer diverses stratégies pour caler le paramètre $[x]$, dans nos implémentations nous avons choisi $[x]$ de la forme $\beta[-1,1]^n$, $\beta \in \mathbb{R}$. Mais on peut imaginer une méthode qui agrandit $[\tilde{x}_t]$ dans des directions privilégiées. La seconde méthode est sans doute plus satisfaisante car pour des raisons de continuité si $[\tilde{x}_t]$ contient $[x_0]$ et si $\partial[\tilde{x}_t] \cap [x_0] = \emptyset$ alors cette méthode termine.

Etape 2

On suppose la première étape accomplie. On a donc à notre disposition un intervalle $[\tilde{x}_t]$ qui contient $\varphi^{[0,t]}([x_0])$. Dans cette seconde étape, on va utiliser cet intervalle $[\tilde{x}_t]$ pour calculer un intervalle $[x_t]$ tel que

$$\varphi^t([x_0]) \subset [x_t] \subset [\tilde{x}_t]. \tag{4.29}$$

On commence par supposer que f est une fonction suffisamment régulière de telle manière qu'une solution x de (4.19) soit de classe $\mathcal{C}^{n+1}$. La formule de Taylor donne pour tout t l'existence d'un réel ξ tel que

$$x(t) = \sum_{i=0}^{i=n} \frac{x^{(i)}(0)}{i!} t^i + \frac{x^{(i+1)}(\xi)}{(i+1)!} t^{i+1} \tag{4.30}$$

Comme la fonction x est l'inconnue de l'équation (4.19), les fonctions $x^{(i)}$ ne sont pas plus connues. Néanmoins, comme $\dot{x} = f(x)$, on peut en déduire que $x'(0) = f(x(0))$. De même, on en déduit que

$$x''(0) = \frac{d}{dt}\left(f(x(t))\right)_{t=0} = \left(\frac{df}{dx}\right)_{x=x(0)} \cdot \left(\frac{dx}{dt}\right)_{t=0} = \left(\frac{df}{dx}\right)_{x=x(0)} \cdot f(x(0)). \tag{4.31}$$

On généralise le processus précédent en posant

$$
\begin{aligned}
f^{[0]}(x) &= x && (4.32)\\
f^{[1]}(x) &= f(x) && (4.33)\\
&\vdots && (4.34)\\
f^{[i+1]}(x) &= \frac{1}{i+1}\frac{\partial f^{[i]}}{\partial x}\cdot f(x) && (4.35)\\
& && (4.36)
\end{aligned}
$$

Finalement, l'équation 4.30 devient :

$$x(t) = \sum_{i=0}^{i=n} f^{[i]}(x(0))t^i + f^{[i+1]}(x(\xi))t^{i+1} \tag{4.37}$$

Or, par hypothèse, $x(\xi) \in [\tilde{x}_t], \forall \xi \in [0,t]$, on a donc :

$$x(t) \in \sum_{i=0}^{i=n} f^{[i]}(x(0))t^i + f^{[i+1]}([\tilde{x}_t])t^{i+1} \tag{4.38}$$

Si on note par T_n la fonction qui à x associe $\sum_{i=0}^{i=n} f^{[i]}(x)t^i$, et $[T_n]$ une fonction d'inclusion pour T_n, alors la fonction qui à un intervalle $[x]$ associe l'intervalle $[T_n]([x]) + f^{[i+1]}([\tilde{x}_t])t^{i+1}$ est une fonction d'inclusion pour φ^t.

Remarque 4.4.3

En pratique, il est vivement déconseillé d'utiliser la fonction d'inclusion naturelle comme fonction d'inclusion pour T_n. En effet, comme la fonction T_n est de la forme : $T_n(x) = x + \ldots$, utilisant la fonction d'inclusion naturelle, on a alors :

$$m([T_n]([x])) \geq m([x]),\ \forall [x] \in \mathbb{IR}^n \tag{4.39}$$

Ce qui est peu satisfaisant, par exemple, dans le cas où le flot est contractant ! Dans les exemples présentés dans la suite, on a développé à l'ordre 2 et utilisé la fonction d'inclusion de la proposition 2.2.25 donnée page 25.

Remarque 4.4.4

Comme indiqué dans l'introduction de cette section, il existe de multiples variantes de la méthodologie présentée. On pourra consulter la thèse de Nedialkov [6] qui en donne un excellent aperçu.

Dans la section suivante, on donne une manière de discrétiser l'équation différentielle ordinaire. On crée un système dynamique non déterministe à temps discret. Ce système dynamique non déterministe contient certaines informations qui peuvent être exploitées pour approcher le bassin d'attration d'un point.

4.4.2 Discrétisation

Il est courant d'utiliser le mot "discrétisation" par exemple pour résoudre numériquement une équation différentielle ordinaire. On peut citer à titre d'exemple les méthodes d'Euler, Runge-Kutta...qui discrétisent le temps. Dans le cas où l'on veut montrer l'existence d'un cycle limite, on peut aussi utiliser le morphisme de premier retour de Poincaré [49] qui, d'une certaine manière, discrétise...Dans cette section, le mot "Discrétisation" a encore une utilisation différente.

On suppose que l'espace des configurations M est compact. On se donne un réel t et $\{\mathbb{S}_i\}_{i\in I}$ un recouvrement fini de M. On définit la relation (le graphe) sur le recouvrement $\{\mathbb{S}_i\}_{i\in I}$ avec

$$\mathbb{S}_i \mathcal{R}_t \mathbb{S}_j \Leftrightarrow \varphi^t(\mathbb{S}_i) \cap \mathbb{S}_j \neq \emptyset \tag{4.40}$$

De cette façon, les éléments du recouvrement vérifient :

$$\varphi^t(\mathbb{S}_i) \subset \bigcup_{\{j|\mathbb{S}_i \mathcal{R}_t \mathbb{S}_j\}} \mathbb{S}_j. \tag{4.41}$$

On peut aussi interpréter la relation $\mathcal{R}_t$ comme un système dynamique discret non déterministe :

$$\begin{array}{rcccl} \Phi^t & : & \{\mathbb{S}_i\}_{i\in I} & \rightarrow & 2^{\{\mathbb{S}_i\}_{i\in I}} \\ & & \mathbb{S}_i & \mapsto & \{\mathbb{S}_j \mid \mathbb{S}_i \mathcal{R}_t \mathbb{S}_j\} \end{array} \tag{4.42}$$

Dans la pratique, nous ne pouvons pas évaluer exactement la relation $\mathcal{R}_t$. En utilisant les outils présentés dans la section 4.4.1, on est capable de calculer une relation $\overline{\mathcal{R}_t}$:

$$\mathbb{S}_i \overline{\mathcal{R}_t} \mathbb{S}_j \Leftrightarrow [\varphi^t](\mathbb{S}_i) \cap \mathbb{S}_j \neq \emptyset \tag{4.43}$$

Il est facile de vérifier que

$$\mathcal{R}_t \subset \overline{\mathcal{R}_t}. \tag{4.44}$$

Dans la suite, on présente un algorithme qui utilise cette relation pour créer une suite d'ensembles qui tend vers le bassin d'attraction d'un point d'équilibre x_∞.

4.4.3 Algorithme

La méthode présentée ici est récursive. On crée une suite d'ensembles $\{A_n\}_{n\in\mathbb{N}}$ qui tend vers le bassin d'attraction de x_∞. On commence par initialiser A_0 avec l'intervalle $[x]$ obtenu via la méthode de la section 4.3.3. On crée un recouvrement de M avec des intervalles et on calcule la relation $\overline{\mathcal{R}_t}$.

L'itération principale se résume à

$$A_{n+1} = A_n \cup \bigcup_{i|\mathbb{S}_i\overline{\mathcal{R}_t}\mathbb{S}_j \Rightarrow \mathbb{S}_j \subset A_n} \mathbb{S}_i \tag{4.45}$$

L'inclusion 4.41 montre que la réunion des $\mathbb{S}_j$ qui sont en relation avec $\mathbb{S}_i$ ($\mathbb{S}_i\overline{\mathcal{R}_t}\mathbb{S}_j$ forme un ensemble qui contient $\varphi^t(\mathbb{S}_j)$. Tous les $\mathbb{S}_i$ qui ont une image via φ^t incluse dans le bassin d'attraction viennent compléter le bassin d'attraction (équation 4.45). L'exécution de ce processus itératif ressemble à un phénomène de diffusion. Au premier pas, seuls les pavés "voisins" de A_0 viennent "gonfler" le bassin d'attraction, puis A_n grandit jusqu'à atteindre un point fixe pour Φ^t. Dans ce cas, on affine le recouvrement et on réitère la démarche. Ce processus est résumé par l'algorithme suivant :

Alg. 11 `Bassin d'attraction`

Entrée:
1. Un système dynamique $\dot{x} = f(x)$ avec f définie sur un ensemble D de $\mathbb{R}^n$.
2. Un point d'équilibre x_∞ et un intervalle $[x]$ tels que x_∞ soit asymptotiquement $([x], D)-$stable.
3. Un réel λ tel que $0 < \lambda < m(D)$ et un réel $t > 0$.
4. Une précision réelle ϵ.

Sortie: Un ensemble A inclus dans le bassin d'attraction de x_∞.

Initialisation : $A_0 \leftarrow [x]$.
$\{\mathbb{S}_i\}_{i_i nI} \leftarrow$ un λ-recouvrement [5] de D.
répéter
  $\overline{\mathcal{R}_t} \leftarrow$ relation sur $\{\mathbb{S}_i\}_{i_i nI}$ définie par 4.43
  répéter
    $A_{n+1} \leftarrow A_n \cup \bigcup_{i|\mathbb{S}_i\overline{\mathcal{R}_t}\mathbb{S}_j \Rightarrow \mathbb{S}_j \subset A_n} \mathbb{S}_i$
  tant que $A_{n+1} \neq A_n$
  $\lambda \leftarrow \lambda/2$.
  $\{\mathbb{S}_i\}_{i \in I} \leftarrow$ un λ-recouvrement de D.
tant que $\lambda > \epsilon$
Retourne A.

Exemple 4.4.5

On considère le système dynamique suivant :

$$\begin{pmatrix} \dot{x} \\ \dot{y} \end{pmatrix} = \begin{pmatrix} x * (x^2 - x * y + 3 * y^2 - 1) \\ y * (x^2 - 4 * y * x + 3 * y^2 - 1) \end{pmatrix} \tag{4.46}$$

où $[x_0] = [-0.6, 0.6]^2$.

La figure suivante montre le champ de vecteurs associé à l'équation 4.46.

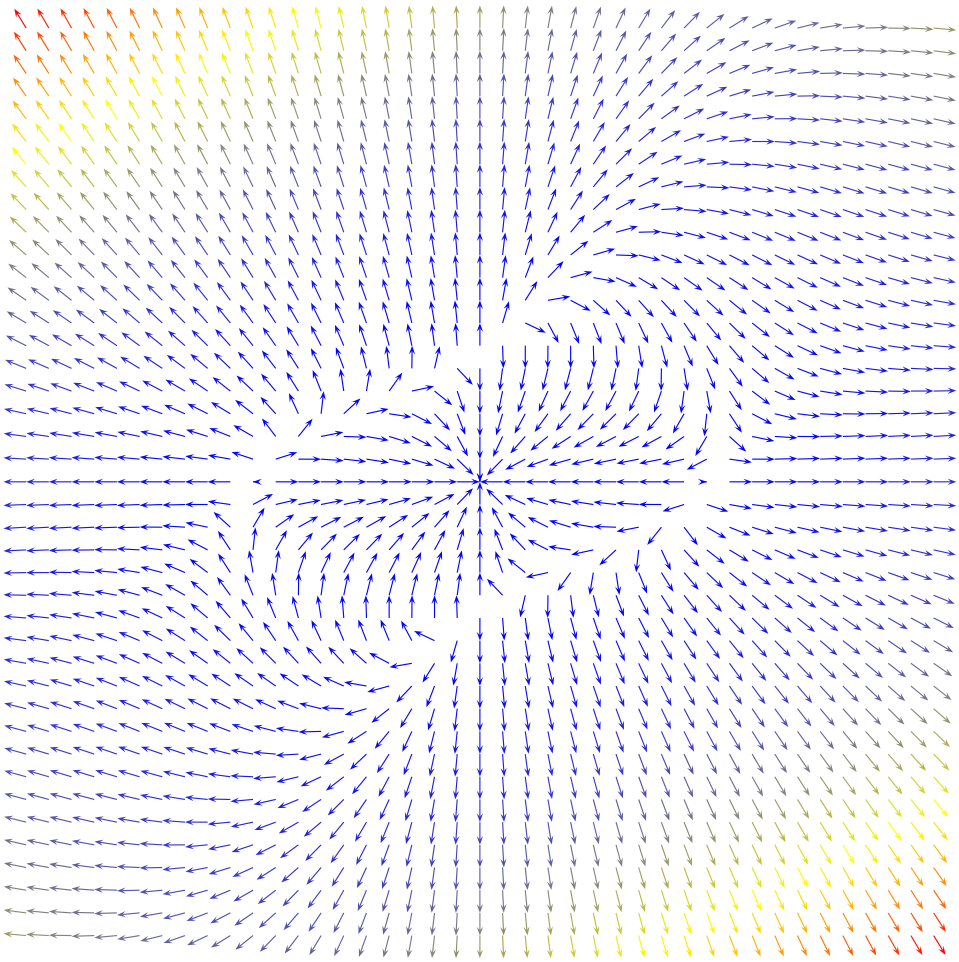


FIG. 4.7 – Champ de vecteurs associé à l'équation différentielle 4.46.

La figure suivante montre le résultat du calcul du bassin d'attraction du point x_∞ de coordonnées $(0, 0)$.

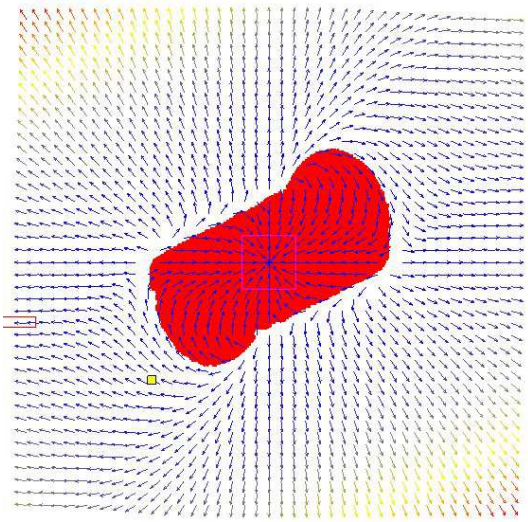


FIG. 4.8 – Bassin d'attraction du point de coordonnées $(0, 0)$.

4.5 Conclusion

Dans ce chapitre, on a proposé une méthode pour approcher le bassin d'attraction d'un point x_∞ asymptotiquement stable. La première contribution est la création effective, à l'aide de la théorie de Lyapunov, d'un voisinage de x_∞ inclus dans le bassin d'attraction de ce point. Puis, les méthodes d'inclusion différentielle nous permettent de créer un système dynamique discret non déterministe (codé par un graphe orienté). Finalement, l'étude de ce graphe nous permet de calculer le bassin d'attraction de x_∞.

Chapitre 5

Conclusion et perspectives

La plupart des problèmes de planification de trajectoires peuvent être modélisés avec des ensembles semi-algébriques, et par conséquent ils existent des méthodes effectives permettant de répondre aux problèmes de planification de trajectoires. Ils arrivent que la modélisation d'un problème de robotique ne conduisent pas naturellement à cette modélisation algébrique. Dans ce cas, le roboticien doit faire un changement de variables pour mettre le problème sous forme algébrique. Cette contrainte de modélisation peut être agaçante voire très compliquée. Dans cette thèse, on propose des algorithmes qui tentent de prendre en entrée des fonctions qui ne pas nécessairement polynomiales. Comme on oppose classiquement les algorithmes numériques aux algorithmes algébriques (ou encore algorithmes du calcul formel) avec comme frontière : les algorithmes numériques fournissent habituellement des valeurs approchées alors que les algorithmes algébriques donnent des résultats garantis.

Contrairement à cette idée, ici, on emploie des algorithmes numériques qui sont garantis. On s'appuit sur un outil qui permet de certifier les calculs numériques : le calcul par intervalles. C'est un outil fort intéressant qui a déjà montré sa puissance. On peut par exemple penser aux méthodes d'optimisations globales qui ont permis en outre de montrer la conjecture de Kepler.

La contribution principale de la thèse est de montrer comment la combinaison du calcul par intervalles et de la théorie des graphes permet la mise en oeuvre de nouveaux algorithmes. Ici, le calcul par intervalles est utilisé pour extraire des propriétés plus "locales", alors que les graphes sont employés pour compiler toutes ces informations et ainsi avoir une vue "globale".

Comme indiqué précédemment, ces algorithmes dévellopés sont garantis : les résultats fournis par ces méthodes ne sont pas discutables. A titre d'exemple, le nombre de composantes connexes calculé dans le chapitre 3 est donné avec certitude. Lorsque l'on planifie une chemin entre une position initiale et une position finale, nous avons la preuve que le robot ne va toucher aucun obstacle. De même, lorsque l'on montre qu'un point d'équilibre est asymptotiquement stable, la preuve fournie a autant de valeurs que n'importe quelle preuve.

Ces méthodes ont néanmoins des inconvénients, les algorithmes ne terminent pas toujours et les temps de calculs sont difficiles à évaluer. Ceci peut s'expliquer par la richesse des fonctions en entrée de ces méthodes (en général de classe $\mathcal{C}^\infty$). Cette variété rend une étude de la complexité quasi-impossible.

Lorsque l'on est en face d'algorithmes qui ne terminent pas pour toutes les entrées, l'une des questions naturelles que l'on peut se poser est : *Quelle est la taille de l'ensemble des entrées pour lesquelles cet algorithme ne termine pas ?*

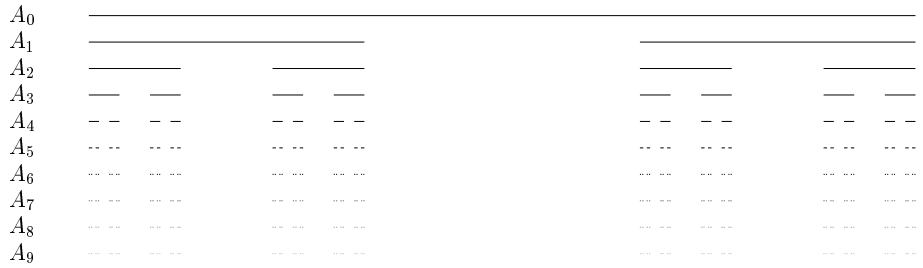
En conclusion, il me semble que les algorithmes présentés dans cette thèse terminent pour une partie résiduelle de l'ensemble des fonctions $\mathcal{C}^\infty$.
Mais ceci reste à prouver.

Annexe A

Cette annexe a pour but de montrer qu'il est relativement facile de créer des ensembles de $\mathbb{R}$ dont la géométrie est particulièrement délicate. On montrera aussi que ce genre d'ensembles peut être obtenu comme les zéros d'une certaine fonction de classe $\mathcal{C}^\infty$. Ce qui laisse penser qu'il est inimaginable de créer une méthode effective capable de classifier topologiquement tous les sous-ensembles de $\mathbb{R}^n$ obtenu comme 0 de fonctions $\mathcal{C}^\infty$.

L'ensemble de Cantor (ou *ensemble triadique de Cantor, ou poussière de Cantor*) est un sous-ensemble remarquable de la droite réelle construit par le mathématicien allemand Georg Cantor.

On le construit de manière itérative à partir du segment [0,1], en enlevant le tiers central ; puis on réitère l'opération sur les deux segments restants, et ainsi de suite. On peut voir les neuf premières itérations du procédé sur le schéma suivant :



Construction itérative On dénote par $\mathcal{T}$ l'opérateur "enlever le tiers central".
$\mathcal{T} : I \to I_0 \cup I_1, [a,b] \mapsto [a, a + \frac{b-a}{3}] \cup [b - \frac{b-a}{3}, b]$

On note $A_0 = [0,1]$ et on définit par récurrence une suite de parties de $[0,1]$ par la relation : $\forall n \in \mathbb{N}, A_{n+1} = \mathcal{T}(A_n)$.

On a : $A_1 = [0, \frac{1}{3}] \cup [\frac{2}{3}, 1]$.
$A_2 = [0, \frac{1}{9}] \cup [\frac{2}{9}, \frac{1}{3}] \cup [\frac{2}{3}, \frac{7}{9}] \cup [\frac{8}{9}, 1]$.
$A_3 = [0, \frac{1}{27}] \cup [\frac{2}{27}, \frac{1}{9}] \cup [\frac{2}{9}, \frac{7}{27}] \cup [\frac{8}{27}, \frac{1}{3}] \cup [\frac{2}{3}, \frac{19}{27}] \cup [\frac{20}{27}, \frac{7}{9}] \cup [\frac{8}{9}, \frac{25}{27}] \cup [\frac{26}{27}, 1]$.

Alors l'ensemble de Cantor K est "la limite" de A_n quand n tend vers $+\infty$:

$$K = \bigcap_{n\in\mathbb{N}} A_n. \tag{5.1}$$

Il s'agit d'un ensemble fermé de $[0,1]$, d'intérieur vide. Il sert d'exemple pour montrer qu'il existe des ensembles non dénombrables mais négligeables au sens de la mesure de Lebesgue. Il admet enfin une interprétation en terme de développement des réels en base 3. Pour cette raison, il est souvent noté K_3.

Écriture en base 3 On peut aussi définir l'ensemble de Cantor via l'écriture en base 3 : tout réel $x \in [0,1]$ s'écrit de manière : $x = \sum_{n=1}^{\infty} \frac{x_n}{3^n}$ avec $x_n \in \{0,1,2\}$

On écrit alors $x = 0,x_1x_2x_3x_4x_5\ldots$

Cette écriture est unique à ceci près : on peut remplacer $1000000\ldots$ par $0222222\ldots$ (et $2000000\ldots$ par $1222222\ldots$) à la fin d'une écriture, de la même manière que $0,999999\cdots = 1$ en base 10. L'ensemble de Cantor est formé des réels de $[0,1]$ ayant une écriture en base 3 ne contenant que des 0 et des 2.

C'est-à-dire

$$K_3 = \left\{ \sum_{n=1}^{\infty} \frac{x_n}{3^n}\ ,\ x_n \in \{0,2\} \right\} \tag{5.2}$$

Donc $1/3$ est dans cet ensemble, puisqu'il admet les deux écritures $0,1000\ldots$ et $0,02222\ldots$ en base 3. $2/3$ également ($0,2000\ldots$ ou $0,12222\ldots$). Parmi les nombres admettant un développement propre et un développement impropre, il n'en existe aucun dont les deux écritures vérifient la propriété demandée.

Propriétés L'ensemble de Cantor a de nombreuses propriétés particulières.

1. L'ensemble de Cantor est de mesure nulle, c'est-à-dire négligeable au sens de la mesure de Lebesgue.

   En effet en notant m la mesure de Lebesgue sur $\mathbb{R}$, on a :

   – $m(([0,1]) = 1$,

   – pour une réunion A_n d'intervalles : $m\left(\mathcal{T}(A_n)\right) = m(A_{n+1}) = \frac{2}{3}m(A_n)$.

   On en déduit que pour les étapes de la construction itérative ci-dessus : $\forall n \in \mathbb{N}$, $l\left(A_n\right) = \left(\frac{2}{3}\right)^n$

   Et comme l'ensemble de Cantor est inclus dans tous les A_n : $m\left(K\right) = 0$. L'ensemble de Cantor est donc "petit" au sens de la mesure de Lebesgue.

2. Cependant l'ensemble de Cantor n'est pas dénombrable ; il a la puissance du continu. En effet on peut montrer que les ensembles K_3 et $[0,1]$ sont équipotents.

   Pour cela on associe à tout élément $x = 0,x_1x_2x_3x_4\cdots \in K_3$ écrit en base 3, l'élément $f(x) = 0,x'_1x'_2x'_3x'_4\cdots \in [0,1]$ écrit en base 2, avec :
   - $x'_i = 0$ si $x_i = 0$,
   - $x'_i = 1$ si $x_i = 2$

   Par exemple l'élément $0,0202200222000\ldots$ de l'ensemble de Cantor correspondra à l'élément $0,0101100111000\ldots$ du segment unité $[0,1]$.

   Il est facile de voir que cette application est surjective mais non injective (l'élément $0,1$ étant l'image de $0,0222222\ldots$ comme de $0,2$). De l'existence d'une surjection de K_3 dans $[0,1]$ et en admettant l'axiome du choix, on déduit l'existence d'une injection de $[0,1]$ dans K_3, et comme l'application identité induit clairement une injection de K_3 dans $[0,1]$, alors d'après le théorème de Cantor-Bernstein, on en déduit que K_3 et $[0,1]$ sont équipotents. Donc l'ensemble de Cantor a la puissance du continu.

   Ainsi l'ensemble de Cantor est "grand" au sens de la théorie des ensembles.

3. Propriétés topologiques :

   (a) L'ensemble de Cantor est compact, et n'a que des points d'accumulation. On dit que c'est un ensemble parfait. Par ailleurs, il est d'intérieur vide,

   Démonstration : soit P un point de K_3, et soit une boule ouverte (intervalle ouvert) centrée en P. Cet ouvert contient nécessairement un réel dont le développement en base 3 contient le chiffre 1, qui n'est pas élément de K_3. Donc P n'est pas intérieur à K_3. Par ailleurs, dans ce même intervalle, il existe toujours un réel dont le développement en base 3 s'écrit uniquement avec des 0 ou des 2. Donc P n'est pas un point isolé.

   (b) L'ensemble de Cantor est également totalement discontinu c'est-à-dire que chaque singleton est sa propre composante connexe, et homéomorphe à l'espace topologique $\{0,1\}^{\mathbb{N}}$.

   (c) L'ensemble de Cantor est "universel dans la catégorie des espaces métriques", autrement dit tout espace métrique est l'image de l'ensemble de Cantor par une application continue.

Annexe B

Optimisation

Soit J une fonction $J : V \to \mathbb{R}$ avec V une partie de $\mathbb{R}^n$, J est souvent appelée fonction coût. Dans cette partie, nous nous intéressons au problème suivant :

$$\text{Trouver } \{u \in V, \text{ tel que } J(u) = \inf_{v \in V} J(v)\} \tag{5.3}$$

Le fait d'avoir à notre disposition l'image directe d'un intervalle via une fonction f ouvre aussi la possibilité de faire de l'*optimisation globale*. Ce terme d'optimisation globale vient en opposition avec les autres méthodes itératives dites locales. (Méthode de relaxation, méthode de gradient à pas optimal, méthode du gradient conjugué ...) Les propriétés de ces méthodes locales ont été largement étudiées. La plus intéressantes des propriétés étant sans doute la méthode du gradient conjugué qui converge en n itérations vers le minimum dans le cas d'une fonction coût elliptique. Quand la fonction coût n'a pas les bonnes propriétés (convexité, ellipticité ...), alors les résultats obtenus avec ces méthodes dépendent cruellement du point de départ de ces algorithmes. La plupart du temps, elles ne convergent que vers un minimum local.

L'algorithme classique d'optimisation globale, du calcul par intervalle, est un algorithme de type *branch and bound* (la méthode générale de branch and bound a été proposé pour la première fois par A. H. Land et A. G. Doig. [27]). Afin de pouvoir garantir l'existence d'un minimum de la fonction J sur un intervalle, nous supposerons que J est continue. Nous noterons par J_D le minimum de J sur D.

L'idée principale de cet algorithme est d'exclure de notre domaine de recherche toutes régions qui ne contiennent que des valeurs supérieures à J_D (amélioré au

cours de l'algorithme). Ici, on présente cette méthode dans le cas où f est définie sur une partie de $\mathbb{R}$:

Alg. 12 Algorithme d'optimisation globale - *branch and bound*

Entrée: D un compact de $\mathbb{R}$. $J : D \to \mathbb{R}$.

Entrée: Une précision ϵ.

Sortie: Une famille finies d'intervalles $\mathcal{P} = \{[x_j]\}_j$ telle que

$$\{u \in V, \text{ tel que } J(u) = \inf_{v \in V} J(v)\} \subset \bigcup [x_j], \text{ et } \forall j, m([x_j]) < \epsilon \tag{5.4}$$

1: *Initialisation* : $\mathcal{P} \leftarrow \emptyset$,
2: *Initialisation* : $\mathcal{P}_\Delta \leftarrow \{[x_i]\}_i$, avec $D = \bigcup_{1 \leq i \leq n} [x_i]$
3: *Initialisation* : $J_D \leftarrow \inf J(D)$.
4: **tant que** $\mathcal{P}_\Delta \neq \emptyset$ **faire**
5: $[x_t] \leftarrow [x]$ où $[x] \in \mathcal{P}_\Delta$.
6: $\mathcal{P}_\Delta \leftarrow \mathcal{P}_\Delta - \{[x]\}$.
7: **si** $\inf J([x_t]) = J_D$ **alors**
8: **si** $m([x_t]) < \epsilon$ **alors**
9: $\mathcal{P} \leftarrow \mathcal{P} \cup \{[x_t]\}$
10: **sinon**
11: $[x_{t_1}] \leftarrow [\ \underline{x_t}\ ;\ 0.5 \cdot (\underline{x_t} + \overline{x_t})\ ]$. et $[x_{t_2}] \leftarrow [\ 0.5 \cdot (\underline{x_t} + \overline{x_t})\ ;\ \overline{x_t}\ ]$.
12: $\mathcal{P}_\Delta \leftarrow \mathcal{P}_\Delta \cup \{[x_{t_1}]\} \cup \{[x_{t_2}]\}$
13: **fin si**
14: **fin si**
15: **fin tant que**

Annexe C

Dans cette annexe, on regroupe les preuves de l'étude de l'erreur commise lors de la multiplication et l'addition de réels en utilisant les flottants.

Proposition 5.0.1 (Erreur lors de l'approximation d'un réel à l'aide d'un flottant) *L'approximation d'un nombre réel x par un flottant de $F_{n,q}$ est à précision relative ;*

$$\frac{\Delta x}{|x|} \leq 10^{1-n}. \tag{5.5}$$

Preuve : Si l'on note par x' le flottant le plus proche de x, alors on a $\Delta x = |x - x'|$. On peut toujours écrire x et x' sous la forme $x = m \cdot 10^p$, $x' = m' \cdot 10^p$ avec m et m' :

$$\begin{aligned} m &= \pm 0, a_0 a_1 \ldots a_n a_{n+1} \ldots \\ m' &= \pm 0, a_0 a_1 \ldots a_n \end{aligned}$$

et donc

$$\begin{aligned} \frac{\Delta x}{|x|} &\leq \frac{|x - x'|}{|x|} \\ &\leq \frac{|m \cdot 10^p - m' \cdot 10^p|}{m \cdot 10^p} \\ &\leq \frac{|m - m'|}{m} \end{aligned}$$

or $|m - m'| \leq 10^{-n}$ et $m \geq 10^{-1}$ et donc :

$$\frac{\Delta x}{|x|} \leq 10^{1-n}. \tag{5.6}$$

□

On notera par $\epsilon = 10^{1-n}$. Dans la suite, on se propose de majorer l'erreur sur une somme et sur un produit lorsque l'on utilise des flottants.

Propriétés 5.0.2 (Erreur lors de la somme de flottants)
Soient x et y deux nombres réels flottants, avec $x +' y$ la somme flottante de x et de y, on note $\Delta(x+y)$ la distance qui sépare $x+y$ de $x +' y$, on a

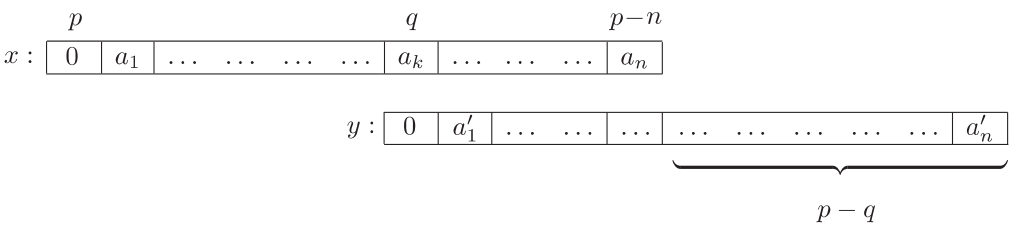
$$\Delta(x+y) \leq \epsilon(|x|+|y|). \tag{5.7}$$

Preuve : On a :

$$\begin{aligned} x &= \pm 0, a_0 a_1 \ldots a_n \cdot 10^p, \quad 10^{p-1} \leq x < 10^p \\ y &= \pm 0, a'_0 a'_1 \ldots a'_n \cdot 10^q, \quad 10^{q-1} \leq y < 10^q \end{aligned}$$

Supposons par exemple que $p \geq q$. Nous allons distinguer deux situations :

- Supposons qu'il n'y ait pas de débordement, (i.e. $x+y < 10^p$), le calcul de $x+y$ s'accompagne de la perte des $p-q$ derniers chiffres de y correspondant aux puissances 10^l où $l \leq p-n$.



  Ainsi $\Delta(x+y) \leq 10^{p-n}$.
- En cas de débordement, $x+y \geq 10^p$, (ce qui se produit par exemple, si $p=q$ et si $a_1 + a'_1 \geq 10$), la décimale correspondant à la puissance 10^{p-n} est elle aussi perdue, d'où $\Delta(x+y) \leq 10^{p-n+1}$.

Dans les deux cas :

$$\Delta(x+y) \leq \epsilon(|x|+|y|). \tag{5.8}$$

□

En général, les réels x et y ne sont pas des flottants, ils sont approchés par des nombres x' et y'. Mais on peut tout de même majorer l'erreur commise en les additionnant avec des flottants.

Propriétés 5.0.3 (Erreur de la somme de deux réels en utilisant des flottants)
Si x et y sont des réels, l'erreur commise en les additionnant avec des flottants est donnée par

$$\Delta(x+y) \leq \epsilon(\epsilon+2)(|x|+|y|) \tag{5.9}$$

Preuve : On a $x = x' + e_x, y = y' + e_y$ avec $|e_x| \leq \Delta x$ et $|e_y| \leq \Delta y$. On note par $x +' y$ le flottant obtenu en additionnant x et y, et $e_{x'+y'}$ l'erreur commise lors de cette addition. On a :

$$\begin{aligned} x + y &= x' + y' + e_x + e_y \\ &= (x' +' y') + e_{x'+y'} + e_x + e_y \end{aligned}$$

donc :

$$\begin{aligned} (x + y) - (x' +' y') &= e_{x'+y'} + e_x + e_y \\ |(x + y) - (x' +' y')| &\leq \underbrace{|e_{x'+y'}|}_{\leq \Delta(x'+y')} + \underbrace{|e_x|}_{\leq \Delta x} + \underbrace{|e_y|}_{\leq \Delta y} \\ &\leq \Delta(x' + y') + \Delta x + \Delta y \end{aligned}$$

on en déduit :

$$\begin{aligned} \Delta(x + y) &\leq \epsilon(|x'| + |y'|) + \Delta x + \Delta y \\ &\leq \epsilon(|x| + \Delta x + |y| + \Delta y) + \Delta x + \Delta y \\ &\leq \epsilon(|x| + \epsilon|x| + |y| + \epsilon|y|) + \epsilon|x| + \epsilon|y| \\ &\leq \epsilon(\epsilon + 2)(|x| + |y|) \end{aligned}$$

□

Propriétés 5.0.4 (Erreur lors du produit de deux flottants)
Soient x et y deux nombres réels flottants avec une mantisse à n chiffres,

$$\Delta(x \times y) \leq \epsilon|x||y|. \tag{5.10}$$

Propriétés 5.0.5 (Erreur lors du produit de deux réels en utilisant des flottants)
Soient x et y deux nombres réels avec une mantisse à n chiffres,

$$\Delta(x \times y) \leq \epsilon(\epsilon^2 + 5\epsilon + 3)|x||y|. \tag{5.11}$$

Preuve : On note par x' et y' deux flottants les plus proches de x et de y. On a $x = x' + e_x, y = y' + e_y$ avec $|e_x| \leq \Delta x$ et $|e_y| \leq \Delta y$. On en déduit :

$$\begin{aligned} x \times y &= (x' + e_x) \times (y' + e_y) \\ &= x' \times y' + x' \times e_y + e_x \times y' + e_x \times e_y \\ &= x' \times' y' + \epsilon|x'y'| + x' \times e_y + e_x \times y' + e_x \times e_y. \end{aligned}$$

On en déduit :

$$\begin{aligned} |x \times y - x' \times' y'| &= \epsilon|x'y'| + x' \times e_y + e_x \times y' + e_x \times e_y \\ |x \times y - x' \times' y'| &\leq \epsilon(|x| + \Delta x)(|y| + \Delta y) + (|x| + \Delta x)\Delta y + \Delta x(|y| + \Delta y) + \Delta x \Delta y \end{aligned}$$

d'où :

$$\begin{aligned} \Delta(x \times y) &\leq (1 + \epsilon)(|x|\Delta y + |y|\Delta x) + \epsilon|x||y| + (\epsilon + 3)\Delta x \Delta y \\ &\leq (1 + \epsilon)(|x|\epsilon|y| + |y|\epsilon|x|) + \epsilon|x||y| + (\epsilon + 3)\epsilon|x|\epsilon|y| \\ &\leq \epsilon(\epsilon^2 + 5\epsilon + 3)|x||y|. \end{aligned}$$

□

Bibliographie

[1] Volker Stahl, *Interval methods for bounding the range of polynomials and solving systems of nonlinear equations*, Ph.D. thesis.

[2] Barton T. Stander, John C. Hart *Guaranteeing the Topology of an Implicit Surface Polygonization for Interactive Modeling*

[3] John Willard Milnor *Morse Theory* Princeton Univ Pr, 1963, isbn 0-691-08008-9

[4] Roberto Genesio, Michele Tartaglia, Antonio Vivino, *On the estimation of the asymptotic stability regions : state of the art and new proposals* IEEE Transaction On Automatic Control, Vol AC-30, NO. 8, 1985.

[5] Tucker, W. *A Rigorous ODE Solver and Smale's 14th Problem.* Found. Comput. Math. 2, 53-117, 2002.

[6] S. Nedialkov *Ph.D. Thesis : Computing Rigorous Bounds on the Solution of an Initial Value Problem for an Ordinary Differential Equation* Computer Science Dept., Univ. of Toronto, 1999.

[7] J. P. Demailly *Analyse numérique et équations différentielles, Manuel pour le Second Cycle de Mathématiques*, Presses Universitaires de Grenoble, Septembre 1991, 309 pages

[8] E. Hansen *Global Optimization using Interval Analysis.* Marcal Dekker, Inc., 1992.

[9] A. Bjorner. *Topological methods. In Handbook of combinatorics, Vol. 1, 2,* pages 1819-1872. Elsevier, Amsterdam, 1995

[10] Herbert Edelsbrunner, Nimish R. Shha, *Triangulating topological spaces* Source Annual Symposium on Computational Geometry archive Proceedings of the tenth annual symposium on Computational geometry table of contents Stony Brook, New York, United States Publication : 1994 0-89791-648-4

[11] *Planning Algorithms* Cambridge University Press, 2006

[12] Michel Demazure *Bifurcations and Catastrophes : Geometry of Solutions to Nonlinear Problems* Springer, janvier 2000, ISBN : 3540521186

[13] J.P. Dedieu *Points fixes, Zéros et la Méthode de Newton* Springer-Verlag, mars 2006, ASIN : 3540309950

[14] Delanoue, N., Jaulin,L., Cottenceau,B. Using interval arithmetic to prove that a set is path-connected. *Theoretical computer science, Special issue : Real Numbers and Computers.*, 2004.

[15] L. Jaulin. Path planning using intervals and graphs. *Reliable Computing, issue 1, volume 7*, 2001

[16] Khatib, O. Real-time obstacle avoidance for manipulators and mobile robots. *International Journal Of Robotics Research*, 1986.

[17] Janich, K. *Topology (Undergraduate Texts in Mathematics)* Springer Verlag

[18] Jaulin, L. and Walter, E., Set inversion via interval analysis for nonlinear bounded- error estimation, *Automatica*, 29(4), 1993, 1053-1064.

[19] Dijkstra, E.W.,. A note on two problems in connection with graphs, *Numerische Math*, 1, 1959, 269-271.

[20] R. E. Moore, 1979, *Methods and Applications of Interval Analysis* SIAM, Philadelphia, PA

[21] F. Rouillier, M.-F. Roy, M. Safey. Finding at least one point in each connected component of a real algebraic set defined by a single equation, *Journal of Complexity* 16 716-750 (2000)

[22] S. Basu, R. Pollackz, M.-F. Roy. *Computing the first Betti number and the connected components of semi-algebraic sets,*

[23] T. Lozano-Pérez, Spatial Planning : A Confi uration Space Approach. 1983, IEEETC

[24] R. E. Moore, *Interval Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1996

[25] R. E. Moore, Practical aspects of interval computation, *Appl. Math.*, 13, pages 52-92

[26] Schwartz and Sharir, *On the Piano movers Problem II : General techniques for computing topological properties of algebraic manifolds,* Advances in Applied Mathematics, vol 4, pp. 298-351, 1983.

[27] A. H. Land and A. G. Doig, *An Automatic Method for Solving Discrete Programming Problems* Econometrica, Vol.28 (1960), pp. 497-520.

[28] B. Chazelle, Approximation and decomposition of shapes, *In J.T Schwaertz and C.-K Yap, editors, Advances in Robotics 1 : Algorithmic and Geometric Aspect of Robotics*, pages 145-185. Lawrence Erlbaum Assciates, Hillsdale, NJ, 1987.

[29] G.E. Collins
Quantifier Elimination for Real Closed Fields by Cylindrical Algebraic Decomposition.
In *Lecture Notes In Computer Science*, volume Vol. 33, pages 134-183. Springer-Verlag, Berlin, 1975.

[30] A. Neumaier *Interval Methods for Systems of Equations* Encyclopedia of Mathematics and its Applications 37, Cambridge Univ. Press, Cambridge 1990 255 pp.

[31] G. Alefeld, *Inclusion methods for systems of nonlinear equations - the interval Newton method and modifications.*, In "Topics in Validated Computation". Proceedings of the IMACS-GAMM International Workshop on Validated Computation, Oldenburg, Germany, August 30 - September 3, 7-26, 1993. (Editor : J. Herzberger. Elsevier Amsterdam 1994).

[32] J.H. Davenport, Y.Siret, Eournier, *Computer Algebra, Systems and Algorithms for Algebraic Computation*, Academic Pr ; 2nd edition (June 1993), ISBN : 0-12204-232-8.

[33] Jiri Rohn, Positive Definiteness and Stability of Interval Matrices, *SIAM Journal on Matrix Analysis and Applications*, Volume 15, Number 1, pp. 175-184, © 1994 Society for Industrial and Applied Mathematics,

[34] Reinhard Diestel, *Graph Theory (Graduate Texts in Mathematics, 173)* 2000, Springer Verlag, 0-387-98976-5.

[35] 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.

[36] B. A. Davey, H. A. Priestley, *Introduction to Lattices and Order* 2002, Cambridge University Press, 0-521-78451-4

[37] G. Birkhoff, *Lattice theory* American Mathematical Society Colloquim Publications, 1940, XXV, Providence, Rhode Island

[38] T. Raïssi, N. Ramdani and Y. Candau. Set membership state and parameter estimation for systems described by nonlinear differential equation. *Automatica*, 40 : 1771-1777,2004.

[39] Slotine, J.J.E., and Li, W., *Applied Nonlinear Control,* Prentice-Hall, 1991.

[40] *Jaulin L., M. Kieffer, O. Didrit and E. Walter Applied Interval Analysis with Examples in Parameter and State Estimation, Robust Control and Robotics,* Springer-Verlag, ISBN : 1-85233-219-0, (2001).

[41] Revol N. Interval Newton iteration in multiple precision for the univariate case, *Numerical Algorithms*, vol 34, no 2, pp 417–426, 2003

[42] Lingas A. The power of non-rectilinear holes. *Proceedings of the 9th International Colloquium on Automata, Languages and Programmings, vol 140 of Lecture Notes in Computer Sciences*, pages 369-383, 1982

[43] L. Jaulin, and D. Henrion, Contracting optimally an interval matrix without loosing any positive semi-definite matrix is a tractable problem, *Reliable Computing, Volume 11, issue 1, pages 1-17.* (2005)

[44] Kharitonov V.L., About an asymptotic stability of the equilibrium position of linear differential equations systems family *Diffrential equations.* 1978. 14. N 11. pp.2086-2088 (in Russian).

[45] Bialas S., A necessary and sufficient condition for stability of interval matrices *Int. J. Contr.* 1983.

[46] Karl W.C., Greschak J.P., Verghese G.C., Comments on a necessary and sufficient condition for stability of interval matrices *Int. J. Contr.* 1984.

[47] Kreinovich V., Lakeyev A., Rohn J., Kahl P., *Computational complexity and feasibility of data processing and interval computations.* Kluwer, Dordrecht, 1997.

[48] John Willard Milnor.
C.R.F. Maunder.
Algebraic topology. London : Van Nostrand Reinhold Co., 1970.

[49] Henri Poincare, *New Methods of Celestial Mechanics, Volume 1-3,* American Institute of Physics, 1993.

[50] V. Lakshmikantham, V. M. Matrosov, and S. Sivasundaram, *Vector Lyapunov Functions and Stability Analysis of Nonlinear Systems,* Kluwer, 1991.

[51] Richard M. Karp and Robert Endre Tarjan, Linear expected-time algorithms for connectivity problems (Extended Abstract), *Proceedings of the twelfth an-*

nual ACM symposium on Theory of computing, 1980 isbn 0-89791-017-6 pages = 368–377, Los Angeles, California, United States, ACM Press,

[52] Ortolf H.-J. *Eine Verallgemeinerung der Intervallarithmetik.* Geselschaft fuer Mathematik und Datenverarbeitung, Bonn, 11 :1-71, 1969

[53] Kaucher E. *Uber metrische und algebraische Eigenschaften einiger beim numerischen Rechnen auftretender Raume.* PhD thesis, Karlsruhe, 1973

[54] Kaucher E. *Interval Analysis in the Extended Interval Space* $\mathbb{R}^n$. Computing, Suppl. 2 :33-49, 1980.

[55] Shary S.P. *A new technique in systems analysis under interval uncertainty and ambiguity.* Reliable computing, 8 :321418, 2002.