MATHÉMATIQUES
&
APPLICATIONS

Directeurs de la collection:
G. Allaire et M. Benaïm

54

MATHÉMATIQUES & APPLICATIONS

Comité de Lecture / Editorial Board

Directeurs de la collection:
G. ALLAIRE et M. BENAÏM

Instructions aux auteurs:

Les textes ou projets peuvent être soumis directement à l'un des membres du comité de lecture avec copie à G. ALLAIRE OU M. BENAÏM. Les manuscrits devront être remis à l'Éditeur sous format LaTeX 2e.

Jean-Pierre Dedieu

Points fixes, zéros et la méthode de Newton

Springer

Jean-Pierre Dedieu
MIP. Département de Mathématiques
Université Paul Sabatier
118 route de Narbonne
31062 Toulouse Cedex 9
France
e-mail : jean-pierre.dedieu@math.ups-tlse.fr

Library of Congress Control Number: 2005938218

Mathematics Subject Classification (2000): 37Cxx, 49Mxx, 58Cxx, 65Hxx

ISSN 1154-483X
ISBN-10 3-540-30995-0 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-30995-6 Springer Berlin Heidelberg New York

Springer est membre du Springer Science+Business Media

springer.com
Imprimé en Pays-Bas

Imprimé sur papier non acide 3141/SPI Publisher Services - 5 4 3 2 1 0 -

à Dany Serrato, in memoriam

Preface

The advent of the computer age has had an enormous impact on science and conversely science also has had great importance on the development of the computer. I believe that agreement on this statement is almost universal.

This beautiful book by Dedieu studies this relationship when the computational side is represented by "scientific computation" in the broad sense and the science is represented by mathematics. In this realm the historical roots lie in the times of Kepler and Newton, when some of the earliest computations played a role in the revolution in physics, "classical mechanics". This was the era when Newton's method established itself. That algorithm has now assumed a central place in numerical analysis.

The computer has contributed a new dimension to this picture. It is the contribution of the computer scientists in the last half of the 20th century which has given us a fundamentally new way of looking at computation. "What are the best algorithms ?" "When do they terminate ?" "How well do they perform ?" "How is that perfomance to be measured ?" Eventually such questions lead to what might be called the foundations of computer science. Finally one reaches the most central problem of all. Which algorithms possess the measure of efficiency called "polynomial time" ?

The study of these questions, called complexity theory, has been undertaken in the setting of discrete mathematics, with the 0's and 1's of Turing machines. However very recently, a new element has entered into complexity theory. The influence of computer scientists has become felt in the domain of real number mathematics where continuity and calculus play a dominant role. The old algorithms as Newton's method, the old problems as finding approximate zeros of polynomials, are being considered from the point of view of complexity and efficiency, and the need for new foundations is being realized.

What is the place of Dedieu's book in this picture ? Here we have an introduction to the mathematics sufficient to enter into the world of complexity of real number algorithms. Its study of Newton's method is deep, with its inclusion of both the extension to overdetermined systems and underdetermined

systems. With the simple, direct and elegant treatment found here, with the various examples, one sees the confirmation of the central importance of Newton's method in non-linear algorithmic mathematics.

Chicago, Steve Smale
june 2003

Table des matières

1

Introduction

Ce livre est consacré aux calculs de point fixes, de zéros de systèmes de fonctions et à la méthode de Newton. Il trouve son origine dans un cours professé en maîtrise ayant pour thème la résolution des systèmes d'équations non linéaires. Ce qui devait être une simple rédaction de notes de cours adressée aux étudiants s'en est finalement bien écarté pour devenir au fil des mois un ensemble plus étoffé présentant à la fois des résultats classiques sur les méthodes itératives, des points de vue plus modernes sur les systèmes dynamiques discrets et des travaux récents sur la méthode de Newton.

La première partie de ce texte est consacrée aux points fixes. Nous présentons un théorème d'existence pour une application contractante puis nous décrivons des théorèmes de classification de points fixes pour des applications différentiables définies sur des espaces de Banach. Ces points fixes sont classés en trois catégories : attractifs, répulsifs et hyperboliques. Nous montrons qu'un point fixe est attractif si le spectre de la dérivée en ce point est contenu dans l'ensemble des nombres complexes de module plus petit que 1. En fait ces deux énoncés sont équivalents.

Des résultats similaires ont lieu pour les points fixes répulsifs.

Les points fixes hyperboliques constituent une catégorie qui englobe les deux premières (attractifs et répulsifs) et pour laquelle on a une « bonne » théorie de la linéarisation. C'est le théorème de Grobman-Hartman. Il permet de passer, via un changement de variable bicontinu, de l'application à sa dérivée c'est à dire du non linéaire au linéaire.

Nous décrivons ensuite, dans le « Théorème de la variété stable locale » une décomposition de l'espace, au voisinage d'un point fixe hyperbolique, en deux sous-variétés transverses, les variétés stables et instables. L'application considérée laisse ces sous-variétés invariantes et agit sur l'une en contraction et sur l'autre en dilatation.

La plupart des résultats de cette première partie est présenté dans le cadre des espaces de Banach réels.

Nous donnons ensuite plusieurs exemples d'applications. Les plus significatifs sont issus de l'algèbre linéaire : problème des valeurs propres, calculs

de sous-espaces invariants. Un cadre géométrique naturel pour l'étude de ces exemples est celui de variétés différentiables telles que l'espace projectif réel ou complexe, la variété des drapeaux ou bien la grassmannienne. Nous ne considérons ici que la structure topologique de ces espaces. Elle peut être décrite sans faire appel au formalisme lourd de la géométrie différentielle. Nous mettons alors en évidence que certains algorithmes (QR, LR, Choleski) ne sont, dans un cadre géométrique adéquat, rien d'autre que des avatars de la méthode des approximations successives. Cela facilite la compréhension que l'on a de ces algorithmes et fournit un cadre de pensée pour les analyser et en concevoir d'autres.

La seconde partie de ce texte est consacrée à la méthode de Newton pour la résolution de systèmes d'équations non linéaires. De tels systèmes peuvent avoir autant d'équations que d'inconnues auquel cas, en général, leurs zéros sont des points isolés. Ils peuvent être sous-déterminés et décrivent alors, dans les cas considérés ici, des sous-variétés différentiables de l'espace ambiant, il peuvent enfin être sur-déterminés, donc génériquement sans racines, et dans ce cas on en cherche des solutions au sens des moindres carrés. La méthode de Newton agit dans ces trois cas fondamentaux. Elle est un outil classique et bien étudié dans le premier (Kantorovich, Ostrowski, Smale), classique quoique moins étudiée dans le troisième (méthode de Newton-Gauss), peu connue et encore peu utilisée dans le second.

Pour les systèmes relevant du premier cas (autant d'équations que d'inconnues), nous présentons en premier lieu la théorie de Kantorovich qui précise les propriétés de convergence quadratique de la méthode de Newton pour des fonctions de classe C^2. Nous passons ensuite à la théorie alpha de Smale qui est apparue très récemment, au cours des années 1980-1990. Le cadre de travail est celui des fonctions analytiques au lieu des fonctions de classe C^2. Les propriétés de convergence de la suite de Newton sont obtenues à partir du comportement du système au point initial de la suite au lieu d'une boule centrée en ce point comme c'est le cas dans le cadre C^2. Il y a là comme un effet de bascule : plus le problème est régulier et moins les hypothèses sont fortes...

Nous considérons ensuite le cas de systèmes sous-déterminés, c'est à dire dont le nombre d'inconnues est plus grand que celui des équations. Comme nous l'avons déjà dit, l'ensemble des zéros est, dans les cas considérés ici, une sous-variété différentiable. Nous montrons comment certaines caractéristiques géométriques de ces sous-variétés peuvent être décrites par l'invariant γ introduit par Shub et Smale dans leur série de papiers sur la complexité du Théorème de Bézout. Nous introduisons ensuite la méthode de Newton pour de tels systèmes et étudions ses propriétés de convergence du point de vue de la théorie alpha. Nous montrons qu'elle agit comme une projection sur cette sous-variété.

La dernière partie de ce texte a pour thème la méthode de Newton-Gauss pour des problèmes de type «moindre carrés non linéaires». Nous y présentons

des résultats de convergence « à la Kantorovich » et aussi le point de vue de la théorie alpha.

L'essentiel des résultats sur la méthode de Newton ont pour cadre les espaces de Banach lorsqu'il s'agit de systèmes « bien déterminés » et les espaces de Hilbert pour les systèmes sur-déterminés ou sous-déterminés. On utilise en effet l'inverse généralisé d'un opérateur linéaire et ce concept n'a de sens que dans un cadre hilbertien.

La lecture de ce texte suppose une bonne connaissance de l'algèbre linéaire telle qu'elle est enseignée dans les deux premières années de nos universités, de topologie générale et de calcul différentiel (niveau licence). Nous utilisons quelques outils d'analyse fonctionnelle et quelques rudiments de variable complexe comme la représentation locale d'une fonction analytique par sa série de Taylor. Afin de rendre ce livre aussi « auto-contenu » que possible, un appendice en fin d'ouvrage vient préciser les principaux résultats utilisés.

Ce texte s'adresse à des étudiants de maîtrise ou de troisième cycle ou ceux préparant l'agrégation de mathématique et bien sûr aux enseignants chercheurs. Le contenu des trois chapitres sur la méthode de Newton, qui présente la théorie alpha de Smale, n'est publiée, à ce jour, dans aucun autre ouvrage à l'exception d'une partie du chapitre 3 qui figure dans « Complexity and Real Computation » de Blum-Cucker-Shub-Smale.

Je remercie enfin Steve Smale qui a accepté d'écrire la préface de ce livre.

2

Points fixes

2.1 Introduction

Ce chapitre est consacré au calcul des zéros d'un système de fonctions $F(x) = 0$ ainsi qu'au calcul de points fixes $f(x) = x$. Comme une équation de point fixe peut s'écrire $f(x) - x = 0$ les deux points de vue sont équivalents dès lors que les espaces source et image sont identiques et que la soustraction existe. Mais ce n'est pas forcément le cas. Par exemple, l'itération de Rayleigh pour le calcul de vecteurs propres est une méthode de recherche de points fixes qui se déroule sur la sphère sur laquelle on ne dispose pas de structure vectorielle. L'autre aspect de non équivalence est relatif au fait qu'il faut parfois considérer des systèmes où le nombre d'équations et celui des inconnues ne sont pas nécessairement égaux : $F(x) = 0$ où $F = (F_1, \ldots, F_m) : \mathbb{R}^n \to \mathbb{R}^m$. Mais dans ce chapitre nous ferons l'économie de telles situations.

Les théorèmes les plus généraux sont formulés dans des espaces métriques complets et lorsqu'intervient le calcul différentiel nous nous situons dans le cadre des espaces de Banach.

Le premier des résultats que nous étudions est le « Théorème des Applications Contractantes ». Il fournit une foule de résultats d'existence mais aussi une méthode d'approximation et de calcul. L'idée de base est la suivante : partant d'un $x_0 \in \mathbb{E}$ on construit la suite des approximations successives $x_{k+1} = f(x_k)$. Si cette suite converge vers x et si f est continue en ce point alors $f(x) = x$: nous avons trouvé un point fixe x et nous disposons d'approximations de x, à savoir les points de la suite x_k. Dans le souci de tenir compte des calculs approchés, nous verrons aussi ce qu'il advient lorsqu'on remplace le schéma théorique $x_{k+1} = f(x_k)$ par un schéma perturbé.

Dans un second temps nous étudions la structure d'un point fixe x de f. Une classification est établie en fonction du portrait spectral de l'opérateur linéaire $Df(x)$. Pour cette raison nous commençons par étudier le cas des itérations définies par les automorphismes linéaires d'un espace vectoriel de dimension finie. Lorsque le spectre d'un tel opérateur ne rencontre pas le cercle unité, on obtient trois types de points fixes : attractifs, répulsifs et

hyperboliques. Nous établissons l'existence d'une décomposition de l'espace en somme directe de deux sous-espaces vectoriels invariants, l'opérateur étant une contraction sur l'un et une dilatation sur l'autre. Ces résultats sont étendus, toujours dans le cas linéaire, aux endomorphismes bornés d'un espace de Banach.

Dans le cas non linéaire, la classification des points fixes demeure, ainsi que l'existence des deux sous-espaces invariants : la restriction de f est une contraction sur l'un et une dilatation sur l'autre. Toutefois la situation est beaucoup plus compliquée : ces sous-espaces sont désormais des sous-variétés différentiables au lieu d'espaces linéaires. Ils sont appelés variété stable et variété instable. Pour aboutir à ce résultat nous passons par le théorème de Grobman-Hartman qui permet d'élucider la nature d'un point fixe x de f à partir de la dérivée $Df(x)$. La structure des variétés stable et instable est étudiée dans la dernière section de ce chapitre, c'est le « Théorème de la variété stable », dû à Perron.

2.2 Le théorème des applications contractantes

2.2.1 Enoncé du théorème

Notons $\mathbb{E}$ un espace métrique complet et d sa distance.

Définition 1. *Une application $f : \mathbb{E} \to \mathbb{E}$ est lipschitzienne s'il existe une constante $\lambda \geq 0$, appelée constante de Lipschitz, telle que pour tout x et $y \in \mathbb{E}$ on ait $d(f(x), f(y)) \leq \lambda d(x, y)$. Une application $f : \mathbb{E} \to \mathbb{E}$ est une contraction si elle est lipschitzienne pour une constante $\lambda < 1$. On dit aussi que f est contractante.*

La plus petite des constantes de Lipschitz est donnée par

$$\mathrm{Lip}(f) = \sup \frac{d(f(x), f(y))}{d(x, y)}$$

où le sup est pris pour tous les x et $y \in \mathbb{E}$, $x \neq y$.

Définition 2. *Une application $f : \mathbb{E} \to \mathbb{E}$ est dilatante s'il existe une constante $\Lambda > 1$ telle que, pour tout x et $y \in \mathbb{E}$, on ait $d(f(x), f(y)) \geq \Lambda d(x, y)$.*

Une application dilatante est injective. Une application bijective $f : \mathbb{E} \to \mathbb{E}$ est dilatante si et seulement si f^{-1} est contractante puisque

$$d(f(x), f(y)) \geq \Lambda d(x, y)$$

équivaut à

$$d(f^{-1}(x), f^{-1}(y)) \leq \Lambda^{-1} d(x, y),$$

la condition $\Lambda > 1$ devenant $\Lambda^{-1} < 1$.

Proposition 3. *Une application $f : \mathbb{E} \to \mathbb{E}$ qui est lipschitzienne est uniformément continue.*

Théorème 4. *(Théorème des approximations successives) Soit $f : \mathbb{E} \to \mathbb{E}$ une contraction de constante $0 \leq \lambda < 1$.*

1. *Pour tout $x_0 \in \mathbb{E}$ la suite $x_{k+1} = f(x_k)$ converge vers un point fixe $x \in \mathbb{E}$, autrement dit $f(x) = x$,*
2. *Ce point fixe est unique,*
3. *Pour tout $q \geq 0$, $d(x_q, x) \leq \dfrac{\lambda^q}{1-\lambda} d(x_0, x_1)$,*
4. $\dfrac{d(x_0, x_1)}{1+\lambda} \leq d(x_0, x) \leq \dfrac{d(x_0, x_1)}{1-\lambda}$.

Preuve On a

$$d(x_{k+1}, x_k) = d(f(x_k), f(x_{k-1})) \leq \lambda d(x_k, x_{k-1}) \leq \ldots \leq \lambda^k d(x_1, x_0)$$

de sorte que pour des entiers $p \geq q \geq 0$ l'inégalité triangulaire donne

$$\begin{aligned} d(x_p, x_q) &\leq \sum_{k=q}^{p-1} d(x_{k+1}, x_k) \leq \sum_{k=q}^{p-1} \lambda^k d(x_1, x_0) \\ &\leq \lambda^q d(x_1, x_0) \sum_{k=0}^{\infty} \lambda^k = \frac{\lambda^q}{1-\lambda} d(x_0, x_1). \end{aligned}$$

Ceci prouve que la suite (x_p) est de Cauchy et comme $\mathbb{E}$ est complet elle possède une limite x. Par continuité de f et passage à la limite, l'égalité $x_{k+1} = f(x_k)$ conduit à $x = f(x)$: x est un point fixe. Il ne peut y en avoir d'autre puisque f est contractante. En effet, si $f(x) = x$ et $f(y) = y$ alors

$$d(x, y) = d(f(x), f(y)) \leq \lambda d(x, y),$$

ce qui ne saurait se produire avec $d(x, y) \neq 0$ et $\lambda < 1$. L'inégalité $d(x_q, x) \leq \lambda^q d(x_0, x_1)/(1-\lambda)$ se prouve en passant à la limite pour $p \to \infty$ dans l'inégalité précédente. En particulier $d(x_0, x) \leq d(x_0, x_1)/(1-\lambda)$. Pour finir on note que

$$d(x_0, x_1) \leq d(x_0, x) + d(x, x_1) \leq d(x_0, x) + \lambda d(x_0, x) = (1+\lambda) d(x_0, x). \quad \square$$

Remarque 1. L'aspect complexité du calcul approché du point fixe x par cette méthode est aussi donné par le Théorème 4. On obtiendra une précision $\epsilon > 0$ sur le calcul du point fixe, c'est-à-dire $d(x_q, x) \leq \epsilon$, dès que

$$q \geq \frac{1}{\log \lambda} \log \left(\frac{\epsilon(1-\lambda)}{d(x_0, x_1)} \right).$$

Le résultat que l'on vient de présenter a ceci d'exceptionnel que l'on converge vers un unique point fixe x quel que soit le point initial x_0 choisi. C'est loin d'être une situation générale. Une application peut avoir plusieurs points fixes et les suites des approximations successives peuvent ne pas nécessairement converger.

2.2.2 Comment vérifier l'hypothèse de contraction ?

Changeons de cadre : au lieu d'un espace métrique, nous considérons ici un espace de Banach et nous utilisons le calcul différentiel. L'inégalité des accroissements finis 6.1.5 donne la clé du problème.

Proposition 5. *Soient $x_0 \in \mathbb{E}$ et $r > 0$. Prenons pour C la boule ouverte de centre x_0 et de rayon r. Soit $f : C \to \mathbb{E}$ une application différentiable dont la norme de la dérivée est bornée sur C : $\|Df(x)\| \leq \lambda < 1$. Si de plus*

$$\lambda r + \|x_0 - f(x_0)\| \leq r$$

alors f est une application contractante de C dans C de constante de contraction λ.

Preuve L'inégalité des accroissements finis 6.1.5 prouve que f est une application contractante de constante λ. Montrons que son image est contenue dans C. Pour tout $x \in C$ on a :

$$\begin{aligned}\|f(x) - x_0\| &\leq \|f(x) - f(x_0)\| + \|f(x_0) - x_0\| \leq \lambda\|x - x_0\| + \|f(x_0) - x_0\| \\ &< \lambda r + \|f(x_0) - x_0\| \leq r.\square\end{aligned}$$

2.2.3 Méthode des approximations successives et calcul approché

Revenons à un cadre très général, nous supposons que $\mathbb{E}$ est un espace métrique complet. Au lieu de considérer le schéma classique $x_{k+1} = f(x_k)$ pour le calcul du point fixe x de f, nous introduisons un schéma approché. Le calcul de x_{k+1} est fait avec une erreur $\epsilon > 0$, autrement dit :

$$d(x_{k+1}, f(x_k)) \leq \varepsilon.$$

Trois sources d'erreurs conduisent à de tels shémas.
Primo : le modèle mathématique étudié, représenté ici par la fonction f, peut dépendre de paramètres eux-mêmes affectés d'erreurs. C'est le cas lorsque ceux-ci sont le résultat de calculs approchés ou bien sont des données expérimentales ou des résultats de mesures faites avec une précision finie.

Secondo : les erreurs d'approximation et de troncature (les processus limites sont arrêtés après un nombre fini d'étapes, les fonctions transcendantes sont remplacées par des approximations, et cetera).
Tercio : les erreurs d'arrondis dues à certaines arithmétiques utilisées par les ordinateurs (arithmétique virgule flottante par exemple).
Bien sûr, ces trois séries de causes peuvent être présentes simultanément, ce qui conduit à considérer un shéma itératif approché. Une étude plus approfondie de tels shémas est faite par Chatelin-Frayssé [12] pour des problèmes non linéaires et par N. Higham [23] pour l'algèbre linéaire.

Le shéma que nous allons étudier privilégie les erreurs absolues. On peut lui préférer un shéma adapté aux erreurs relatives comme (dans un cadre d'espaces normés)

$$\|x_{k+1} - f(x_k)\| \leq \varepsilon \|x_k\|.$$

Nous laissons au lecteur intéressé le soin de formuler le résultat correspondant à la proposition qui suit.

Proposition 6. *Soit f une contraction de l'espace métrique complet $\mathbb{E}$ de constante $0 \leq \lambda < 1$ et de point fixe x. Soit $\epsilon > 0$ et soit (x_k) une suite de points de $\mathbb{E}$ qui vérifie*

$$d(x_{k+1}, f(x_k)) \leq \varepsilon.$$

On a

$$d(x_k, x) \leq \lambda^k d(x_0, x) + \frac{\epsilon}{1-\lambda}$$

pour tout $k \geq 0$.

Preuve Nous allons montrer, par récurrence sur k, que

$$d(x_k, x) \leq \lambda^k d(x_0, x) + \epsilon \sum_{i=0}^{k-1} \lambda^i.$$

L'inégalité en résulte puisque $\sum_{i=0}^{k-1} \lambda^i \leq 1/(1-\lambda)$. Pour $k = 0$ il n'y a rien à démontrer. Le passage de k à $k+1$ se fait ainsi :

$$\begin{aligned} d(x_{k+1}, x) &\leq d(x_{k+1}, f(x_k)) + d(f(x_k), f(x)) \leq \epsilon + \lambda d(x_k, x) \\ &\leq \epsilon + \lambda \left(\lambda^k d(x_0, x) + \epsilon \sum_{i=0}^{k-1} \lambda^i \right) = \lambda^{k+1} d(x_0, x) + \epsilon \sum_{i=0}^{k} \lambda^i. \ \square \end{aligned}$$

Ce résultat montre que la suite (x_k) « converge » vers la boule de centre x et de rayon $\epsilon/(1-\lambda)$ au lieu de converger vers x. Notons que ce rayon ne dépend pas de la suite (x_k).

2.2.4 Convergence quadratique

La vitesse avec laquelle la suite des approximations successives converge vers un point fixe est estimée par le Théorème 4 : c'est celle de la convergence d'une suite géométrique $C\lambda^k$ où λ est la constante de contraction. On qualifie de linéaire ce type de convergence, le nombre de décimales exactes augmente linéairement. Dans certains cas on peut aller beaucoup plus vite, c'est ce que nous allons exposer dans le contexte d'une fonction de classe C^2 définie sur un espace de Banach. L'outil essentiel pour cette étude est la formule de Taylor qui est exposée en appendice.

Venons-en au résultat principal de ce paragraphe l'expression $\|D^2f(y)\|$ qui y figure est définie par :

$$\|D^2f(y)\| = \sup_{\|u\|=\|v\|=1} \|D^2f(y)(u,v)\|.$$

Théorème 7. *(Convergence quadratique) Soit $f : \mathbb{E} \to \mathbb{E}$ de classe C^2 et soit x un point fixe de f tel que $Df(x) = 0$. Soient $M > 0$ et $r > 0$ deux nombres pour lesquels les conditions suivantes sont satisfaites :*

1. *$\|D^2f(y)\| \leq 2M$ pour tout y tel que $\|y - x\| \leq r$,*
2. *$2Mr \leq 1$.*

Sous ces hypothèses, pour tout x_0 tel que $\|x_0 - x\| \leq r$, la suite des approximations successives $x_{k+1} = f(x_k)$ vérifie

$$\|x_k - x\| \leq \left(\frac{1}{2}\right)^{2^k-1} \|x_0 - x\|.$$

La suite (x_k) converge donc très rapidement vers x : le théorème précédent montre que le nombre de décimales exactes est multiplié par 2 à chaque itération. On qualifie de quadratique une telle vitesse de convergence.

Preuve Elle repose sur la formule de Taylor 6.1.7 à l'ordre 2 et au voisinage de x :

$$f(y) = f(x) + Df(x)(y-x) + \int_0^1 (1-t)D^2f(x+t(y-x))(y-x)^2dt,$$

ce qui donne, lorsque $\|y - x\| \leq r$,

$$\begin{aligned}
\|f(y) - x\| = \|f(y) - f(x)\| &= \left\| \int_0^1 (1-t)D^2f(x+t(y-x))(y-x)^2dt \right\| \\
&\leq \int_0^1 \left\|(1-t)D^2f(x+t(y-x))(y-x)^2\right\| dt \\
&\leq \int_0^1 (1-t)\left\|D^2f(x+t(y-x))\right\| \|y-x\|^2 dt \\
&\leq \int_0^1 (1-t)2M\|(y-x)\|^2 dt = M\|y-x\|^2.
\end{aligned}$$

A ce stade on raisonne par récurrence. Soit x_0 avec $\|x_0 - x\| \leq r$. Pour $k = 0$ il n'y a rien à démontrer. Le passage de k à $k+1$ se fait ainsi : supposons que $\|x_k - x\| \leq r$ et que l'inégalité du théorème soit vraie pour k. L'inégalité que l'on vient de prouver appliquée à $y = x_k$ donne :

$$\|x_{k+1} - x\| \leq M\|x_k - x\|^2 \leq M\left(\left(\frac{1}{2}\right)^{2^k-1}\|x_0 - x\|\right)^2$$
$$\leq M\|x_0 - x\|\left(\frac{1}{2}\right)^{2^{k+1}-2}\|x_0 - x\| \leq \left(\frac{1}{2}\right)^{2^{k+1}-1}\|x_0 - x\|.$$

Ceci prouve aussi que $\|x_{k+1} - x\| \leq \|x_0 - x\| \leq r$ et achève la démonstration. □

Nous allons maintenant examiner quelles modifications apportent l'introduction d'une erreur de calcul dans un tel schéma.

Théorème 8. *Soit $f : \mathbb{E} \to \mathbb{E}$ de classe C^2 et soit x un point fixe de f tel que $Df(x) = 0$. Soient $M > 0$, $r > 0$ et $\epsilon > 0$ trois nombres pour lesquels les conditions suivantes sont satisfaites :*

1. *$\|D^2 f(y)\| \leq 2M$ pour tout y tel que $\|y - x\| \leq r$,*
2. *$2Mr \leq 1$,*
3. *$4\epsilon \leq r$.*

Soit x_0 tel que $\|x_0 - x\| \leq r$ et soit (x_k) une suite qui vérifie

$$\|x_{k+1} - f(x_k)\| \leq \epsilon.$$

Sous ces hypothèses, pour tout $k \geq 1$,

$$\|x_k - x\| \leq 2\epsilon + \left(\frac{1}{2}\right)^{2^{k-1}}\|x_0 - x\|.$$

Preuve Par récurrence sur k, nous allons prouver qu'il existe une suite (θ_k) de nombres réels > 0 telle que

$$\|x_k - x\| \leq \theta_k\epsilon + \left(\frac{1}{2}\right)^{2^{k-1}}\|x_0 - x\|$$

et que $\|x_k - x\| \leq r$ pout tout k. Il faut noter que, si $\|x_k - x\| \leq r$, on a

$$\|x_{k+1} - x\| \leq \|x_{k+1} - f(x_k)\| + \|f(x_k) - x\| \leq \epsilon + M\|x_k - x\|^2$$

en vertu de l'inégalité prouvée dans la démonstration du théorème précédent. Remarquons aussi que $2M\epsilon \leq 1/4$. On a

$$\|x_1 - x\| \leq \epsilon + M\|x_0 - x\|^2 \leq \epsilon + \frac{1}{2}\|x_0 - x\|,$$

qui correspond à la formule souhaitée avec $\theta_1 = 1$. De plus

$$\|x_1 - x\| \leq \epsilon + \frac{1}{2}\|x_0 - x\| \leq r/4 + r/2 \leq r,$$

ce qui prouve le cas $k = 1$. Le passage de k à $k + 1$ se fait ainsi :

$$\begin{aligned}\|x_{k+1} - x\| &\leq \epsilon + M\|x_k - x\|^2 \leq \epsilon + M\left(\theta_k\epsilon + \left(\frac{1}{2}\right)^{2^{k-1}}\|x_0 - x\|\right)^2 \\ &\leq \epsilon + 2M\theta_k^2\epsilon^2 + 2M\left(\frac{1}{2}\right)^{2^k}\|x_0 - x\|^2 \\ &\leq \epsilon\left(1 + \frac{\theta_k^2}{4}\right) + \left(\frac{1}{2}\right)^{2^k}\|x_0 - x\|,\end{aligned}$$

qui donne la valeur $\theta_{k+1} = 1 + \theta_k^2/4$. La suite (θ_k) est croissante et a pour limite $\theta = 2$ qui est aussi la valeur de $1 + \theta^2/4$. On a donc prouvé que

$$\|x_{k+1} - x\| \leq 2\epsilon + \left(\frac{1}{2}\right)^{2^k}\|x_0 - x\|$$

qui est l'inégalité souhaitée. Enfin, en utilisant $4\epsilon \leq r$ et $\|x_0 - x\| \leq r$, on obtient

$$\|x_{k+1} - x\| \leq \frac{r}{2} + \frac{r}{2} = r,$$

ce qui termine la démonstration. □

Ce résultat prouve que la convergence quadratique n'est pas détruite par l'introduction d'erreurs : la suite des itérés (x_k) va « converger » vers la boule de centre x (la limite exacte) et de rayon 2ϵ. Cette information permet de prévoir avec quelle précision il faut calculer les itérés afin d'obtenir une qualité donnée des résultats.

2.3 Classification des points fixes : définitions

Dans cette section, plutôt théorique, nous allons décrire une classification des points fixes d'une application f en fonction des propriétés de convergence des suites $(f^k(x_0))$ où x_0 est pris dans un voisinage du point fixe x. Nous commencerons par étudier le cas où f est linéaire et $x = 0$. Nous ferons apparaître une décomposition de l'espace ambiant en somme directe de deux sous-espaces : le sous-espace dilaté et le sous-espace contracté et appelés encore sous-espace stable et sous-espace instable (« stable and unstable subspaces » pour les anglophones). Nous prouverons ensuite des résultats similaires dans le cas non linéaire. Le calcul différentiel étant omniprésent dans cette étude, nous nous plaçons dans le cadre d'espaces de Banach. Notons $\mathbb{E}$ et $\mathbb{F}$ deux tels espaces.

Définition 9. *Nous dirons qu'une application f définie sur un ouvert $U \subset \mathbb{E}$ et à valeurs dans un ouvert $V \subset \mathbb{F}$ est un homéomorphisme lorsque f est une bijection continue de U sur V dont l'inverse $f^{-1} : V \to U$ est aussi continu.*

Définition 10. *Nous dirons qu'une application f définie sur un ouvert $U \subset \mathbb{E}$ et à valeurs dans un ouvert $V \subset \mathbb{F}$ est un difféomorphisme lorsque f est une bijection de U sur V, de classe C^1 sur U ainsi que son inverse.*

Nous savons, par le « Théorème d'inversion locale » 185, que si f est de classe C^1 sur U et si $Df(x)$ est un isomorphisme alors f est un difféomorphisme d'un voisinage de x dans $\mathbb{E}$ sur un voisinage de $f(x)$ dans $\mathbb{F}$. De plus, la dérivée de l'application inverse f^{-1} est donnée par $D(f^{-1})(f(x)) = (Df(x))^{-1}$.

Nous allons nous intéresser aux suites $(x_k) = (f^k(x))$. On ne souhaite pas seulement étudier le devenir de x_k lorsque $k \to \infty$ mais aussi leur origine « en remontant le temps » c'est à dire lorsque $k \to -\infty$. Il faut donc étendre la définition de x_k au cas d'entiers k négatifs ce qui suppose que f est bijective. Si l'on regarde ces suites comme décrivant un processus spatio-temporel, l'état spatial est donné par x_k et les entiers k modélisent les différents instants considérés. Les valeurs positives de k décrivent les instants à venir, les valeurs négatives ceux du passé.

Définition 11. *Les itérés de f sont définis par $f^0 = id$ et*

1. *$f^k = f \circ f^{k-1}$ pour tout entier $k \geq 1$,*
   et, lorsque f est une bijection, par
2. *$f^k = f^{-1} \circ f^{k+1}$ pour tout entier $k \leq -1$.*

Définition 12. *Nous dirons qu'un point fixe x de $f : D \subset \mathbb{E} \to \mathbb{E}$ est attractif si toutes les suites $(x_k) = (f^k(x_0))$ sont définies et convergent vers x lorsque $k \to \infty$ quel que soit x_0 dans un voisinage de x dans D.*

Définition 13. *Lorsque $f : D \subset \mathbb{E} \to f(D) \subset \mathbb{E}$ est bijective, nous dirons qu'un point fixe x de f est répulsif si toutes les suites $(x_k) = (f^{-k}(x_0))$ sont définies et convergent vers x lorsque $k \to \infty$ quel que soit x_0 dans un voisinage de x dans D.*

Les concepts « attractif » et « répulsif » s'échangent dans le passage de f à f^{-1} : un point fixe attractif pour f^{-1} est répulsif pour f et un point fixe répulsif pour f^{-1} est attractif pour f.

Un exemple élémentaire est donné par $f : [0, \infty[\to [0, \infty[$ définie par $f(x) = x^2$; 0 est un point fixe attractif puisque $\lim_{k\to\infty} f^k(x) = 0$ pour tout $x \in [0, 1[$, 1 est un point fixe répulsif puisque $\lim_{k\to-\infty} f^k(x) = 1$ pour tout $x \in [0, \infty[$. On voit donc que toutes les suites $(f^k(x))$ pour $x \neq 1$ « proviennent » de 1 et « se dirigent » vers 0 ou vers ∞ ($\lim_{k\to\infty} f^k(x) = \infty$ pour tout $x > 1$).

Nous envisageons maintenant une troisième catégorie de points fixes plus générale que les deux premières : les points fixes hyperboliques. Nous commençons par traiter le cas d'une application linéaire sur lequel nous nous appuierons pour traiter le cas non linéaire.

Définition 14. *Nous dirons qu'une application linéaire $L : \mathbb{E} \to \mathbb{E}$, où $\mathbb{E}$ est un espace de Banach, est hyperbolique si elle est continue et s'il existe une décomposition de $\mathbb{E}$ en somme directe topologique de deux sous-espaces fermés (c'est-à-dire que la somme est directe et que les projecteurs associés sont continus)*

$$\mathbb{E} = E_c \oplus E_d$$

telle que

1. *E_c et E_d soient invariants par L,*
2. *$L|_{E_c}$ soit une contraction,*
3. *$L|_{E_d}$ soit une dilatation.*

Notons que l'un des espaces E_c et E_d peut être égal à $\{0\}$. C'est le cas, par exemple, lorsque L est une homothétie : $L(x) = \lambda x$ avec $\lambda > 1$. On a alors $E_c = \{0\}$ et $E_d = \mathbb{E}$.

2.3.1 Les sous-espaces contractés et dilatés

Les sous-espaces E_c et E_d introduits dans la définition 14 sont caractérisés par la proposition suivante :

Proposition 15. *Soit L un endomorphisme hyperbolique d'un espace de Banach $\mathbb{E}$ et soit $\mathbb{E} = E_c \oplus E_d$ une décomposition de $\mathbb{E}$ associée à L telle qu'en Définition 14. On a*

1. *$E_c = E_c(L)$ où $E_c(L) = \{x \in \mathbb{E} : \lim_{k\to\infty} L^k(x) = 0\}$,*
2. *Si $L : \mathbb{E} \to \mathbb{E}$ est bijective alors, $E_d = E_d(L)$ où $E_d(L) = \{x \in \mathbb{E} : \lim_{k\to-\infty} \|L^k(x)\| = 0\}$.*

$E_c(L)$ et $E_d(L)$ s'appellent les sous-espaces contractés et dilatés associés à L.

Preuve

1. Par hypothèse il existe $0 < \lambda < 1$ et $\Lambda > 1$ tels que $\|Lx\| \leq \lambda\|x\|$ et $\|Ly\| \geq \Lambda\|y\|$ pour tout $x \in E_c$ et $y \in E_d$. Si $x \in E_c$ on a

$$\left\|L^k x\right\| \leq \lambda^k \|x\| \to 0$$

lorsque $k \to \infty$ de sorte que $x \in E_c(L)$. Réciproquement, soit $x \in E_c(L)$. Ecrivons

$$x = x_c + x_d \in E_c \oplus E_d.$$

On a

$$x - x_c = x_d \in E_c(L) \cap E_d$$

de sorte que $x_d = 0$ et que $x = x_c \in E_c$.

2. La seconde assertion se prouve par échange des rôles de L et L^{-1}. □

Dans le cas non linéaire on introduit la définition suivante :

Définition 16. *Soit f définie sur un ouvert U d'un espace de Banach $\mathbb{E}$, à valeurs dans un autre ouvert $V \subset \mathbb{E}$ et qui soit de classe C^1 sur U. Nous dirons qu'un point fixe x de f est hyperbolique lorsque la dérivée $Df(x)$ de f en x est un endomorphisme hyperbolique de $\mathbb{E}$.*

2.3.2 Exemple : les endomorphismes diagonalisables

Soit $\mathbb{E}$ un espace vectoriel de dimension finie et soit $L : \mathbb{E} \to \mathbb{E}$ un endomorphisme diagonalisable de $\mathbb{E}$. Il existe une base $e_1, \ldots, e_n$ de $\mathbb{E}$ et des scalaires $\lambda_1, \ldots, \lambda_n$ tels que

$$L(e_i) = \lambda_i e_i$$

pour tout i.

- Cette application est contractante si $|\lambda_i| < 1$ pour tout i. Dans ce cas

$$\|L(x) - L(y)\| \leq (\max |\lambda_i|) \ \|x - y\| .$$

- Elle est dilatante si $|\lambda_i| > 1$ pour tout i. On a alors

$$\|L(x) - L(y)\| \geq (\min |\lambda_i|) \ \|x - y\| .$$

- Elle est hyperbolique si $|\lambda_i| \neq 1$ pour tout i. Dans ce cas

$$\mathbb{E} = E_c \oplus E_d$$

où E_c (resp. E_d) est engendré par les vecteurs e_i avec $|\lambda_i| < 1$ (resp. $|\lambda_i| > 1$).

- Elle n'est pas hyperbolique s'il existe i avec $|\lambda_i| = 1$. Raisonnons par l'absurde : si $\mathbb{E} = E_c \oplus E_d$ et si L est une contraction sur E_c et une dilatation sur E_d écrivons $e_i = e_c + e_d \in E_c \oplus E_d$. On a $L^k(e_i) = \lambda_i^k e_i$ donc $\left\|L^k(e_i)\right\| = \|e_i\| \neq 0$ pour tout k. D'autre part

$$L^k(e_i) = L^k(e_c) + L^k(e_d).$$

Si $e_d \neq 0$ alors

$$\lim \left\|L^k(e_i)\right\| = \lim \left\|L^k(e_d)\right\| = \infty$$

et si $e_d = 0$ alors

$$\lim \left\|L^k(e_i)\right\| = \lim \left\|L^k(e_c)\right\| = 0$$

ce qui, dans ces deux cas, contredit $\lim \left\|L^k(e_i)\right\| = \|e_i\| \neq 0$.

2.3.3 Exemple : les endomorphismes du plan

Soit L un endomorphisme du plan $\mathbb{R}^2$. A quelle condition 0 est-il un point fixe attractif, répulsif, hyperbolique ? Nous allons déjà rencontrer des situations très générales. Commençons par un théorème de structure de ces endomorphismes.

Théorème 17. *Il existe une base de $\mathbb{R}^2$ dans laquelle la matrice J de L ait l'une des formes suivantes :*

$$\begin{pmatrix} \lambda & 0 \\ 0 & \mu \end{pmatrix} \quad \textit{ou bien} \quad \begin{pmatrix} \lambda & 1 \\ 0 & \lambda \end{pmatrix} \quad \textit{ou bien} \quad \rho \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

pour des nombres λ, $\mu \in \mathbb{R}$, $\rho > 0$ et $0 < \theta < 2\pi$.

Preuve Considérons les deux valeurs propres de L. Si elles sont réelles et distinctes on peut diagonaliser L et on obtient le premier cas. Si elles sont réelles et égales on obtient le premier cas si le sous-espace propre associé est de dimension 2 et le second cas si ce sous-espace propre est de dimension 1. Si les deux valeurs propres de L sont complexes conjuguées $\rho\exp(\pm i\theta)$, $\rho > 0$, $0 < \theta < 2\pi$, il existe deux vecteurs propres complexes conjugués $x \pm iy$, où les vecteurs x et y sont réels et indépendants. Dans la base $\{x, y\}$ on obtient le troisième cas. □

En vertu de ce théorème, il existe une matrice inversible P telle que L (identifié à sa matrice dans la base canonique de $\mathbb{R}^2$) s'écrive $L = PJP^{-1}$. La suite des itérés $(L^k(x))$ est donnée par $L^k(x) = PJ^kP^{-1}x$ ce qui ramène notre étude, via le changement de variable $x = Py$, aux suites $(J^k y)$, $y \in \mathbb{R}^2$.

Premier cas $J = \begin{pmatrix} \lambda & 0 \\ 0 & \mu \end{pmatrix}$. Pour tout $y \in \mathbb{R}^2$ on a $J^k y = \begin{pmatrix} \lambda^k y_1 \\ \mu^k y_2 \end{pmatrix}$ ce qui fait de l'origine un point fixe attractif si $|\lambda|$ et $|\mu| < 1$, répulsif si $|\lambda|$ et $|\mu| > 1$ et hyperbolique si $|\lambda| > 1$ et $|\mu| < 1$ ou bien si $|\lambda| < 1$ et $|\mu| > 1$.

Deuxième cas $J = \begin{pmatrix} \lambda & 1 \\ 0 & \lambda \end{pmatrix}$. Pour tout $y \in \mathbb{R}^2$ on a $J^k y = \begin{pmatrix} \lambda^k y_1 + k\lambda^{k-1} y_2 \\ \lambda^k y_2 \end{pmatrix}$ ce qui fait de l'origine un point fixe attractif si $|\lambda| < 1$ et répulsif si $|\lambda| > 1$.

Troisième cas Les matrices J suivantes correspondent aux deux premiers cas avec des coefficients λ et μ pouvant être de valeur absolue égale à 1 :

$$J = \begin{pmatrix} 1 & 0 \\ 0 & \mu \end{pmatrix}, \quad \begin{pmatrix} -1 & 0 \\ 0 & \mu \end{pmatrix}, \quad \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \quad \begin{pmatrix} -1 & 1 \\ 0 & -1 \end{pmatrix}.$$

Nous laissons étudier au lecteur, à titre d'exercice, les suites $(J^k y)$ qui leurs sont associées. Ce sont des cas de non hyperbolicité.

Quatrième cas $J = \rho \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$. Pour tout $y \in \mathbb{R}^2$ on a

$$J^k y = \rho^k \begin{pmatrix} \cos k\theta & -\sin k\theta \\ \sin k\theta & \cos k\theta \end{pmatrix} y$$

ce qui fait de l'origine un point fixe attractif si $\rho < 1$ et répulsif si $\rho > 1$.

Cinquième cas $J = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$. C'est une rotation d'angle θ autour de l'origine. La suite des itérés $(J^k y)$ reste enfermée dans le cercle centré à l'origine et de rayon $\|y\|$. Cette suite est périodique si θ est un multiple rationnel de π, elle est dense dans ce cercle sinon. C'est un cas de non hyperbolicité.

2.4 Endomorphismes contractants, dilatants et hyperboliques

L'exemple des endomorphismes du plan ainsi que celui des endomorphismes diagonalisables montrent que les propriétés « contractant », « dilatant » et « hyperbolique » se lisent sur le spectre de cet endomorphisme : L est contractant si ses si ses valeurs propres sont à l'intérieur du disque unité, dilatant si elles sont à l'extérieur de ce disque et hyperbolique si aucune des valeurs propres n'est située sur le cercle unité.

Cela est-il encore vrai pour un endomorphisme continu L d'un espace de Banach réel $\mathbb{E}$? Nous allons voir que la réponse est « oui » pour les propriétés « contractant » et « dilatant ». La réponse est encore « oui » dans le cas hyperbolique à condition de supposer que $\mathbb{E}$ soit de dimension finie.

Pour démontrer ces résultats nous devons utiliser le concept de spectre d'un opérateur linéaire continu d'un espace de Banach dont nous allons décrire les aspects les plus élémentaires. Nous renvoyons le lecteur, pour une étude plus complète, aux ouvrages suivants : Bollobàs [7], Dieudonné [18] et Yosida [56].

2.4.1 Spectre d'un opérateur

Soit $\mathbb{F}$ un espace de Banach complexe et soit $M : \mathbb{F} \to \mathbb{F}$ une application linéaire continue. Un nombre complexe $\zeta \in \mathbb{C}$ est une *valeur régulière* de M si $M - \zeta$ id possède un inverse $(M - \zeta \text{ id})^{-1}$. Un tel inverse est nécessairement continu en vertu du Théorème de l'inverse continu (Théorème 198). Les nombres complexes qui ne sont pas des valeurs régulières sont des *valeurs spectrales* de M et leur ensemble est noté

$$\text{Spec } (M).$$

Lorsque $\zeta \in \text{Spec } (M)$ et que $\ker(M - \zeta \text{ id})$ n'est pas réduit à $\{0\}$ on dit que ζ est une *valeur propre* de M. On a alors $Mu = \zeta u$ pour un vecteur $u \neq 0$.

Un tel vecteur est un *vecteur propre* associé à ζ. Mais une valeur spectrale n'est pas nécessairement une valeur propre sauf lorsque M est de dimension finie. Dans ce cas les valeurs spectrales (propres) sont les racines du *polynôme caractéristique*

$$P_M(z) = \det(M - z\ \mathrm{id}).$$

2.4.2 Rayon spectral

Le spectre de M est un ensemble non vide et compact dans $\mathbb{C}$. Pour cette raison on définit le *rayon spectral* de M par

$$\rho(M) = \max_{\zeta \in \mathrm{Spec}\ (M)} |\zeta|\,.$$

Le rayon spectral possède plusieurs propriétés que nous utiliserons par la suite :

Théorème 18.

1. $\rho(M) = \lim_{k\to\infty} n(M^k)^{1/k}$ *où* n *est n'importe quelle norme sur* $\mathbb{F}$ *équivalente à la norme de* $\mathbb{F}$ *(*n *désigne à la fois une norme sur* $\mathbb{F}$ *et la norme d'opérateur associée).*
2. *Pour tout entier* $p \geq 0$, $\rho(M^p) = \rho(M)^p$.
3. $\rho(M) = \inf n(M)$ *où l'infimum est pris pour toutes les normes* n *équivalentes à la norme de* $\mathbb{F}$.

Preuve Notons $\|.\|$ la norme de $\mathbb{F}$ et n une norme équivalente. L'égalité $\rho(M) = \lim_{k\to\infty} \left\|M^k\right\|^{1/k}$ est prouvée par Yosida [56], Chap. VIII-2, Théorèmes 3 et 4. Pour passer à une norme équivalente, on note que si

$$\alpha \|x\| \leq n(x) \leq \beta \|x\|$$

pour tout $x \in \mathbb{F}$ alors

$$\frac{\alpha}{\beta} \|u\| \leq n(u) \leq \frac{\beta}{\alpha} \|u\|$$

pour tout endomorphisme continu u de $\mathbb{F}$. Ainsi

$$\left(\frac{\alpha}{\beta}\right)^{1/k} \left\|M^k\right\|^{1/k} \leq n(M^k)^{1/k} \leq \left(\frac{\beta}{\alpha}\right)^{1/k} \left\|M^k\right\|^{1/k}$$

ce qui prouve que $\lim_k \left\|M^k\right\|^{1/k} = \lim_k n(M^k)^{1/k}$.

La seconde assertion utilise la première :

$$\rho(M^p) = \lim_k \left\|M^{pk}\right\|^{1/k} = \left(\lim_k \left\|M^{pk}\right\|^{1/pk}\right)^p = \rho(M)^p.$$

Prouvons la troisième assertion. Comme $n(M^k) \leq n(M)^k$ on obtient

$$n(M^k)^{1/k} \leq n(M)$$

pour tout $k > 0$ d'où

$$\rho(M) \leq \inf_n n(M).$$

Pour prouver que cet infimum est égal à $\rho(M)$ on se donne un réel $\alpha > \rho(M)$ et on construit une norme n équivalente à $\|.\|$ telle que

$$n(M) \leq \alpha.$$

Par la première assertion, il existe $p > 0$ tel que $\|M^p\|^{1/p} < \alpha$. On a

$$\|M^p x\| \leq \alpha^p \|x\|$$

pour tout $x \in \mathbb{F}$. Posons

$$n(x) = \sum_{i=0}^{p-1} \alpha^{p-i-1} \left\|M^i x\right\|.$$

C'est une norme sur $\mathbb{F}$ équivalente à $\|.\|$. De plus, pour tout $x \in \mathbb{F}$,

$$n(Mx) = \sum_{i=0}^{p-1} \alpha^{p-i-1} \left\|M^{i+1} x\right\| = \alpha n(x) + \|Mx\| - \alpha^p \|x\| \leq \alpha n(x).$$

Ainsi $n(M) \leq \alpha$. □

2.4.3 Spectre d'un endomorphisme réel

Donnons nous maintenant une application linéaire et continue $L : \mathbb{E} \to \mathbb{E}$ où $\mathbb{E}$ est un espace de Banach réel. A cet espace nous associons son complexifié

$$\mathbb{F} = \mathbb{E} \oplus i\mathbb{E}.$$

C'est un espace vectoriel complexe pour l'addition

$$(x + iy) + (x' + iy') = (x + x') + i(y + y')$$

et la multiplication externe

$$(\alpha + i\beta)(x + iy) = (\alpha x - \beta y) + i(\beta x + \alpha y)$$

où x, x', y et $y' \in \mathbb{E}$, α et $\beta \in \mathbb{R}$. $\mathbb{F}$ est un espace de Banach complexe pour la norme

$$\|x + iy\|_{\mathbb{F}} = \left(\|x\|_{\mathbb{E}}^2 + \|y\|_{\mathbb{E}}^2\right)^{1/2}.$$

Nous omettrons désormais les indices $\mathbb{E}$ et $\mathbb{F}$ dans l'écriture de ces normes. Le prolongement M de L à $\mathbb{F}$ tout entier est défini par

$$M : \mathbb{F} \to \mathbb{F}, \quad M(x+iy) = L(x) + iL(y).$$

Notons que $M(x) = L(x)$ pour tout $x \in \mathbb{E}$ et que

$$\|M\| = \|L\|.$$

Le spectre de L est défini par

$$\text{Spec } (L) = \text{Spec } (M)$$

et le rayon spectral de L par

$$\rho(L) = \rho(M).$$

Les propriétés suivantes se déduisent facilement du Théorème 18 :

Théorème 19.

1. $\rho(L) = \lim_{k\to\infty} n(L^k)^{1/k}$ *où* n *est n'importe quelle norme sur* $\mathbb{E}$ *équivalente à la norme de* $\mathbb{E}$.
2. $\rho(L^p) = \rho(L)^p$ *pour tout entier* $p \geq 0$.
3. $\rho(L) = \inf n(L)$ *où l'infimum est pris pour toutes les normes* n *équivalentes à la norme de* $\mathbb{E}$.

2.4.4 Endomorphismes contractants

Les endomorphismes contractants sont caractérisés par le théorème suivant :

Théorème 20. *Pour un endomorphisme continu* L *d'un espace de Banach* $\mathbb{E}$ *il y a equivalence entre*

1. *Pour tout* $x \in \mathbb{E}$, $\lim_{k\to\infty} L^k(x) = 0$,
2. $\rho(L) < 1$,
3. *Il existe une norme* n *sur* $\mathbb{E}$ *équivalente à la norme de* $\mathbb{E}$ *et un scalaire* λ, $0 \leq \lambda < 1$, *tels que, pour tout* $x \in \mathbb{E}$, *on ait* $n(Lx) \leq \lambda n(x)$.

Preuve $1 \Rightarrow 2$. Si $\lim_{k\to\infty} L^k(x) = 0$ pour tout $x \in \mathbb{E}$, par le Théorème de Banach-Steinhaus (Théorème 197), la convergence est uniforme en x et $\lim_{k\to\infty} \left\|L^k\right\| = 0$. On peut donc supposer que $\left\|L^k\right\| < 1$ pour un entier k assez grand. Par le Théorème 19 on obtient

$$\rho(L) = \rho(L^k)^{1/k} \leq \left\|L^k\right\|^{1/k} < 1.$$

$2 \Rightarrow 3$. Soit λ tel que $\rho(L) < \lambda < 1$. Par le Théorème 19 il existe une norme n sur $\mathbb{E}$ équivalente à la norme de $\mathbb{E}$ telle que

$$\rho(L) \leq n(L) < \lambda < 1$$

d'où

$$n(Lx) \leq \lambda n(x)$$

pour tout $x \in \mathbb{E}$.

$3 \Rightarrow 1$. On a $n(L^k x) \leq \lambda^k n(x) \to 0$ lorsque $k \to \infty$. □

2.4.5 Endomorphismes dilatants

Les endomorphismes dilatants sont caractérisés par le théorème suivant :

Théorème 21. *Pour un endomorphisme bijectif et continu L d'un espace de Banach $\mathbb{E}$ il y a equivalence entre*

1. *Pour tout $x \in \mathbb{E}$, $\lim_{k\to\infty} L^{-k}(x) = 0$,*
2. *Il existe $\Lambda > 1$ tel que $|\lambda| \geq \Lambda$ pour toute valeur spectrale $\lambda \in Spec\ (L)$,*
3. *Il existe une norme n sur $\mathbb{E}$ équivalente à la norme de $\mathbb{E}$ et un scalaire $\Lambda > 1$, tels que, pour tout $x \in \mathbb{E}$, on ait $n(Lx) \geq \Lambda n(x)$.*

Sous ces conditions, $\lim_{k\to\infty} \left\|L^k x\right\| = \infty$ pour tout $x \neq 0$.

Preuve L'équivalence de ces énoncés est une conséquence du Théorème 20 appliqué à L^{-1} et de l'équivalence entre $\lambda \in \text{Spec}\ (L)$ et $\lambda^{-1} \in \text{Spec}\ (L^{-1})$. □

Corollaire 22. *Pour un endomorphisme L d'un espace vectoriel de dimension finie $\mathbb{E}$ il y a equivalence entre*

1. *L est bijectif et, pour tout $x \in \mathbb{E}$, $\lim_{k\to\infty} L^{-k}(x) = 0$,*
2. *Il existe $\Lambda > 1$ tel que $|\lambda| \geq \Lambda$ pour tout valeur propre λ de L,*
3. *Il existe une norme n sur $\mathbb{E}$ et un scalaire $\Lambda > 1$, tels que $n(Lx) \geq \Lambda n(x)$ pour tout $x \in \mathbb{E}$*
4. *$\lim_{k\to\infty} \left\|L^k x\right\| = \infty$ pour tout $x \neq 0$.*

Preuve Les assertions 1, 2 et 3 sont équivalentes par le Théorème 21 et parce que les conditions 2 et 3 impliquent l'inversibilité de L. Montrons que $3 \Rightarrow 4$ (facile) et que $4 \Rightarrow 2$. Raisonnons par l'absurde. S'il existe $\lambda \in \text{Spec}\ (L)$ avec $|\lambda| \leq 1$, prenons un vecteur propre $u = x + iy$ de l'application M qui prolonge L sur le complexifié de $\mathbb{E}$. On a $M^k u = L^k x + iL^k y$ et

$$\left\|L^k x\right\| \text{ et } \left\|L^k y\right\| \leq \left\|M^k u\right\| = |\lambda|^k \|u\| \leq \|u\|.$$

Comme u est un vecteur propre, x ou $y \neq 0$ et cela contredit 4. □

2.4.6 Endomorphismes hyperboliques

Passons au cas hyperbolique qui est beaucoup plus compliqué à étudier.

Théorème 23. *Soit $\mathbb{E}$ un espace vectoriel normé réel de dimension finie. Un endomorphisme L de $\mathbb{E}$ est hyperbolique si et seulement si toutes ses valeurs propres sont de module $\neq 1$.*

Preuve Pour montrer que cette condition est nécessaire on raisonne par l'absurde : supposons que L soit hyperbolique, que $\mathbb{E} = E_c \oplus E_d$ comme dans la définition 14 et qu'une valeur propre λ de L soit de module 1. Notons $M : \mathbb{F} \to \mathbb{F}$ les complexifiés de L et $\mathbb{E}$ (paragraphe 2.4.3). Il existe

$$u = x + iy \in \mathbb{F}, \quad u \neq 0,$$

avec $Mu = \lambda u$. Comme $M^k u = \lambda^k u$ et que $M^k u = L^k x + iL^k y$ on obtient

$$\left\|M^k u\right\|^2 = \left\|L^k x\right\|^2 + \left\|L^k y\right\|^2 = \|x\|^2 + \|y\|^2 \neq 0.$$

Les vecteurs x et y sont eux-mêmes décomposés en $x = x_c + x_d$ et $y = y_c + y_d \in \mathbb{E}_c \oplus \mathbb{E}_d$ d'où

$$\left\|L^k x_c + L^k x_d\right\|^2 + \left\|L^k y_c + L^k y_d\right\|^2 = \|x\|^2 + \|y\|^2 \neq 0.$$

Comme les suites $(L^k x_c)$ et $(L^k y_c)$ ont pour limite 0 on obtient

$$\lim_k \left\|L^k x_d\right\|^2 + \left\|L^k y_d\right\|^2 = \|x\|^2 + \|y\|^2 \neq 0,$$

mais d'après le Corollaire 22 une telle limite ne peut valoir que 0 ou ∞ : contradiction !

Montrons maintenant que la condition est suffisante. Notons

$$C = \{\lambda \in \text{Spec } (L) : |\lambda| < 1\} \text{ et } D = \{\lambda \in \text{Spec } (L) : |\lambda| > 1\}$$

de sorte que

$$\text{Spec } (L) = D \cup C.$$

Notons $n = \dim \mathbb{E}$ et considérons les polynômes

$$P_D(z) = \Pi_{\lambda \in D}(z - \lambda)^n \text{ et } P_C(z) = \Pi_{\lambda \in C}(z - \lambda)^n.$$

Ce sont des polynômes réels parce que C et D sont invariants par conjugaison complexe. Les polynômes d'endomorphisme associés sont

$$P_D(L) = \Pi_{\lambda \in D}(L - \lambda \text{ id})^n \text{ et } P_C(L) = \Pi_{\lambda \in C}(L - \lambda \text{ id})^n$$

dont les noyaux sont notés

$$E_c = \ker P_C(L) \text{ et } E_d = \ker P_D(L).$$

Nous allons prouver que $\mathbb{E} = E_c \oplus E_d$, que E_c et E_d sont invariants par L, que $L|_{E_c}$ est une contraction et que $L|_{E_d}$ est une dilatation. Ainsi L sera hyperbolique.

1. $L(E_c) \subset E_c$ et $L(E_d) \subset E_d$. Soit $x \in E_c$ de sorte que $P_C(L)x = 0$. On a

$$P_C(L)(L(x)) = (P_C(L) \circ L)(x) = (L \circ P_C(L))(x) = 0$$

et donc $L(x) \in E_c$. Idem pour E_d.

2. $L|_{E_c}$ est une contraction. Soit λ une valeur propre de $L|_{E_c}$. On a

$$P_C(\lambda) \in P_C(\text{Spec } (L|_{E_c})) = \text{Spec } (P_C(L|_{E_c})) = 0$$

puisque $P_C(L|_{E_c}) = 0$. Ceci prouve que $\lambda \in C$ et donc que $|\lambda| < 1$. Par le Théorème 20 $L|_{E_c}$ est une contraction. L'égalité

$$P_C(\text{Spec } (v)) = \text{Spec } (P_C(v))$$

utilisée ci-dessus est vraie pour tout polynôme P et pour tout endomorphisme v de $\mathbb{E}$. Il suffit de le prouver pour une matrice $n \times n$ complexe A. On écrit $A = BTB^{-1}$ avec T triangulaire supérieure de sorte que

$$\text{Spec } (A) = \text{Spec } (T) = \{t_{ii} : 1 \leq i \leq n\}.$$

Comme $P(A) = BP(T)B^{-1}$ on obtient

$$\text{Spec } (P(A)) = \text{Spec } (P(T)) = \{P(t_{ii}) : 1 \leq i \leq n\} = P(\text{Spec } (A)).$$

Par un argument similaire on montre que

3. $L|_{E_d}$ est une dilatation.

4. $\mathbb{E} = E_c \oplus E_d$. Comme les polynômes $P_C(z)$ et $P_D(z)$ n'ont pas de racine commune, ils sont premiers entre-eux et, par le théorème de Bézout, il existe deux polynômes réels $A(z)$ et $B(z)$ tels que

$$A(z)P_C(z) + B(z)P_D(z) = 1.$$

Les polynômes d'endomorphismes correspondant vérifient

$$A(L)P_C(L) + B(L)P_D(L) = \text{id}$$

de sorte que

$$A(L)P_C(L)x + B(L)P_D(L)x = x$$

pour tout $x \in \mathbb{E}$. Cette identité prouve que $E_c \cap E_d = \{0\}$. Montrons que $A(L)P_C(L)x \in E_d$ et que $B(L)P_D(L)x \in E_c$ ce qui prouvera que $\mathbb{E} = E_c \oplus E_d$. Le polynôme

$$P_C(z)P_D(z) = \Pi_{\lambda \in \text{Spec } (L)}(z - \lambda)^n$$

est un multiple du polynôme caractéristique de L

$$P_L(z) = \Pi_{\lambda \in \text{Spec } (L)}(z - \lambda)^{n(\lambda)}$$

où $n(\lambda) \leq n$ est la multiplicité de le valeur propre λ. Par le Théorème de Cayley-Hamilton, $P_L(L) = 0$ et donc $P_C(L)P_D(L) = 0$. On en déduit que

$$P_D(L)\left(A(L)P_C(L)x\right) = P_C(L)P_D(L)\left(A(L)x\right) = 0$$

c'est à dire que $A(L)P_C(L)x \in E_d$. De la même manière

$$P_C(L)\left(B(L)P_D(L)x\right) = P_C(L)P_D(L)\left(B(L)x\right) = 0$$

et donc $B(L)P_D(L)x \in E_c$, ce qui termine cette démonstration. □

2.5 Le cas non linéaire : le théorème de Grobman-Hartman

Soit $\mathbb{E}$ un espace de Banach. Lorsque $f : \mathbb{E} \to \mathbb{E}$ n'est plus un opérateur linéaire la situation est-elle différente ? Nous allons voir que l'on a une bonne théorie de la linéarisation, qui permet de déduire la structure d'un point fixe hyperbolique x de f des propriétés de la dérivée $Df(x)$. Ceci se fait au travers du théorème de Grobman-Hartman qui permet de passer de f à $Df(x)$ par un changement de variable h bijectif et bicontinu.

Théorème 24. *(Grobman-Hartman) Soit f un difféomorphisme de classe C^1 défini sur un ouvert U de $\mathbb{E}$ et soit x un point fixe hyperbolique de f dans U. Il existe un homéomorphisme h d'un voisinage ouvert de 0 dans $\mathbb{E}$ sur un voisinage ouvert de x dans U tel que $f = h \circ Df(x) \circ h^{-1}$. On dit alors que f et $Df(x)$ sont topologiquement conjugués.*

Il n'est pas toujours possible pour h d'être de classe C^1. Nous renvoyons le lecteur intéressé par ces questions à Demazure [15] ou à Hartman [22].

Quel est le comportement de la suite des itérés $x_k = f^k(x_0)$ où x_0 est pris dans un voisinage du point fixe x ? Par le changement de variable $x = h(y)$ on se ramène à la suite $y_k = Df(x)^k y_0$, c'est-à-dire au cas linéaire. On peut alors utiliser les Théorèmes 20 et 21 qui donnent les deux théorèmes suivants :

Théorème 25. *Soit f un difféomorphisme de classe C^1 défini sur un ouvert U de $\mathbb{E}$ et soit x un point fixe hyperbolique de f dans U. Il y a équivalence entre :*

1. *x est un point fixe attractif,*
2. *Le rayon spectral de $Df(x)$ vérifie $\rho(Df(x)) < 1$,*
3. *Il existe une distance définie sur un voisinage de x pour laquelle f est une contraction.*

Preuve L'équivalence des deux premiers énoncés vient d'être justifiée et le théorème des approximations successives prouve que l'existence d'une distance pour laquelle f est une contraction fait de x un point fixe attractif. La construction de la distance d utilise la norme n du Théorème 20 ainsi que

le changement de variable h du Théorème de Grobman-Hartman : on pose $d(u,v) = n(h^{-1}(u) - h^{-1}(v))$. Ainsi

$$\begin{aligned} d(f(y), f(z)) &= n(Df(x)(h^{-1}(y)) - Df(x)(h^{-1}(z))) \leq \lambda n(h^{-1}(y) - h^{-1}(z)) \\ &= \lambda d(y,z) \end{aligned}$$

dans le cas contractant. □

Théorème 26. *Soit f un difféomorphisme de classe C^1 défini sur un ouvert U de $\mathbb{E}$ et soit x un point fixe hyperbolique de f dans U. Il y a équivalence entre :*

1. *x est un point fixe répulsif,*
2. *Il existe $\Lambda > 1$ tel que $|\lambda| \geq \Lambda$ pour tout $\lambda \in Spec\ (Df(x))$*
3. *Il existe une distance définie sur un voisinage de x pour laquelle f est une dilatation.*

Preuve On procède comme pour le théorème précédent en utilisant le Théorème de Grobman-Hartman et le Théorème 21. □

Nous allons prouver un théorème plus général que celui de Grobman-Hartman, puis nous déduirons celui-ci de celui-là. Nous suivons l'exposé de M. Shub, 1978, [41], dont nous recommandons la lecture.

Définition 27. *Notons $C_b(\mathbb{E})$ l'espace des fonctions $f : \mathbb{E} \to \mathbb{E}$ qui sont continues et bornées. Cet espace est muni de la norme uniforme*

$$\|f\| = \sup_{x \in \mathbb{E}} \|f(x)\|$$

qui en fait un espace de Banach.

Rappelons que l'on note $\mathrm{Lip}(f)$ la plus petite constante de Lipschitz pour f. L'énoncé central que nous allons prouver est le suivant :

Théorème 28. *Soit $L : \mathbb{E} \to \mathbb{E}$ un automorphisme hyperbolique. Il existe $\epsilon > 0$ tel que, pour toute fonction $k \in C_b(\mathbb{E})$ qui soit lipschitzienne avec $\mathrm{Lip}(k) \leq \epsilon$, et pour $f = L + k$, il existe un homéomorphisme $h : \mathbb{E} \to \mathbb{E}$ qui conjugue f et L au sens où $f = h \circ L \circ h^{-1}$.*

Commençons par une proposition qui prouve que lorsque l'on perturbe un homéomorphisme par une « petite » application lipschitzienne on obtient encore un homéomorphisme.

Proposition 29. *Soient U et V des ouverts de $\mathbb{E}$ et f un homéomorphisme de U sur V dont l'inverse est lipschitzien. Soit $h : U \to \mathbb{E}$ lipschitzienne et telle que $\mathrm{Lip}(f^{-1})\mathrm{Lip}(h) < 1$. Alors $g = f + h$ est un homéomorphisme de U sur V et son inverse est lipschitzien de constante*

$$\mathrm{Lip}(g^{-1}) \leq \frac{\mathrm{Lip}(f^{-1})}{1 - \mathrm{Lip}(f^{-1})\mathrm{Lip}(h)}.$$

De plus, lorsque h est bornée sur U, on peut écrire $g^{-1} = f^{-1} + k$ et la fonction k ainsi définie est elle-même bornée : soit $x_0 \in U$ et supposons que $\|h(x)\| \leq C$ pour tout $x \in U$. Alors

$$\|k(y)\| \leq C \,\mathrm{Lip}\ (f^{-1})$$

pour tout y.

Preuve On va commencer par prouver que g est injective et calculer la constante de Lipschitz de g^{-1} par la même occasion. On a

$$\begin{aligned}\|g(x) - g(y)\| &= \|f(x) - f(y) + h(x) - h(y)\| \\ &\geq \|f(x) - f(y)\| - \|h(x) - h(y)\| \\ &\geq \frac{1}{\mathrm{Lip}(f^{-1})}\|x - y\| - \mathrm{Lip}(h)\|x - y\| \\ &= \frac{1 - \mathrm{Lip}(f^{-1})\mathrm{Lip}(h)}{\mathrm{Lip}(f^{-1})}\|x - y\|\end{aligned}$$

et comme $\mathrm{Lip}(f^{-1})\mathrm{Lip}(h) < 1$, cette dernière constante est positive. D'où l'injectivité de g.

Avant de montrer la surjectivité, remarquons que l'on peut remplacer la situation $g = f + h$ par $f^{-1}g = \mathrm{id} + f^{-1}h$ c'est-à-dire prendre pour f l'application identité et $V = U$. L'hypothèse est alors $\mathrm{Lip}(h) < 1$. Etant donné $v \in U$ on veut prouver qu'il existe $x \in U$ tel que $x + h(x) = v$. C'est une équation de point fixe que l'on écrit $x = v - h(x)$. On va utiliser le théorème 4 pour prouver l'existence d'un tel point fixe. Quitte à composer par des translations on suppose que $v = 0$ et $h(v) = 0$. Notons $\bar{B}_r$ une boule fermée de centre 0 et de rayon $r > 0$ qui soit contenue dans U. On a

$$-h(\bar{B}_r) \subset \bar{B}_{\mathrm{Lip}(h)r} \subset \bar{B}_r$$

donc $-h$ est une contraction de $\bar{B}_r$ dans elle-même, elle possède un unique point fixe dans cette boule : la surjectivité est établie.

L'inégalité

$$\|g(x) - g(y)\| \geq \frac{1 - \mathrm{Lip}(f^{-1})\mathrm{Lip}(h)}{\mathrm{Lip}(f^{-1})}\|x - y\|$$

appliquée à $x = g^{-1}(u)$ et $y = g^{-1}(v)$ donne

$$\|g^{-1}(u) - g^{-1}(v)\| \leq \frac{\mathrm{Lip}(f^{-1})}{(1 - \mathrm{Lip}(f^{-1})\mathrm{Lip}(h))}\|u - v\|$$

de sorte que g^{-1} est lipschitzienne. Elle est donc continue et g est un homéomorphisme.

La dernière assertion de la proposition résulte de l'argument suivant. Posons $y = g(x)$ alors :

$$k(y) = g^{-1}(y) - f^{-1}(y) = f^{-1}(f(x)) - f^{-1}(g(x))$$

de sorte que

$$\|k(y)\| \leq \mathrm{Lip}(f^{-1})\|h(x)\| \leq C\mathrm{Lip}(f^{-1}). \quad \square$$

Le lemme suivant sera utilisé dans la suite :

Lemme 30. *Soit $h : U \to V$ un homéomorphisme entre deux ouverts de $\mathbb{E}$ dont l'inverse est lipschitzien et vérifie $\mathrm{Lip}(h^{-1}) < \mu$. Notons $\bar{B}(x,r)$ la boule fermée de centre x et de rayon $r > 0$. Alors $h(\bar{B}(x,r)) \supset \bar{B}(h(x), r/\mu)$.*

Preuve On suppose que $x = 0 = h(0)$ et on note $\bar{B}_r = \bar{B}(0,r)$. Soit $v \in \bar{B}_{r/\mu}$, $v \neq 0$, et posons

$$t = \sup\{s > 0 : [0, sv] \subset h(\bar{B}_r)\}.$$

$t > 0$ puisque l'image $h(\bar{B}_r)$ contient un voisinage de 0. Ce supremum est atteint : la condition de Lipschitz

$$\|h^{-1}(sv) - h^{-1}(s'v)\| \leq \mu|s - s'| \, \|v\|$$

prouve que la limite $\lim_{s\to t} h^{-1}(sv)$ existe et appartient à $\bar{B}_r$. Autrement dit $h^{-1}(tv) \in \bar{B}_r$ ou encore $tv \in h(\bar{B}_r)$. Nous allons voir que, pour tout v, on a $t \geq 1$ ce qui prouve le lemme. Raisonnons par l'absurde. Si $t < 1$ alors

$$\|h^{-1}(tv)\| = \|h^{-1}(tv) - h^{-1}(0)\| \leq \mu t\|v\| < \mu\|v\| \leq r$$

et donc $tv \in h(B_r)$ où B_r est la boule ouverte de centre 0 et de rayon r. Comme cet ensemble est ouvert il existe $u > t$ tel que $[t, u[v \subset h(B_r)$ ce qui contredit la maximalité de t. $\square$

La proposition suivante donne un théorème de point fixe pour une petite perturbation d'un endomorphisme hyperbolique. Commençons par décrire le contexte de cette proposition.

Soit $L : \mathbb{E} \to \mathbb{E}$ un automorphisme hyperbolique (donc continu) et soit $\mathbb{E} = E_c \oplus E_d$ la décomposition de $\mathbb{E}$ en somme directe topologique des sous-espaces contractés et dilatés. Notons p_c et p_d les projections de $\mathbb{E}$ sur E_c et E_d parallèlement à E_d et E_c. Ce sont des applications linéaires et continues.

Pour tout $x \in \mathbb{E}$ écrivons

$$x = p_c(x) + p_d(x) = x_c + x_d \in E_c \oplus E_d$$

la décomposition de x sur cette somme directe.

Introduisons une nouvelle norme sur $\mathbb{E}$ adaptée à cette décomposition. Il s'agit de

$$\|x\|_{ad} = \max(\|x_c\|, \|x_d\|).$$

Cette nouvelle norme est équivalente à l'ancienne. En effet

$$\|x\| = \|x_c + x_d\| \leq \|x_c\| + \|x_d\| \leq 2\|x\|_{ad}$$

et

$$\|x\|_{ad} = \max(\|x_c\|, \|x_d\|) = \max(\|p_c(x)\|, \|p_d(x)\|) \leq \max(\|p_c\|, \|p_d\|)\, \|x\|\,.$$

A cette « norme adaptée » nous associons une norme d'endomorphisme $\|L\|_{ad}$ qui vérifie, par définition,

$$\|L(x)\|_{ad} \leq \|L\|_{ad}\|x\|_{ad}$$

pour tout $x \in \mathbb{E}$.

Les restrictions de L sont notées $L_c = L|_{E_c} : E_c \to E_c$ et $L_d = L|_{E_d} : E_d \to E_d$. Notons que L est hyperbolique pour la norme adaptée. En effet les deux normes $\|.\|$ et $\|.\|_{ad}$ coïncident sur les sous-espaces E_c et E_d. Puisque L est hyperbolique, il existe deux constantes $0 \leq \lambda, \mu < 1$ telles que

$$\|L_c\| = \|L_c\|_{ad} \leq \lambda < 1$$

et

$$\|L_d^{-1}\| = \|L_d^{-1}\|_{ad} \leq \mu < 1.$$

Quitte à prendre pour λ la plus grande de ces deux constantes, on peut supposer que

$$\|L_c\| = \|L_c\|_{ad} \leq \lambda < 1$$

et

$$\|L_d^{-1}\| = \|L_d^{-1}\|_{ad} \leq \lambda < 1.$$

Ces considérations montrent qu'il est équivalent de prouver le Théorème 28 pour la norme initiale ou pour la norme adaptée. Pour cette raison, dans les lignes qui suivent, nous supposons que $\|.\|$ est une norme adaptée c'est à dire que

$$\|x\| = \|x\|_{ad}$$

pour tout $x \in \mathbb{E}$.

Nous notons $\bar{B}_r$ la boule fermée de centre 0 et de rayon r et B_r la boule ouverte.

Proposition 31. *Soit $f : \bar{B}_r \to \mathbb{E}$ une perturbation de L : $f = L + h$ avec h lipschitzienne et vérifiant* $\mathrm{Lip}(h) \leq \varepsilon$ et $\|f(0)\| \leq \delta$. *Supposons que les inégalités suivantes soient satisfaites :*

$$\lambda + \epsilon < 1 \quad et \quad \delta \leq r(1 - \lambda - \epsilon).$$

Alors f a un unique point fixe x_f contenu dans $\bar{B}_r$, sa norme est majorée par

$$\|x_f\| \leq \frac{\|f(0)\|}{1-\lambda-\epsilon}$$

et, pour deux fonctions f et g perturbations de L vérifiant les hypothèses ci-dessus, on a

$$\|x_f - x_g\| \leq \frac{d(f,g)}{1-\lambda-\epsilon}$$

où d est la distance associée à la convergence uniforme des fonctions $\bar{B}_r \to \mathbb{E}$ c'est-à-dire,

$$d(f,g) = \sup_{x\in\bar{B}_r} \|f(x) - g(x)\|.$$

Preuve Notons $f_c = p_c \circ f$ et $f_d = p_d \circ f$ les projections de f sur E_c et E_d et $x_d = p_d(x)$. On définit une application $\bar{f} : \bar{B}_r \to \mathbb{E}$ par

$$\bar{f}(x) = L_d^{-1}\left(x_d + L_d(x_d) - f_d(x)\right) + f_c(x).$$

f et $\bar{f}$ ont les mêmes points fixes. On va montrer que $\bar{f}$ est contractante et que $\bar{f}\left(\bar{B}_r\right) \subset \bar{B}_r$. Le Théorème 4 permettra de conclure. Notons que f_c est lipschitzienne de constante $\mathrm{Lip}(f_c) \leq \lambda+\varepsilon$. Puisque $\|.\|$ est une norme adaptée on a

$$\begin{aligned}&\|\bar{f}(x) - \bar{f}(y)\| \\ &= \max\Big(\big\|L_d^{-1}\big((x_d - y_d) + (L_d(x_d) - f_d(x)) - (L_d(y_d) - f_d(y))\big)\big\|, \|f_c(x) - f_c(y)\|\Big).\end{aligned}$$

Comme

$$L_d(x_d) - f_d(x) = L_d(x_d) - p_d(L(x) + h(x)) = -p_d(h(x))$$

est lipschitzienne de constante ε on a

$$\|\bar{f}(x) - \bar{f}(y)\| \leq \max\left(\lambda(1+\epsilon)\|x-y\|, (\lambda+\varepsilon)\|x-y\|\right) \leq (\lambda+\varepsilon)\|x-y\|$$

et comme on a supposé que $\lambda + \epsilon < 1$, $\bar{f}$ est une contraction.

Montrons que $\bar{f}\left(\bar{B}_r\right) \subset \bar{B}_r$. D'après ce qui précède

$$\begin{aligned}\|\bar{f}(x)\| \leq (\lambda+\varepsilon)\|x\| + \|\bar{f}(0)\| &= (\lambda+\varepsilon)\|x\| + \max\left(\|L_d^{-1} f_d(0)\|, \|f_c(0)\|\right) \\ &\leq (\lambda+\varepsilon)\|x\| + \|f(0)\|.\end{aligned}$$

Lorsque $x \in \bar{B}_r$ on obtient

$$\|\bar{f}(x)\| \leq (\lambda+\varepsilon)r + \|f(0)\| \leq (\lambda+\varepsilon)r + \delta \leq r$$

puisque l'on a supposé que $\delta \leq r(1-\lambda-\epsilon)$.

Estimons la norme du point fixe. Comme x_f est la limite de la suite $(\bar{f}^k(0))$, en utilisant l'inégalité $\|\bar{f}(x)\| \leq (\lambda+\varepsilon)\|x\| + \|f(0)\|$ on montre par

récurrence que

$$\|\bar{f}^{k+1}(0)\| \leq \|f(0)\| \sum_{i=0}^{k} (\lambda+\varepsilon)^i.$$

Cette dernière expression se majore par la somme de la série égale à $1/(1-(\lambda+\varepsilon))$ d'où

$$\|x_f\| \leq \frac{\|f(0)\|}{1-\lambda-\epsilon}.$$

Montrons la continuité du point fixe par rapport à f. Définissons $\bar{g}$ comme $\bar{f}$:

$$\begin{aligned}\|\bar{f}(x)-\bar{g}(x)\| &= \|L_d^{-1}(g_d(x)-f_d(x))+(f_c(x)-g_c(x))\| \\ &= \max\left(\|L_d^{-1}(g_d(x)-f_d(x))\|, \|f_c(x)-g_c(x)\|\right) \\ &\leq \max\left(\lambda\|g_d(x)-f_d(x)\|, \|f_c(x)-g_c(x)\|\right) \leq \|f(x)-g(x)\| \leq d(f,g)\end{aligned}$$

d'où

$$\begin{aligned}\|x_f - x_g\| = \|\bar{f}(x_f)-\bar{g}(x_g)\| &\leq \|\bar{f}(x_f)-\bar{g}(x_f)\| + \|\bar{g}(x_f)-\bar{g}(x_g)\| \\ &\leq d(f,g) + (\lambda+\varepsilon)\|x_f - x_g\|,\end{aligned}$$

cette dernière inégalité venant du fait que $\mathrm{Lip}(g) \leq \lambda+\varepsilon$. On en déduit l'inégalité souhaitée :

$$\|x_f - x_g\| \leq \frac{d(f,g)}{1-\lambda-\epsilon}. \ \square$$

Le dernier des résultats intermédiaires qui nous conduisent au Théorème 28 est le suivant :

Lemme 32. *Soit ϵ tel que $\epsilon\|L^{-1}\| < 1$. Soient k et $k' \in C_b(\mathbb{E})$ de L qui satisfont $Lip(k) \leq \epsilon$ et $Lip(k') \leq \epsilon$. Il existe une unique application $g \in C_b(\mathbb{E})$ qui vérifie*

$$(L+k)(id+g) = (id+g)(L+k').$$

Preuve La Proposition 29 prouve que, puisque $\epsilon\|L^{-1}\| < 1$, $L+k$ et $L+k'$ sont des homéomorphismes que l'on peut donc inverser. L'égalité précédente devient

$$(L+k)(\mathrm{id}+g)(L+k')^{-1} = \mathrm{id}+g$$

ou encore

$$Lg(L+k')^{-1} + kg(L+k')^{-1} + (L+k)(L+k')^{-1} - \mathrm{id} = g$$

qui est une équation de point fixe dans l'espace de Banach $C_b(\mathbb{E})$. Nous allons prouver l'existence d'un tel point fixe g en utilisant la Proposition 31.

Définissons deux applications :

$$\mathcal{L} : C_b(\mathbb{E}) \to C_b(\mathbb{E}) \text{ par } \mathcal{L}(g) = Lg(L+k')^{-1},$$

qui est linéaire, et

$$\mathcal{H} : C_b(\mathbb{E}) \to C_b(\mathbb{E}) \text{ par } \mathcal{H}(g) = kg(L+k')^{-1},$$

et notons $c = (L+k)(L+k')^{-1} - \mathrm{id}$. Nous recherchons un point fixe de $g \to \mathcal{L}(g) + \mathcal{H}(g) + c$. Notons que cette application est bien définie sur $C_b(\mathbb{E})$. En effet $\mathcal{L}(g)$ et $\mathcal{H}(g)$ sont des fonctions bornées sur $\mathbb{E}$ dès que $g \in C_b(\mathbb{E})$, et c est une fonction bornée sur $\mathbb{E}$ d'après la dernière assertion de la Proposition 29.

L'application $\mathcal{L}$ est hyperbolique. Notons que $\mathcal{L}$ est inversible, continue et à inverse continu : $\mathcal{L}^{-1}(g') = k^{-1}g'(L+k')$ qui est de même type que $\mathcal{L}$. L'espace $C_b(\mathbb{E})$ est scindé en

$$C_b(\mathbb{E}) = C_b(\mathbb{E}, E_c) \oplus C_b(\mathbb{E}, E_d) = \mathcal{E}_c \oplus \mathcal{E}_d$$

où $\mathbb{E} = E_c \oplus E_d$ est la décomposition associée à L. Pour une fonction $g \in C_b(\mathbb{E})$, les coordonnées de g dans cette décomposition sont données par $g = p_c \circ g + p_d \circ g$ où p_c et p_d sont les projections de $\mathbb{E}$ sur E_c et E_d associées à la décomposition $\mathbb{E} = E_c \oplus E_d$.

Nous allons vérifier que $\mathcal{L}$ est une contraction sur $\mathcal{E}_c$ et une dilatation sur $\mathcal{E}_d$. Pour g et $g' \in \mathcal{E}_c$ on a :

$$\begin{aligned}\|\mathcal{L}|_{\mathcal{E}_c}(g) - \mathcal{L}|_{\mathcal{E}_c}(g')\| &= \|L(g-g')(L+k')^{-1}\| = \sup_{y\in\mathbb{E}} \|L(g-g')(L+k')^{-1}(y)\| \\ &\leq \sup_{x\in\mathbb{E}} \|L(g-g')(x)\| \leq \|L\| \sup_{x\in\mathbb{E}} \|(g-g')(x)\| \leq \lambda \|g-g'\|,\end{aligned}$$

ce qui prouve que

$$\|\mathcal{L}|_{\mathcal{E}_c}\| \leq \lambda < 1.$$

On procède de même avec $(\mathcal{L}|_{\mathcal{E}_d})^{-1}$ et l'on obtient :

$$\|(\mathcal{L}|_{\mathcal{E}_d})^{-1}\| \leq \lambda < 1.$$

L'application $\mathcal{H}$ est lipschitzienne et vérifie $\mathrm{Lip}(\mathcal{H}) \leq \mathrm{Lip}(k)$. Cela provient de

$$\begin{aligned}\|\mathcal{H}(g) - \mathcal{H}(g')\| &= \sup_{y\in\mathbb{E}} \|kg(L+k')^{-1}(y) - kg'(L+k')^{-1}(y)\| \\ &\leq \sup_{x\in\mathbb{E}} \|kg(x) - kg'(x)\| \leq \mathrm{Lip}(k) \sup_{x\in\mathbb{E}} \|g(x) - g'(x)\| = \mathrm{Lip}(k)\|g-g'\|.\end{aligned}$$

On peut maintenant utiliser la Proposition 31 dont nous venons de vérifier les hypothèses (notons que la condition $\delta \leq r(1-\lambda-\epsilon)$ est automatiquement satisfaite puisque l'on a ici $r = \infty$). On a donc prouvé l'existence d'une unique

fonction $g \in C_b(\mathbb{E})$ telle que

$$(L+k)(\mathrm{id}+g) = (\mathrm{id}+g)(L+k'). \ \square$$

Démonstration du Théorème 28. Montrons que, dans le lemme précédent, $\mathrm{id}+g$ est un homéomorphisme. Commençons par noter $\mathrm{id}+g = G_{k,k'}$ de sorte que

$$(L+k)G_{k,k'} = G_{k,k'}(L+k')$$

et échangeons les rôles de k et k' dans le résultat que l'on vient d'obtenir. Il existe une nouvelle unique fonction $G_{k',k} \in C_b(\mathbb{E})$ telle que

$$(L+k')G_{k',k} = G_{k',k}(L+k).$$

On a :

$$G_{k',k}G_{k,k'}(L+k') = G_{k',k}(L+k)G_{k,k'} = (L+k')G_{k',k}G_{k,k'}$$

et de façon similaire

$$(L+k)G_{k,k'}G_{k',k} = G_{k,k'}(L+k')G_{k',k} = G_{k,k'}G_{k',k}(L+k).$$

Mais puisque les égalités $(L+k)\mathrm{id} = \mathrm{id}(L+k)$ et $(L+k')\mathrm{id} = \mathrm{id}(L+k')$ ont lieu et qu'il y a unicité d'une telle fonction on en déduit que

$$G_{k,k'}G_{k',k} = G_{k',k}G_{k,k'} = \mathrm{id}.$$

Ceci prouve que $G_{k,k'}$ et $G_{k',k}$ sont inverses l'un de l'autre, ce sont donc des homéomorphismes puisque par construction ils sont continus.

On obtient le Théorème 28 en prenant $k' = 0$ et $h = \mathrm{id}+g$. $\square$

Démonstration du Théorème 24. Soit f un difféomorphisme de classe C^1 défini sur un ouvert U de $\mathbb{E}$ et soit x un point fixe hyperbolique de f dans U. On va appliquer le Théorème 28 à la situation $f = Df(x) + k$. Nous allons en vérifier les hypothèses. Tout d'abord $Df(x)$ est continue, inversible et hyperbolique par hypothèse. L'application $k = f - Df(x)$ est de classe C^1 et de dérivée nulle en x. Soit ϵ la constante introduite dans le Théorème 28. Par continuité $\|Dk(y)\| \leq \epsilon/2$ pour tout $y \in \bar{B}(x,r) \subset U$ et pour un $r > 0$ convenable. En utilisant l'inégalité des accroissements finis 6.1.5, on prouve que k est lipschitzienne et bornée sur $\bar{B}(x,r)$ avec $\mathrm{Lip}(k) \leq \epsilon/2$. On peut étendre k à $\mathbb{E}$ tout entier en une fonction $\tilde{k}$ bornée et lipschitzienne en posant par exemple

$$\tilde{k}(y) = k(y) \ \text{ si } \ y \in \bar{B}(x,r) \ \text{ et } \ \tilde{k}(y) = k\left(r\frac{y-x}{\|y-x\|} + x\right) \ \text{ sinon.}$$

La constante de Lipschitz de $\tilde{k}$ vérifie

$$\mathrm{Lip}(\tilde{k}) \leq 2\mathrm{Lip}(k) \leq \epsilon.$$

On applique alors le Théorème 28 à $\tilde{f} = Df(x) + \tilde{k}$, puis on restreint le résultat obtenu à $B(x,r)$. $\square$

2.6 Les variétés stables et instables

Cette section est consacrée à l'étude des ensembles stables et instables associés à un point fixe. Ce sont les équivalents non linéaires des sous-espaces contractés et dilatés d'un endomorphisme hyperbolique.

2.6.1 Définition des ensembles stables et instables

Définition 33. *Soit f un difféomorphisme de classe C^1, défini sur un ouvert U de $\mathbb{E}$ et soit x un point fixe hyperbolique de f dans U. On définit l'ensemble stable de f en x par*

$$V^s(f,x) = \{y \in U : \lim_{k\to\infty} f^k(y) = x\}$$

et l'ensemble instable par

$$V^i(f,x) = \{y \in U : \lim_{k\to-\infty} f^k(y) = x\}.$$

Nous retrouvons pour un f linéaire les concepts de sous-espace contracté pour V^s et de sous-espace dilaté pour V^i.

2.6.2 Le théorème de la variété stable locale

La proposition précédente n'est pas très précise quant à la structure des ensembles $V^s(f,x)$ et $V^i(f,x)$. En effet, l'image par un homéomorphisme d'un sous-espace vectoriel peut être extrêmement irrégulière.

Le théorème de la variété stable locale précise un peu mieux les choses. Ce théorème est dû à O. Perron (1928-1930) ainsi que l'idée de la démonstration basée sur une transformation de graphe.

Définition 34. *Soit f un difféomorphisme, de classe C^1, défini sur un ouvert U de $\mathbb{E}$ et soit x un point fixe hyperbolique de f dans U. La variété stable locale est l'ensemble défini pour tout $r > 0$ par*

$$V^s_r(f,x) = \{y \in \mathbb{E} : \forall n \geq 0 \ f^n(y) \text{ est défini et } \|f^n(y) - x\| < r\}$$

et la variété instable locale est l'ensemble défini par

$$V^i_r(f,x) = \big\{y \in \mathbb{E} : \forall n \geq 0 \ f^{-n}(y) \text{ est défini et } \big\|f^{-n}(y) - x\big\| < r\big\}.$$

Le théorème de la variété stable locale prouve que, lorsque r est assez petit, la variété stable locale est une sous-variété différentiable de $\mathbb{E}$ contenue dans $V^s(f,x)$, de même classe de régularité que f et tangente en x au sous-espace contracté E_c de la dérivée $Df(x)$. Une définition précise des mots «sous-variété différentiable» et «espace tangent» est en appendice mais nous n'aurons pas

besoin de ces concepts pour pour formuler et prouver le théorème de la variété stable locale.

La démonstration que nous donnons ici est différente de celle de Perron. Elle est adaptée des deux articles d'Irwin [26] et [27], ainsi que du livre de Shub [41].

Illustrons ce théorème par l'exemple suivant : soit

$$f : \mathbb{R}^2 \to \mathbb{R}^2, \quad f(x,y) = \begin{pmatrix} x/2 \\ -15x^3/8 - x + 2y \end{pmatrix}.$$

L'origine est un point fixe hyperbolique de f puisque la dérivée en ce point est égale à

$$Df(0,0) = \begin{pmatrix} 1/2 & 0 \\ -1 & 2 \end{pmatrix}.$$

Les sous-espaces contractés et dilatés sont donnés par les deux directions propres associées aux valeurs propres $1/2$ et 2 respectivement :

$$E_c = \{(x,y) : 2x - 3y = 0\}, \quad E_d = \{(x,y) : x = 0\}.$$

Un calcul direct montre que les sous-espaces stables et instables sont décrits par les équations suivantes :

$$V^s = \{(x,y) : x^3 + 2x - 3y = 0\}, \quad V^i = \{(x,y) : x = 0\}$$

et que les variétés stables et instables locales sont

$$V_r^s = \{(x,y) \in V^s : |x| < t\}, \quad V_r^i = \{(x,y) \in V^i : |y| < r\}$$

où t est l'unique réel positif pour lequel $\left\| f(t, (t^3 + 2t)/3) \right\| = r$.

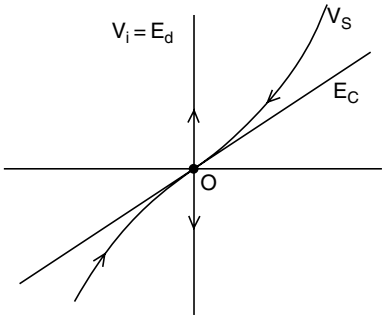


Fig. 2.1.

Les espaces tangents à V_r^s et V_r^i en l'origine sont E_c et E_d. Dans cet exemple f est topologiquement conjugué à $Df(0,0)$ via le changement de variable

$$h(x,y) = \begin{pmatrix} x \\ x^3 + y \end{pmatrix}.$$

Avant de donner l'énoncé du théorème de la variété stable locale, précisons son contexte géométrique.

Soit f un difféomorphisme de classe C^k, $k \geq 1$, défini sur un ouvert U d'un espace de Banach $\mathbb{E}$ et à valeurs dans $\mathbb{E}$. Soit x un point fixe hyperbolique de f. On simplifie l'exposé et les notations en supposant que $x = f(x) = 0$. La dérivée $L = Df(0)$ est un automorphisme hyperbolique. L'espace est donc scindé en somme directe topologique des sous-espaces contractés et dilatés : $\mathbb{E} = E_c \oplus E_d$ que l'on préfère écrire ici comme un produit cartésien : $\mathbb{E} = E_c \times E_d$. Equipons $\mathbb{E}$ de la norme adaptée

$$\|x\| = \max(\|x_d\|, \|x_c\|)$$

comme cela a été fait à la section précédente. La norme d'endomorphisme $\|L\|$ qui est considérée est associée à cette norme vectorielle de sorte que

$$\|Lx\| \leq \|L\|\|x\|.$$

La restriction de L à E_c (en fait à $E_c \times \{0\} \subset E_c \times E_d$) est notée

$$L_c = L|_{E_c} : E_c \to E_c.$$

On définit de même L_d. Puisque L est hyperbolique il existe une constante $\lambda < 1$ telle que

$$\|L_c\| \leq \lambda < 1 \text{ et } \|L_d^{-1}\| \leq \lambda < 1.$$

Enfin $E_c(r)$ est la boule ouverte dans E_c de centre 0 et de rayon r. On définit de même $E_d(r)$ et $E(r) = E_c(r) \times E_d(r)$. On note p_d et p_c les projections de $\mathbb{E}$ sur E_d et E_c ainsi que $f_d = p_d \circ f$ et $f_c = p_c \circ f$.

Pour tout $\varepsilon > 0$, soit $r > 0$ tel que $E(r) \subset U$ et que

$$k = f - L : E(r) \to \mathbb{E}$$

soit lipschitzienne de constante

$$\mathrm{Lip}(f - L) < \varepsilon.$$

La construction de r et de ϵ qui permet d'assurer cette inégalité a déjà été justifiée au cours de la démonstration du Théorème 24. On prend ϵ arbitraire puis on choisit $r > 0$ de sorte que $\|Df(y) - L\| \leq \epsilon$ pour tout $y \in E(r)$. Cette construction est rendue possible par le fait que f est au moins de classe C^1 et que $Df(0) - L = 0$.

Nous allons prouver le théorème suivant :

Théorème 35. *(Théorème de la variété stable locale) Avec les hypothèses ci-dessus, soient $\varepsilon > 0$ et $r > 0$ tels que $f : E(r) \to \mathbb{E}$ vérifie*

$$\mathrm{Lip}(f - L) < \varepsilon < \frac{(1-\lambda)^2}{2-\lambda} < 1 - \lambda.$$

Alors $V_r^s = V_r^s(f, 0)$ est le graphe d'une fonction lipschitzienne

$$g : E_c(r) \to E_d(r)$$

de constante $Lip(g) \leq 1$. Cette fonction est de classe C^k et elle vérifie $g(0) = 0$ et $Dg(0) = 0$. Enfin, $f\big|_{V_r^s}$ est une contraction de constante $\lambda + \varepsilon < 1$ et de point fixe 0.

Remarque 2.

1. Le fait que $f\big|_{V_r^s}$ soit une contraction de point fixe 0 montre que $V_r^s \subset V^s$, autrement dit la variété stable locale est contenue dans l'ensemble stable lorsque r est assez petit.
2. Par échange de f et de f^{-1} on montre le « Théorème de la variété instable locale » : « V_r^i est le graphe d'une application lipschitzienne $h : E_d(r) \to E_c(r)$, de constante de Lipschitz $\mathrm{Lip}(h) \leq 1$. De plus h est de classe C^k et elle vérifie $h(0) = 0$ et $Dh(0) = 0$. Enfin, $f^{-1}\big|_{V_r^i}$ est une contraction de constante $\lambda + \varepsilon < 1$ et de point fixe 0 ».

Corollaire 37. *V_r^s est une sous-variété différentiable de $\mathbb{E}$, de classe C^k, et son espace tangent en 0 est l'espace contracté par $Df(0)$:*

$$T_0 V_r^s = E_c.$$

Preuve C'est une conséquence du théorème précédent et de l'exemple 4 de l'appendice sur les sous-variétés différentiables. □

Corollaire 38. *V_r^i est une sous-variété différentiable de $\mathbb{E}$, de classe C^k, et son espace tangent en 0 est l'espace dilaté par $Df(0)$:*

$$T_0 V_r^i = E_d.$$

2.6.3 Démonstration du théorème de la variété stable

Cette démonstration, extrêmement dense, occupe une dizaine de pages. Elle est découpée en une succession de lemmes.

Lemme 39. *Si $x = (x_c, x_d)$ et $y = (y_c, y_d) \in E(r)$ alors*

$$\|x_c - y_c\| < \|x_d - y_d\|$$

implique

$$\|f_c(x) - f_c(y)\| < (\lambda + \varepsilon)\,\|x_d - y_d\| < (\lambda^{-1} - \varepsilon)\,\|x_d - y_d\| \leq \|f_d(x) - f_d(y)\|\,.$$

Preuve Ecrivons

$$f_c(x) = p_c f(x) = p_c(f - L)(x) + p_c L(x) = p_c(f - L)(x) + L_c(x_c).$$

Ainsi

$$\begin{aligned}\|f_c(x) - f_c(y)\| &\leq \|p_c(f - L)(x) - p_c(f - L)(y)\| + \|L_c(x_c) - L_c(y_c)\| \\ &\leq \varepsilon\,\|x - y\| + \lambda\,\|x_c - y_c\| < (\lambda + \varepsilon)\,\|x_d - y_d\|\,.\end{aligned}$$

Par un argument similaire

$$\begin{aligned}\|f_d(x) - f_d(y)\| &\geq -\|p_d(f - L)(x) - p_d(f - L)(y)\| + \|L_d(x_d) - L_d(y_d)\| \\ &\geq -\varepsilon\,\|x - y\| + \lambda^{-1}\,\|x_d - y_d\| = (\lambda^{-1} - \varepsilon)\,\|x_d - y_d\|\,.\end{aligned}$$

On remarque enfin que

$$\lambda + \varepsilon < 1 < \lambda^{-1} - \varepsilon.\ \square$$

Lemme 40. *V_r^s est le graphe d'une fonction lipschitzienne*

$$g : p_c(V_r^s) \to E_d(r)$$

de constante $Lip(g) \leq 1$. De plus, $f\big|_{V_r^s}$ est une contraction de constante $\lambda + \varepsilon < 1$ et de point fixe 0.

Preuve Soient $x = (x_c, x_d)$ et $y = (y_c, y_d)$ deux points dans V_r^s. Les assertions sur g reviennent à prouver que $\|x_d - y_d\| \leq \|x_c - y_c\|$ et à prendre $g(x_c) = x_d$. Raisonnons par l'absurde. Si $\|x_c - y_c\| < \|x_d - y_d\|$, par le Lemme 39 on a

$$(\lambda^{-1} - \varepsilon)\,\|x_d - y_d\| \leq \|f_d(x) - f_d(y)\|$$

ainsi que

$$\|f_c(x) - f_c(y)\| < \|f_d(x) - f_d(y)\|\,.$$

Puisque $f(x)$ et $f(y) \in V_r^s$ on peut itérer ce raisonnement ce qui donne, par récurrence,

$$(\lambda^{-1} - \varepsilon)^n\,\|x_d - y_d\| \leq \|f_d^n(x) - f_d^n(y)\|\,.$$

Comme $\|f_d^n(x) - f_d^n(y)\| \leq 2r$ et $1 < \lambda^{-1} - \varepsilon$, on doit conclure que $\|x_d - y_d\| = 0$ ce qui contredit notre hypothèse.

Montrons que $f\big|_{V_r^s}$ est une contraction. Nous avons vu que si x et $y \in V_r^s$ alors $\|x_d - y_d\| \leq \|x_c - y_c\|$ et donc

$$\|x - y\| = \|x_c - y_c\|.$$

Puisque $f(V_r^s) \subset V_r^s$ on a aussi

$$\|f(x) - f(y)\| = \|f_c(x) - f_c(y)\|$$

et donc

$$\begin{aligned}\|f(x) - f(y)\| &= \|f_c(x) - f_c(y)\| \\ &\leq \|p_c(f-L)(x) - p_c(f-L)(y)\| + \|L_c(x_c) - L_c(y_c)\| \\ &\leq \|(f-L)(x) - (f-L)(y)\| + \|L_c(x_c) - L_c(y_c)\| \\ &\leq \varepsilon \|x - y\| + \lambda \|x_c - y_c\| = (\varepsilon + \lambda) \|x - y\|.\end{aligned}$$

Puisque $\varepsilon + \lambda < 1$, $f\big|_{V_r^s}$ est une contraction. □

Notons

$$\mathbf{C} = \left\{\gamma = (\gamma_n)_{n\geq 1} : \gamma(n) \in \mathbb{E} \text{ et } \lim_{n\to\infty} \gamma(n) \text{ existe}\right\}.$$

C'est un espace de Banach pour la norme

$$\|\gamma\| = \sup_{n\geq 1} \left(\max\left(\|\gamma_c(n)\|, \|\gamma_d(n)\|\right)\right).$$

On note

$$C(r) = \{\gamma \in \mathbf{C} : \|\gamma\| < r\}.$$

Les espaces $E_c \times E_d \times \mathbf{C}$ et $E_c \times \mathbf{C}$ sont munis des normes $\|(x_c, x_d, \gamma)\| = \max(\|x_c\|, \|x_d\|, \|\gamma\|)$ et $\|(x_c, \gamma)\| = \max(\|x_c\|, \|\gamma\|)$. Ces normes font de ces espaces des espaces de Banach. On définit ensuite

$$\mathbf{F} : E_c \times E_d \times \mathbf{C} \to E_c \times \mathbf{C}$$

par

$$\mathbf{F}(x, \gamma) = \mathbf{F}(x_c, x_d, \gamma) = (x_c, \mathbf{F}_{x_c}(x_d, \gamma))$$

où la suite $\mathbf{F}_{x_c}(x_d, \gamma)$ est définie par

$$\mathbf{F}_{x_c}(x_d, \gamma)(1) = f(x) - \gamma(1),$$

$$\mathbf{F}_{x_c}(x_d, \gamma)(n) = f(\gamma(n-1)) - \gamma(n),\ n \geq 2.$$

Il est clair que cette suite est convergente. Cette définition mystérieuse et son rapport avec g sont justifiés par le lemme suivant :

Lemme 41. *$(x_c, x_d) \in V_r^s$ si et seulement s'il existe $\gamma \in C(r)$ pour lequel $\mathbf{F}(x_c, x_d, \gamma) = (x_c, 0)$.*

Preuve En effet, $\mathbf{F}(x_c, x_d, \gamma) = (x_c, 0)$ si et seulement si $\gamma(n) = f^n(x)$ pour tout $n \geq 1$. □

Si l'on montre que $\mathbf{F}$ est inversible et que l'image de $\mathbf{F}$ contient $E_c(r) \times \{0\}$ on aura montré que

$$g = \Pi_d \mathbf{F}^{-1} \big|_{E_c(r)\times\{0\}}$$

($\Pi_d : E_c \times E_d \times \mathbf{C} \to E_d$ est la projection sur E_d) et que g est défini sur $E_c(r) \times \{0\}$. C'est le programme que nous allons réaliser.

Définissons $\mathbf{L}$ à partir de L de même que $\mathbf{F}$ à partir de f :

$$\mathbf{L} : E_c \times E_d \times \mathbf{C} \to E_c \times \mathbf{C}, \quad \mathbf{L}(x, \gamma) = \mathbf{L}(x_c, x_d, \gamma) = (x_c, \mathbf{L}_{x_c}(x_d, \gamma))$$

où la suite $\mathbf{L}_{x_c}(x_d, \gamma)$ est donnée par

$$\mathbf{L}_{x_c}(x_d, \gamma)(1) = L(x_c, x_d) - \gamma(1),$$

$$\mathbf{L}_{x_c}(x_d, \gamma)(n) = L(\gamma(n-1)) - \gamma(n), \ n \geq 2.$$

Il est facile de voir que cette suite est convergente.

Lemme 42.

1. $\mathbf{L}$ *est un opérateur linéaire et continu et* $\|\mathbf{L}\| \leq 1 + \|L\|$.

2. $\mathbf{F} - \mathbf{L}$ *est lipschitzienne de constante* $\mathrm{Lip}(\mathbf{F} - \mathbf{L}) \leq \mathrm{Lip}(f - L) < \varepsilon$.

3. $\mathbf{F}$ *est dérivable en* 0 *et* $D\mathbf{F}(0) = \mathbf{L}$.

Preuve Puisque

$$\|\mathbf{L}(x_c, x_d, \gamma)\| \leq \max\left(\|x_c\|, \sup_{n\geq 1} \|\mathbf{L}_{x_c}(x_d, \gamma)(n)\|\right),$$

en vertu de la définition de $\mathbf{L}_{x_c}$ on obtient

$$\|\mathbf{L}_{x_c}(x_d, \gamma)(1)\| \leq \|L(x_c, x_d)\| + \|\gamma(1)\| \leq \|L\| \, \|(x_c, x_d)\| + \|\gamma(1)\|,$$

$$\|\mathbf{L}_{x_c}(x_d, \gamma)(n)\| \leq \|L(\gamma(n-1))\| + \|\gamma(n)\| \leq \|L\| \, \|\gamma(n-1)\| + \|\gamma(n)\|, \ n \geq 2,$$

d'où

$$\|\mathbf{L}(x_c, x_d, \gamma)\| \leq (1+\|L\|) \max(\|(x_c, x_d)\|, \sup_{n\geq 1} \|\gamma(n)\|) = (1+\|L\|) \, \|(x_c, x_d, \gamma)\|$$

et ceci prouve que $\mathbf{L}$ est continue et donne sa norme.

Pour prouver que $\mathbf{F} - \mathbf{L}$ est lipschitzienne, il suffit de remarquer que

$$(\mathbf{F} - \mathbf{L})(x_c, x_d, \gamma) = (0, (\mathbf{F}_{x_c} - \mathbf{L}_{x_c})(x_d, \gamma))$$

et que cette suite est donnée par

$$(\mathbf{F}_{x_c} - \mathbf{L}_{x_c})(x_d, \gamma)(1) = (f - L)(x_c, x_d),$$

$$(\mathbf{F}_{x_c} - \mathbf{L}_{x_c})(x_d, \gamma)(n) = (f - L)(\gamma(n-1)),\ n \geq 2.$$

Calculons la dérivée de $\mathbf{F}$ en 0. On a

$$\mathbf{F}(x_c, x_d, \gamma) - \mathbf{F}(0,0,0) - \mathbf{L}(x_c, x_d, \gamma) = (0, \nu)$$

où la suite ν vérifie

$$\nu(1) = f(x_c, x_d) - L(x_c, x_d),$$

$$\nu(n) = f(\gamma(n-1)) - L(\gamma(n-1)),\ n \geq 2.$$

Ceci donne

$$\frac{\|\mathbf{F}(x_c, x_d, \gamma) - \mathbf{F}(0,0,0) - \mathbf{L}(x_c, x_d, \gamma)\|}{\|(x, \gamma)\|}$$

$$\leq \max\left(\frac{\|f(x_c, x_d) - L(x_c, x_d)\|}{\|(x_c, x_d)\|}, \sup_{n \geq 1} \frac{\|f(\gamma(n-1)) - L(\gamma(n-1))\|}{\|\gamma(n)\|}\right).$$

La limite de ce max est nulle lorsque $(x_c, x_d, \gamma) \to (0,0,0)$ parce que $f(0,0) = 0$ et que $Df(0,0) = L$. □

Pour montrer que $\mathbf{F}$ est inversible, nous allons nous inspirer de la Proposition 29. Elle affirme qu'une perturbation $h + k$ d'un homéomorphisme h reste un homéomorphisme si $\mathrm{Lip}(h^{-1})\mathrm{Lip}(k) < 1$. C'est le programme que nous allons suivre avec ici $h = \mathbf{L}$ et $h + k = \mathbf{F}$.

Lemme 43. $\mathbf{L} : E_c \times E_d \times \mathbf{C} \to E_c \times \mathbf{C}$ *est inversible et* $\left\|\mathbf{L}^{-1}\right\| \leq (1 - \lambda)^{-1}$.

Preuve Posons

$$\nu(1) = \mathbf{L}_{x_c}(x_d, \gamma)(1) = L(x_c, x_d) - \gamma(1),$$

$$\nu(n) = \mathbf{L}_{x_c}(x_d, \gamma)(n) = L(\gamma(n-1)) - \gamma(n),\ n \geq 2.$$

Nous devons exprimer x_d et γ en fonction de x_c et ν. Notons $\gamma(n) = (\gamma_c(n), \gamma_d(n))$ et $\nu(n) = (\nu_c(n), \nu_d(n))$, de sorte que

$$\nu_c(1) = L_c(x_c) - \gamma_c(1),$$

$$\nu_d(1) = L_d(x_d) - \gamma_d(1),$$

$$\nu_c(n) = L_c(\gamma_c(n-1)) - \gamma_c(n),\ n \geq 2,$$

$$\nu_d(n) = L_d(\gamma_d(n-1)) - \gamma_d(n),\ n \geq 2.$$

Les équations 1 et 3 donnent

$$\gamma_c(1) = L_c(x_c) - \nu_c(1),$$

$$\gamma_c(n) = L_c(\gamma_c(n-1)) - \nu_c(n),\ n \geq 2,$$

de sorte que

$$\gamma_c(n) = L_c^n(x_c) - \sum_{j=1}^{n} L_c^{n-j}(\nu_c(j)).$$

La quatrième équation donne

$$\gamma_d(n-1) = L_d^{-1}(\nu_d(n) + \gamma_d(n)) = L_d^{-1}\left(\nu_d(n) + L_d^{-1}\left(\nu_d(n+1) + \gamma_d(n+1)\right)\right)$$

et ainsi de suite. Nous obtenons par récurrence

$$\gamma_d(n) = L_d^{-N}(\gamma_d(n+N)) + \sum_{j=1}^{N} L_d^{-j}(\nu_d(n+j)).$$

Puisque $\left\|L_d^{-1}\right\| \leq \lambda < 1$ et que la suite γ est bornée, on a $L_d^{-N}(\gamma_d(n+N)) \to 0$, la série ci-dessus converge et

$$\gamma_d(n) = \sum_{j=1}^{\infty} L_d^{-j}(\nu_d(n+j)).$$

Enfin, de la seconde équation, on obtient par un procédé identique

$$x_d = \sum_{j=1}^{\infty} L_d^{-j}(\nu_d(j)).$$

Nous avons trouvé l'inverse de $\mathbf{L}$: il est donné par

$$\mathbf{L}^{-1} : E_c \times \mathbf{C} \to E_c \times E_d \times \mathbf{C},\ \mathbf{L}^{-1}(x_c, \nu) = (x_c, x_d, \gamma)$$

avec

$$\begin{aligned} x_d &= \textstyle\sum_{j=1}^{\infty} L_d^{-j}(\nu_d(j)), \\ \gamma_c(n) &= L_c^n(x_c) - \textstyle\sum_{j=1}^{n} L_c^{n-j}(\nu_c(j)), \\ \gamma_d(n) &= \textstyle\sum_{j=1}^{\infty} L_d^{-j}(\nu_d(n+j)). \end{aligned}$$

Nous devons estimer la norme de $\mathbf{L}^{-1}$ et montrer que $\gamma \in \mathbf{C}$ c'est à dire que cette suite converge. Commençons par la norme. On a

$$\|x_d\| \leq \|\nu\| \sum_{j=1}^{\infty} \left\|L_d^{-j}\right\| \leq \|\nu\| \frac{\lambda}{1-\lambda},$$

$$\|\gamma_c(n)\| \leq \lambda^n \|x_c\| + \|\nu\| \sum_{j=1}^{n} \left\|\lambda^{n-j}\right\| \leq \|(x_c, \nu)\| \frac{1}{1-\lambda},$$

$$\|\gamma_d(n)\| \leq \|\nu\| \sum_{j=1}^{\infty} \lambda^j \leq \|\nu\| \frac{\lambda}{1-\lambda}$$

ce qui prouve que

$$\left\|\mathbf{L}^{-1}(x_c, \nu)\right\| \leq \|(x_c, \nu)\| \frac{1}{1-\lambda}.$$

Nous devons enfin montrer que γ est une suite convergente, c'est à dire de Cauchy puisque $\mathbb{E}$ est complet. Puisque $\nu \in \mathbf{C}$ c'est une suite congergente et, pour tout $\varepsilon > 0$, il existe un entier N tel que

$$\|\nu(p) - \nu(q)\| \leq \varepsilon$$

dès que p et $q \geq N$. On en déduit que

$$\|\gamma_d(n) - \gamma_d(m)\| \leq \sum_{j=1}^{\infty} \left\| L_d^{-j} \right\| \|\nu_d(n+j) - \nu_d(m+j)\| \leq \sum_{j=1}^{\infty} \lambda^j \varepsilon = \frac{\lambda\varepsilon}{1-\lambda}$$

dès que n et $m \geq N$. Ceci prouve que γ_d est de Cauchy. Passons à γ_c. On a, avec $n \geq m \geq N$,

$$\begin{aligned}\gamma_c(n) - \gamma_c(m) = {} & (L_c^n - L_c^m)(x_c) \\ & - \textstyle\sum_{j=0}^{m-N} L_c^j(\nu_c(n-j) - \nu_c(m-j)) \\ & - \textstyle\sum_{j=m-N+1}^{m-1} L_c^j(\nu_c(n-j) - \nu_c(m-j)) \\ & - \textstyle\sum_{j=m}^{n-1} L_c^j(\nu_c(n-j))\end{aligned}$$

de sorte que

$$\|\gamma_c(n) - \gamma_c(m)\| \leq \|x_c\|\,(\lambda^n + \lambda^m) + \sum_{j=0}^{m-N} \lambda^j \varepsilon + 2\|\nu\| \sum_{j=m-N+1}^{n} \lambda^j$$

et cette quantité est arbitrairement petite dès que n et m sont assez grands. Ainsi γ_c est de Cauchy et le lemme est prouvé. □

Lemme 44. $\mathbf{F} : E_c(r) \times E_d(r) \times C(r) \to E_c \times \mathbf{C}$ *est injective.*

Preuve. Notons que $\mathbf{L}$ est inversible et que $\mathrm{Lip}(\mathbf{L}^{-1}) = \left\|\mathbf{L}^{-1}\right\| \leq (1-\lambda)^{-1}$ par le Lemme 43. De plus, par le Lemme 42, $\mathbf{K} = \mathbf{F} - \mathbf{L}$ est lipschitzienne de constante $\mathrm{Lip}(\mathbf{K}) < \varepsilon$. On a

$$\begin{aligned}\|\mathbf{F}(x,\mu) - \mathbf{F}(y,\nu)\| &= \|\mathbf{K}(x,\mu) - \mathbf{K}(y,\nu) + \mathbf{L}(x,\mu) - \mathbf{L}(y,\nu)\| \\ &\geq \|\mathbf{L}(x,\mu) - \mathbf{L}(y,\nu)\| - \|\mathbf{K}(x,\mu) - \mathbf{K}(y,\nu)\| \\ &> (\|\mathbf{L}^{-1}\|^{-1} - \varepsilon)\|(x,\mu) - (y,\nu)\| \\ &\geq (1 - \lambda - \varepsilon)\|(x,\mu) - (y,\nu)\|.\end{aligned}$$

Ceci prouve que $\mathbf{F}$ est injective puisque $\lambda + \varepsilon < 1$. □

Lemme 45. *Si* $\varepsilon < (1-\lambda)^2/(2-\lambda)$ *alors* $E_c(r) \times \{0\}$ *est contenu dans l'image de* $\mathbf{F}$.

Preuve Notons que $(x_c, 0)$ est dans l'image de $\mathbf{F}$ si et seulement si 0 est dans l'image de $\mathbf{F}_{x_c}$. De plus, $\mathbf{F}_{x_c}$ est une perturbation lipschitzienne de $\mathbf{L}_{x_c}$:

$$\mathbf{F}_{x_c} = \mathbf{L}_{x_c} + \mathbf{F}_{x_c} - \mathbf{L}_{x_c}$$

avec (Lemme 42)

$$\mathrm{Lip}(\mathbf{F}_{x_c} - \mathbf{L}_{x_c}) \leq \mathrm{Lip}(f - L) < \varepsilon.$$

Soit $\mathbf{L}_0 : E_d \times \mathbf{C} \to \mathbf{C}$ l'application linéaire donnée par

$$\begin{aligned} \mathbf{L}_0(x_d, \gamma)(1) &= L(0, x_d) - \gamma(1), \\ \mathbf{L}_0(x_d, \gamma)(n) &= L(\gamma(n-1)) - \gamma(n), \ n \geq 2. \end{aligned}$$

Remarquons que $\mathbf{L}_0$ et $\mathbf{L}_{x_c}$ ne diffèrent que par une translation de sorte que

$$\mathrm{Lip}(\mathbf{F}_{x_c} - \mathbf{L}_0) \leq \mathrm{Lip}(f - L) \leq \varepsilon.$$

Comme dans la preuve du Lemme 43, on montre que $\mathbf{L}_0$ est inversible et que son inverse

$$\mathbf{L}_0^{-1} : \mathbf{C} \to E_d \times \mathbf{C}$$

vérifie $\mathbf{L}_0^{-1}(\nu) = (x_d, \gamma)$ avec

$$\begin{aligned} x_d &= \textstyle\sum_{j=1}^{\infty} L_d^{-j}(\nu_d(j)), \\ \gamma_c(n) &= -\textstyle\sum_{j=1}^{n} L_c^{n-j}(\nu_c(j)), \\ \gamma_d(n) &= \textstyle\sum_{j=1}^{\infty} L_d^{-j}(\nu_d(n+j)), \end{aligned}$$

ainsi que

$$\left\|\mathbf{L}_0^{-1}\right\| \leq (1 - \lambda)^{-1}.$$

Puisque $\mathrm{Lip}(f - L) < \varepsilon < 1 - \lambda$ nous obtenons

$$\begin{aligned} \mathrm{Lip}(\mathbf{L}_0^{-1}\mathbf{F}_{x_c} - \mathrm{id}) &= \mathrm{Lip}(\mathbf{L}_0^{-1}(\mathbf{F}_{x_c} - \mathbf{L}_0)) \\ &\leq \left\|\mathbf{L}_0^{-1}\right\| \mathrm{Lip}(\mathbf{F}_{x_c} - \mathbf{L}_0) \leq (1 - \lambda)^{-1}\varepsilon < 1. \end{aligned}$$

La Proposition 29 et l'estimation ci-dessus montrent que

$$\mathbf{L}_0^{-1}\mathbf{F}_{x_c} = \mathrm{id} + (\mathbf{L}_0^{-1}\mathbf{F}_{x_c} - \mathrm{id})$$

est une petite perturbation lipschitzienne de l'identité. C'est donc un homéomorphisme, son inverse est lipschitzien et

$$\mathrm{Lip}((\mathbf{L}_0^{-1}\mathbf{F}_{x_c})^{-1}) \leq \frac{1}{1 - \frac{\varepsilon}{1-\lambda}}.$$

Le Lemme 30 permet de prouver que

$$\mathbf{L}_0^{-1}\mathbf{F}_{x_c}(E_d(r) \times C(r)) \supset \mathbf{L}_0^{-1}\mathbf{F}_{x_c}(0) + E_d(s) \times C(s)$$

avec $s = r(1 - \frac{\varepsilon}{1-\lambda})$.

Nous allons calculer $\mathbf{L}_0^{-1}\mathbf{F}_{x_c}$. Tout d'abord, $\mathbf{F}_{x_c}(0,0) = \nu$ avec $\nu(1) = f(x_c,0)$ et $\nu(n) = f(0,0) = 0$ pour tout $n \geq 2$. Nous en déduisons que

$$\mathbf{L}_0^{-1}\mathbf{F}_{x_c}(0,0) = (x_d, \gamma)$$

avec

$$\begin{aligned} x_d &= L_d^{-1} f_d(x_c, 0), \\ \gamma_c(n) &= -L_c^{n-1} f_c(x_c, 0), \\ \gamma_d(n) &= 0. \end{aligned}$$

Nous obtenons

$$\begin{aligned} \|f_d(x_c,0)\| &= \|p_d(f-L)(x_c,0)\| \leq \|(f-L)(x_c,0)\| \\ &= \|(f-L)(x_c,0) - (f-L)(0,0)\| < \varepsilon \|x_c\|, \end{aligned}$$

ainsi que

$$\begin{aligned} \|f_c(x_c,0)\| &= \|p_c(f-L)(x_c,0) + L_c(x_c)\| \leq \|(f-L)(x_c,0)\| + \|L_c(x_c)\| \\ &\leq \|(f-L)(x_c,0) - (f-L)(0,0)\| + \|L_c(x_c)\| < (\varepsilon + \lambda)\|x_c\|. \end{aligned}$$

Ces inégalités et les égalités ci-dessus conduisent à

$$\|x_d\| < \lambda\varepsilon \|x_c\| < \lambda\varepsilon r,$$

$$\|\gamma_c(n)\| < \lambda^{n-1}(\varepsilon + \lambda)\|x_c\| < (\lambda + \varepsilon) r,$$

et enfin à

$$\left\|\mathbf{L}_0^{-1}\mathbf{F}_{x_c}(0,0)\right\| < (\lambda + \varepsilon) r.$$

Nous avons vu que l'image de $\mathbf{L}_0^{-1}\mathbf{F}_{x_c}$ contient la boule de centre $\mathbf{L}_0^{-1}\mathbf{F}_{x_c}(0,0)$ et de rayon $s = r(1 - \frac{\varepsilon}{1-\lambda})$. Cette image contiendra $(0,0)$ si

$$(\lambda + \varepsilon) r < r\left(1 - \frac{\varepsilon}{1-\lambda}\right)$$

c'est à dire si $\varepsilon < (1-\lambda)^2/(2-\lambda)$. Si $(0,0)$ est dans l'image de $\mathbf{L}_0^{-1}\mathbf{F}_{x_c}$, c'est que 0 est dans l'image de $\mathbf{F}_{x_c}$ donc que $(x_c, 0)$ est dans l'image de $\mathbf{F}$ et notre lemme est démontré. □

Les lemmes précédents prouvent que :

Lemme 46. *Si* $\mathrm{Lip}(f-L) < \varepsilon < (1-\lambda)^2/2 - \lambda$ *alors* V_r^s *est le graphe d'une fonction lipschitzienne*

$$g : E_c(r) \to E_d(r),\ g = \Pi_d \mathbf{F}^{-1}\big|_{E_c \times \{0\}},$$

de constante $\mathrm{Lip}(g) \leq 1$ *et* $f\big|_{V_r^s}$ *est une contraction de constante* $\lambda + \varepsilon < 1$ *et de point fixe* 0.

Pour terminer la démonstration du théorème de la variété stable nous devons prouver que :

Lemme 47. *Lorsque f est de classe C^k, g est aussi de classe C^k. De plus, $g(0) = 0$ et $Dg(0) = 0$.*

Preuve Pour prouver que g est de classe C^k, il suffit de montrer que $\mathbf{F}^{-1}$ est de classe C^k. En vertu du Théorème d'inversion locale 185, cette propriété se déduit du fait que $\mathbf{F}$ est elle-même de classe C^k et que $D\mathbf{F}(x,\gamma)$ est un isomorphisme pour tout $(x,\gamma) \in E(r) \times C(r)$.

Montrons ce dernier point. Supposons que $\mathbf{F}$ soit de classe C^1. Par les Lemmes 42 et 43

$$\mathrm{Lip}(\mathbf{F} - \mathbf{L}) \leq \mathrm{Lip}(f - L) < \varepsilon < 1 - \lambda \leq \left\|\mathbf{L}^{-1}\right\|^{-1}$$

de sorte que

$$\|D\mathbf{F}(x,\gamma) - \mathbf{L}\| \leq \mathrm{Lip}(\mathbf{F} - \mathbf{L}) < \left\|\mathbf{L}^{-1}\right\|^{-1}.$$

En effet, la norme de la dérivée en un point est majorée par la constante de Lipschitz de la fonction correspondante. Cette inégalité, la Proposition 29 ou sa version linéaire, le Lemme 86, montrent que $D\mathbf{F}(x,\gamma)$ est un isomorphisme c'est à dire un homéomorphisme linéaire.

Un candidat « évident » pour la dérivée de $\mathbf{F}$ en $(x_c, x_d, \gamma) \in E_c(r) \times E_d(r) \times C(r)$ est l'application linéaire

$$\mathbf{\Lambda} : E_c \times E_d \times \mathbf{C} \to E_c \times \mathbf{C},\ \mathbf{\Lambda}(y,\nu) = \mathbf{\Lambda}(y_c, y_d, \nu) = (y_c, \zeta)$$

où la suite ζ est définie par

$$\zeta(1) = Df(x)y - \nu(1),$$

$$\zeta(n) = Df(\gamma(n-1))\nu(n-1) - \nu(n),\ n \geq 2.$$

Justifions cette « évidence ». On a

$$\mathbf{F}(x+y, \gamma+\nu) - \mathbf{F}(x,\gamma) - \mathbf{\Lambda}(y,\nu) = (0,\phi)$$

avec

$$\phi(1) = f(x+y) - f(x) - Df(x)y = \int_0^1 \left(Df(x+ty) - Df(y)\right) y dt,$$

$$\begin{aligned}\phi(n) &= f(\gamma(n-1) + \nu(n-1)) - f(\gamma(n-1)) - Df(\gamma(n-1))\nu(n-1) \\ &= \int_0^1 \left(Df(\gamma(n-1) + t\nu(n-1)) - Df(\gamma(n-1))\right) \nu(n-1) dt,\ n \geq 2,\end{aligned}$$

ce qui donne l'estimation suivante

$$\frac{\|\mathbf{F}(x+y,\gamma+\nu)-\mathbf{F}(x,\gamma)-\mathbf{\Lambda}(y,\nu)\|}{\|(y,\nu)\|}$$
$$\leq \max\left(\int_0^1 \|Df(x+ty)-Df(y)\|\,dt,\right.$$
$$\left.\sup_{n\geq 1}\int_0^1 \|Df(\gamma(n)+t\nu(n))-Df(\gamma(n))\|\,dt\right).$$

Nous devons montrer que les deux expressions à l'intérieur du max ont pour limite 0 lorsque $y\to 0$ pour la première et $\nu\to 0$ pour la seconde. Calculons cette dernière limite, la précédente s'obtient par un argument similaire.

Soit $(\nu^p)_{p\geq 1}$ une suite dans $\mathbf{C}$ qui a pour limite $0\in\mathbf{C}$. Notons

$$N=\overline{\{\nu^p(m): p\geq 1,\ m\geq 1\}},$$
$$G=\overline{\{\gamma(n):\ n\geq 1\}},$$
$$h: G\times N\times[0,1]\to\mathbb{R},\ h(u,v,t)=\|Df(u+tv)-Df(u)\|,$$
$$H: G\times N\to\mathbb{R},\ H(u,v)=\int_0^1 h(u,v,t)dt.$$

Les ensembles G et N sont compacts dans $\mathbb{E}$ (c'est ici qu'il est fondamental d'avoir pris pour γ et ν^p des suites convergentes) et, puisque f est de classe C^1, h est continue et $h(u,0,t)=0$. Il en résulte que H est continue et que $\lim H(u,v)=0$ lorsque $u,v\to 0$. Une référence pour ce type de résultat est, par exemple, Bourbaki [9] Chap. II, Sect. 3, no. 1, Cor. 2. Puisque H est continue sur le compact $G\times N$, elle est uniformément continue sur cet espace de sorte que

$$\lim_{p\to\infty}\sup_{m\geq 1}\sup_{n\geq 1} H(\nu^p(m),\gamma(n))=0$$

ce qu'il fallait démontrer.

Ainsi $\mathbf{F}$ est dérivable et sa dérivée en (x,γ) est $\mathbf{\Lambda}$. Notons que $\mathbf{\Lambda}$ est continue puisque $Df(x)$ est elle-même continue ainsi, $\mathbf{F}$ est de classe C^1.

Pour prouver que $\mathbf{F}$ est de classe C^k, on itère ce raisonnement en calculant les dérivées successives de $\mathbf{F}$ à l'aide de celles de f. Pour $2\leq p\leq k$ la dérivée $p-$ième de $\mathbf{F}$ a pour expression

$$D^p\mathbf{F}(x,\gamma): (E_c\times E_d\times\mathbf{C})^p\to E_c\times\mathbf{C},$$
$$D^p\mathbf{F}(x,\gamma)((y^1,\nu^1),\ldots,(y^p,\nu^p))=(0,\zeta),$$

où la suite ζ est donnée par

$$\zeta(1)=D^pf(x)(y^1,\ldots,y^p),$$
$$\zeta(n+1)=D^pf(\gamma(n))(\nu^1(n),\ldots,\nu^p(n)),\ n\geq 1.$$

Puis on justifie, comme ci-dessus, que

$$\lim \frac{\left\| \left(D^{p-1}\mathbf{F}(x+y^p, \gamma+\nu^p) - D^{p-1}\mathbf{F}(x,\gamma) \right) ((y^1,\nu^1),\ldots,(y^{p-1},\nu^{p-1})) - (0,\zeta) \right\|}{\|(y^p,\nu^p)\|} = 0$$

lorsque $(y^p, \nu^p) \to (0,0)$.

Notons enfin que $g(0) = 0$ puisque $(0,0) \in V_r^s$ et que $Dg(0) = 0$. En effet,

$$Dg(0) = \Pi_d D\mathbf{F}^{-1}(0)\,\big|_{E_c\times\{0\}} = \Pi_d D\mathbf{F}(0)^{-1}\,\big|_{E_c\times\{0\}} = \Pi_d \mathbf{L}^{-1}\,\big|_{E_c\times\{0\}}$$

par le Lemme 42-3. Pour tout $u_c \in E_c$ on a, à l'aide des expressions données dans la preuve du Lemme 43,

$$Dg(0)u_c = \Pi_d \mathbf{L}^{-1}(u_c, 0) = \sum_{j=1}^{\infty} L_d^{-j}(0) = 0.$$

Ceci termine la démonstration de ce lemme ainsi que celle du théorème de la variété stable. □

2.7 Exemples

2.7.1 Calcul de l'inverse d'un nombre

Le calcul de l'inverse d'un nombre réel $a > 0$ revient à résoudre l'équation $ax = 1$. C'est à priori une équation linéaire sauf lorsqu'on l'écrit $1/x = a$ ou $ax^2 - x = 0$.

1. $1/x - a + x = x$ conduit au schéma itératif $x_{k+1} = 1/x_k - a + x_k = f(x_k)$. La dérivée de f en $1/a$ vaut $f'(1/a) = 1 - a^2$ ce qui donne un point fixe attractif dès que $0 < a < 1$.
2. $ax^2 - x = 0$ conduit au schéma itératif $x_{k+1} = 2x_k - ax_k^2 = f(x_k)$. Ici $f'(1/a) = 0$ et la suite des itérés converge quadratiquement.

Il faut remarquer que le second schéma itératif permet de calculer l'inverse d'un nombre en utilisant uniquement des soustractions et des multiplications contrairement au premier schéma qui utilise aussi des divisions.

2.7.2 Calcul des racines carrées

Le calcul de la racine carrée d'un nombre réel $a > 0$ revient à résoudre l'équation $x^2 = a$. Pour en faire une équation de point fixe plusieurs stratégies sont possibles :

1. $x^2 + x - a = x$ conduit au schéma itératif $x_{k+1} = x_k^2 + x_k - a = f(x_k)$. C'est un désastre : la dérivée de f en le point fixe $\sqrt{a}$ vaut $f'(\sqrt{a}) = 2\sqrt{a} + 1 > 1$ ce qui prouve que ce point fixe est répulsif. Si l'on démarre en $x_0 > \sqrt{a}$ la suite (x_k) est croissante jusqu'à l'infini.

2. $a/x = x$ conduit au schéma itératif $x_{k+1} = a/x_k$ qui n'a aucun intérêt puisqu'il fabrique des cycles d'ordre 2 : x_0, a/x_0, x_0, ... La dérivée de $f(x) = a/x$ en le point fixe est égale à $f'(\sqrt{a}) = -1$ qui n'est donc pas attractif.
3. $\frac{1}{2}\left(x + \frac{a}{x}\right) = x$ conduit au schéma itératif $x_{k+1} = \frac{1}{2}\left(x_k + \frac{a}{x_k}\right)$ connu depuis l'antiquité. Sa convergence est quadratique puisque $f(\sqrt{a}) = \sqrt{a}$ et $f'(\sqrt{a}) = 0$.

Nous retrouverons cet exemple lors de l'étude de la méthode de Newton.

2.7.3 Le problème restreint des trois corps

Ce problème décrit le mouvement d'un corps (un satellite ou une comète) de masse négligeable qui se déplace dans le champ gravitationnel d'un ensemble de deux planètes (Soleil-Terre ou bien Terre-Lune ou bien Soleil-Jupiter ...). Les deux corps principaux, notés S et J, sont animés d'un mouvement circulaire plan autour de leur centre de masse commun M avec une vitesse angulaire normalisée à 1. Leur masse totale est aussi normalisée à 1, celle de J est $m_J = \mu$ et celle de S est $m_S = 1 - \mu$. On convient que la masse de S est plus grande que celle de J, autrement dit que $0 < \mu < 1/2$. Pour le système Soleil-Terre $\mu = \frac{m_T}{m_S + m_T} = 3.03591 \times 10^{-6}$ et pour le système Soleil-Jupiter $\mu = 9.537 \times 10^{-4}$. Dans un système de coordonnées en rotation avec S et T et centré au centre de masse, leurs coordonnées sont : $S = (-\mu, 0)$ et $J = (1 - \mu, 0)$.

On considère maintenant un satellite ou une comète, noté C, qui se déplace dans ce champ gravitationnel et on note (x, y) ses coordonnées, $(\dot{x}, \dot{y})$ les coordonnées de son vecteur vitesse et $(\ddot{x}, \ddot{y})$ celles du vecteur accélération. On note aussi r_1 et r_2 les distances Comète-Soleil $(C - S)$ et Comète-Jupiter $(C - J)$:

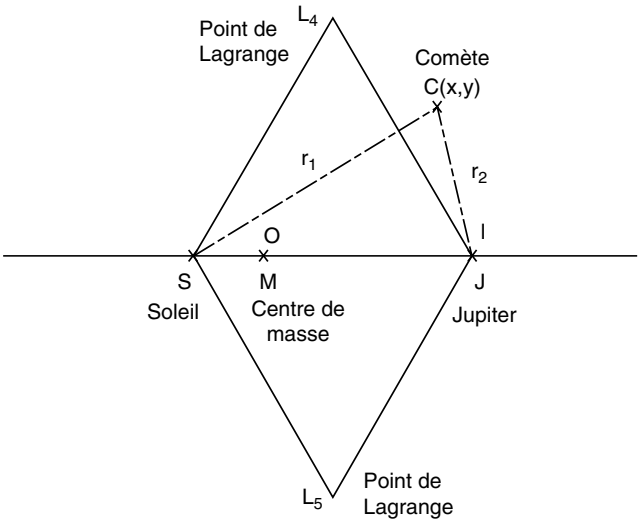
$$r_1 = ((x + \mu)^2 + y^2)^{1/2},$$
$$r_2 = ((x - 1 + \mu)^2 + y^2)^{1/2}.$$

Dans ces coordonnées, le lagrangien du système est donné par

$$\mathcal{L}(x, y, \dot{x}, \dot{y}) = \frac{1}{2}((\dot{x} - y)^2 + (\dot{y} + x)^2) - \mathcal{U}$$

où $\mathcal{U}$ est le potentiel gravitationnel

$$\mathcal{U} = -\frac{1 - \mu}{r_1} - \frac{\mu}{r_2}.$$



Les équations du mouvement dans ce repère (équations d'Euler-Lagrange) sont :

$$\ddot{x} = 2\dot{y} - \frac{\partial \mathcal{V}}{\partial x} = 2\dot{y} - \mathcal{V}_x,$$
$$\ddot{y} = -2\dot{x} - \frac{\partial \mathcal{V}}{\partial y} = -2\dot{x} - \mathcal{V}_y,$$

où $\mathcal{V}$ est le potentiel augmenté :

$$\mathcal{V} = \mathcal{U} - \frac{1}{2}(x^2 + y^2).$$

On fait de ces équations un système différentiel autonome en posant :

$$\begin{aligned} \dot{x} &= u, \\ \dot{y} &= v, \\ \dot{u} &= 2v - \mathcal{V}_x, \\ \dot{v} &= -2u - \mathcal{V}_y \end{aligned}$$

où $\mathcal{V}_x$ et $\mathcal{V}_y$ désignent les dérivées partielles de $\mathcal{V}$ par rapport à x et y. Recherchons les points d'équilibre de ce système. Ce sont, par définition, les solutions « à vitesse nulle » c'est-à-dire ici telles que

$$\begin{aligned} 0 &= u, \\ 0 &= v, \\ 0 &= 2v - \mathcal{V}_x, \\ 0 &= -2u - \mathcal{V}_y. \end{aligned}$$

Ce système devient une équation de point fixe $F(x,y,u,v)=(x,y,u,v)$ si l'on pose

$$\begin{aligned} x &= x+u,\\ y &= y+v,\\ u &= u+2v-\mathcal{V}_x,\\ v &= v-2u-\mathcal{V}_y. \end{aligned}$$

Compte tenu des expressions de $\mathcal{V}_x$ et $\mathcal{V}_y$ on obtient :

$$\begin{aligned} x &= x+u,\\ y &= y+v,\\ u &= u+2v+x-\frac{(1-\mu)(x+\mu)}{r_1^3}-\frac{\mu(x-1+\mu)}{r_2^3},\\ v &= v-2u+y-\frac{(1-\mu)y}{r_1^3}-\frac{\mu y}{r_2^3}. \end{aligned}$$

La dérivée de F est donnée par

$$DF(x,y,u,v)=\begin{pmatrix} 1 & 0 & 1 & 0\\ 0 & 1 & 0 & 1\\ -\mathcal{V}_{xx} & -\mathcal{V}_{xy} & 1 & 2\\ -\mathcal{V}_{yx} & -\mathcal{V}_{yy} & -2 & 1 \end{pmatrix}$$

avec

$$\begin{aligned} \mathcal{V}_{xx} &= -1+(1-\mu)\frac{r_1^2-3(x+\mu)^2}{r_1^5}+\mu\frac{r_2^2-3(x-1+\mu)^2}{r_2^5},\\ \mathcal{V}_{xy}=\mathcal{V}_{yx} &= -3(1-\mu)\frac{(x+\mu)y}{r_1^5}-3\mu\frac{(x-1+\mu)y}{r_2^5},\\ \mathcal{V}_{yy} &= -1+(1-\mu)\frac{(r_1^2-3y^2)}{r_1^5}+\mu\frac{(r_2^2-3y^2)}{r_2^5}. \end{aligned}$$

Le polynôme caractéristique de cette matrice est :

$$\det(DF(x,y,u,v)-\lambda\mathrm{id})=(1-\lambda)^4+(1-\lambda)^2(4+\mathcal{V}_{xx}+\mathcal{V}_{yy})+(\mathcal{V}_{xx}\mathcal{V}_{yy}-\mathcal{V}_{xy}^2).$$

Lagrange et Euler ont montré que ces équations possèdent cinq solutions dont les coordonnées spatiales correspondent à cinq points : L_1, L_2 et L_3 situées sur l'axe Soleil-Jupiter (Euler, 1767), L_4 et L_5 situées symétriquement par rapport à cet axe et formant avec S et J des triangles équilatéraux (Lagrange, 1773). Ce sont les points de Lagrange que nous allons étudier maintenant.

Les solutions équilatérales L_4 et L_5. Nous supposons que $y\neq 0$ de sorte que le système devient

$$
\begin{aligned}
u &= 0,\\
v &= 0,\\
x &= \frac{(1-\mu)(x+\mu)}{r_1^3} + \frac{\mu(x-1+\mu)}{r_2^3},\\
1 &= \frac{(1-\mu)}{r_1^3} + \frac{\mu}{r_2^3}.
\end{aligned}
$$

On en déduit que

$$x = (x+\mu)\left(\frac{1-\mu}{r_1^3} + \frac{\mu}{r_2^3}\right) - \frac{\mu}{r_2^3} = x + \mu - \frac{\mu}{r_2^3}$$

ce qui prouve que $r_2 = 1$ puis, en reportant cette valeur dans la première équation, que $r_1 = 1$. Comme la distance Soleil-Jupiter a été normalisée à 1, on a bien deux triangles équilatéraux : SJL_4 et SJL_5. Les coordonnées de ces points de Lagrange sont :

$$L_4 = \left(\frac{1-2\mu}{2}, \frac{\sqrt{3}}{2}\right) \quad \text{et} \quad L_5 = \left(\frac{1-2\mu}{2}, -\frac{\sqrt{3}}{2}\right).$$

Quelle est la nature de ces points fixes ? Sont-ils attractifs ? Répulsifs ? Hyperboliques ? Lorsque x et y sont les coordonnées de L_4 on obtient $\mathcal{V}_{xx} = -\frac{3}{4}$, $\mathcal{V}_{yy} = -\frac{9}{4}$ et $\mathcal{V}_{xy} = \frac{3\sqrt{3}}{4}(2\mu - 1)$. Le polynôme caractéristique de DF est

$$(1-\lambda)^4 + (1-\lambda)^2 + \frac{27}{4}\mu(1-\mu).$$

On obtient

$$(1-\lambda)^2 = -\frac{1}{2} \pm \frac{1}{2}\sqrt{1 - 27\mu(1-\mu)}.$$

Dans le cas des systèmes Soleil-Jupiter et Soleil-Terre, la quantité $1-27\mu(1-\mu)$ est positive et inférieure à 1 de sorte que $1-\lambda$ est une quantité purement imaginaire et $\lambda = 1 \pm i\alpha$ a un module plus grand que 1. On a donc affaire à un point fixe répulsif. Le cas de L_5 est identique.

Les solutions alignées L_1, L_2 et L_3. Ce sont les solutions pour lesquelles $y = 0$. On a donc $u = v = y = 0$ et la coordonnée x est donnée par l'équation :

$$x = \frac{(1-\mu)(x+\mu)}{|x+\mu|^3} + \frac{\mu(x-1+\mu)}{|x-1+\mu|^3}.$$

Ceci conduit à considérer trois cas qui correspondent à la position respective de la comète, du Soleil et de Jupiter sur l'axe $S-J$. Après disparition des dénominateurs on obtient les équations de degré 5 suivantes :

$$
\begin{aligned}
&x(x+\mu)^2(x-1+\mu)^2 + (1-\mu)(x-1+\mu)^2 + \mu(x+\mu)^2 = 0 \quad \text{si} \quad x < -\mu,\\
&x(x+\mu)^2(x-1+\mu)^2 - (1-\mu)(x-1+\mu)^2\\
&\quad +\mu(x+\mu)^2 = 0 \quad \text{si} \quad -\mu < x < 1-\mu,\\
&x(x+\mu)^2(x-1+\mu)^2 - (1-\mu)(x-1+\mu)^2 - \mu(x+\mu)^2 = 0 \quad \text{si} \quad 1-\mu < x.
\end{aligned}
$$

Chacune de ces équations possède une et une seule racine dans l'intervalle considéré. Prouvons le pour la première équation. On la transforme en l'équation de point fixe

$$x = -\frac{1-\mu}{(x+\mu)^2} - \frac{\mu}{(x-1+\mu)^2}, \quad x < -\mu.$$

L'étude des variations de la fonction décrite dans le membre de droite montre qu'elle ne possède qu'un seul point fixe dans l'intervalle $x < -\mu$.

Dans le cas du système Soleil-Terre, Les points de Lagrange alignés ont pour coordonnées : $L_1 = -1.0000001$, $L_2 = 0.9899092$ et $L_3 = 1.0100701$. Les valeurs propres de DF sont toutes de module > 1. Il s'agit donc de points fixes répulsifs pour F.

2.7.4 Proies et prédateurs

Dans cet exemple nous présentons un modèle d'évolution de deux populations, une de proies x et une de prédateurs y. On note x_k la quantité de proies à l'instant t_k et y_k celui des prédateurs. Le modèle de Volterra et Lokta consiste à supposer que le taux de croissance de ces populations par unité de temps est donné par

$$\frac{x_{k+1} - x_k}{t_{k+1} - t_k} = (A - By_k)x_k,$$

$$\frac{y_{k+1} - y_k}{t_{k+1} - t_k} = (Cx_k - D)y_k,$$

où A, B, C et D sont des constantes positives. Pour des instants régulièrement espacés d'une unité de temps on obtient le modèle :

$$x_{k+1} - x_k = (A - By_k)x_k,$$

$$y_{k+1} - y_k = (Cx_k - D)y_k.$$

On voit que l'augmentation du nombre de proies $x_{k+1} - x_k$ sera d'autant plus grande que le nombre de prédateurs est faible, c'est le sens du terme $A - By_k$, et que le nombre de proies x_k est grand. Une remarque similaire a lieu pour $y_{k+1} - y_k$.

Un équilibre est-il possible ? Les deux populations, suivant ce modèle, peuvent-elles devenir stables ? Il s'agit de trouver les points fixes de

$$F(x,y) = \begin{pmatrix} x + (A - By)x \\ y + (Cx - D)y \end{pmatrix}.$$

Il y en a deux qui sont $(0,0)$ et $(D/C, A/B)$. La dérivée de F est égale à

$$DF(x,y) = \begin{pmatrix} 1 + A - By & -Bx \\ Cy & 1 + Cx - D \end{pmatrix}$$

qui, aux points considérés, vaut

$$DF(0,0) = \begin{pmatrix} 1+A & 0 \\ 0 & 1-D \end{pmatrix} \quad \text{et} \quad DF(D/C, A/B) = \begin{pmatrix} 1 & -BD/C \\ AC/B & 1 \end{pmatrix}.$$

Le premier point fixe, si $D < 2$, est hyperbolique, le second est répulsif puisque les valeurs propres de $DF(D/C, A/B)$ ont un module égal à $(1+AD)^{1/2} > 1$. Donc pas d'équilibre stable possible.

2.8 Les structures topologiques quotient

Nous aurons à considérer des espaces topologiques construits comme suit. E est un ensemble et G est un groupe qui opère à droite sur E. Autrement dit, il existe une application

$$E \times G \to E, \quad (x, g) \to xg,$$

qui vérifie $xe = x$ et $x(gh) = (xg)h$ pour tout $x \in E$, g et $h \in G$ et où e est l'élément neutre du groupe. L'orbite de $x \in E$ est l'ensemble $< x >= xG = \{xg : g \in G\}$. On note E/G l'ensemble des orbites :

$$E/G = \{< x >: x \in E\}.$$

E/G est le quotient de E pour la relation d'équivalence

$$x \equiv y \quad \text{s'il existe} \quad g \in G \quad \text{tel que} \quad y = xg.$$

On note enfin $\pi : E \to E/G$ la surjection canonique : $\pi(x) =< x >$. Voici trois exemples de telles situations.

1. L'espace projectif réel. On le note $\mathbb{P}_{n-1}(\mathbb{R})$, c'est l'ensemble des droites issues de l'origine et contenues dans $\mathbb{R}^n$. Il peut être vu comme l'espace des orbites associées à l'action

$$(\mathbb{R}^n)^* \times \mathbb{R}^* \to (\mathbb{R}^n)^*, \quad (x, \lambda) \to x\lambda.$$

Les orbites sont les droites vectorielles de $\mathbb{R}^n$ privées de l'origine.

2. L'espace projectif complexe. On le note $\mathbb{P}_{n-1}(\mathbb{C})$, c'est l'ensemble des droites complexes issues de l'origine et contenues dans $\mathbb{C}^n$. Il peut être vu comme l'espace des orbites associées à l'action

$$(\mathbb{C}^n)^* \times \mathbb{C}^* \to (\mathbb{C}^n)^*, \quad (x, \lambda) \to x\lambda.$$

3. La sphère. Notons $\mathbb{S}^{n-1}$ la sphère unité dans $\mathbb{R}^n$. Elle peut être décrite comme l'espace des orbites de

$$(\mathbb{R}^n)^* \times \mathbb{R}_+^* \to (\mathbb{R}^n)^*, \quad (x, \lambda) \to x\lambda.$$

Les orbites sont les demi-droites ouvertes de $\mathbb{R}^n$ issues de l'origine. On identifie une telle demi-droite avec son unique point de rencontre avec la sphère unité.

Lorsque E est un espace topologique, E/G hérite de la topologie quotient : les ouverts de E/G pour cette topologie sont les images par π des ouverts A de E saturés pour π c'est-à-dire tels que $A = \pi^{-1}(\pi(A))$. On a les propriétés suivantes, voir par exemple [18] XII. 10.

Lemme 48.

1. *Les fermés de E/G sont les images par π des fermés de E saturés pour π,*
2. *π est continue,*
3. *Soit F un autre espace topologique. Une application $f : E/G \to F$ est continue si et seulement si $f \circ \pi : E \to F$ est continue,*
4. *L'image par π d'un ouvert de E est un ouvert de E/F.*

Il arrive parfois qu'un espace topologique puisse être défini comme espace quotient de plusieurs façons différentes. Par exemple l'espace projectif réel est le quotient de $(\mathbb{R}^n)^*$ par les homothéties mais c'est aussi le quotient de la sphère $\mathbb{S}^{n-1}$ par la relation d'antipodie. La question qui se pose est de savoir sous quelles conditions les structures topologiques quotient sont identiques. Voici un énoncé dans ce sens, voir [8].

Lemme 49. *Soient E un espace topologique, R une relation d'équivalence dans E, $\pi : E \to E/R$ la surjection canonique, F une partie de E et R_F la relation d'équivalence dans F induite par R. Notons h l'application canonique de F/R_F sur $\pi(F)$. S'il existe une application continue $f : E \to F$ telle que $f(x)Rx$ pour tout $x \in E$ alors h est un homéomorphisme de F/R_F sur E/R.*

L'exemple de l'espace projectif réel envisagé ci-dessus entre bien dans ce cadre : la relation d'équivalence induite sur la sphère par « $x \equiv y$ s'il existe $\lambda \neq 0$ avec $y = \lambda x$ » est bien la relation d'antipodie : « $x \equiv y$ si $y = x$ ou $y = -x$ » puisque $\|x\| = \|y\| = 1$. Il suffit de prendre $f(x) = x/\|x\|$ pour voir que les hypothèses du lemme sont satisfaites. Les deux quotients donnent donc la même topologie sur l'espace projectif réel.

Supposons maintenant que E soit muni d'une d'une distance invariante sous l'action de G :

$$d(xg, yg) = d(x, y) \quad \text{pour tout} \quad x, y \in E \quad \text{et} \quad g \in G.$$

Supposons aussi que les orbites $< x >= xG$ soient fermées dans E. Posons alors

$$\delta(< x >, < y >) = \max\left(\sup_{g\in G}\inf_{h\in G} d(xg, yh),\ \sup_{h\in G}\inf_{g\in G} d(xg, yh)\right)$$

la distance de Hausdorff de $< x >$ et $< y >$. On a :

Lemme 50. *Sous les hypothèses ci-dessus*

1. *δ est une distance définissant la topologie de E/G,*
2. *Supposons que les orbites $< x >$ soient compactes dans E. Alors, pour toute suite (x_k) et $x \in E$, $< x_k > \to < x >$ si et seulement s'il existe une suite (g_k) dans G telle que $x_k g_k \to x$ dans E,*
3. *Si E est complet alors E/F est aussi complet.*

Preuve Nous allons prouver que $\delta(< x >, < y >) = d(x, yG) = \inf_{h\in G} d(x, yh)$. On a $d(x, yG) = d(xg, yG)$ pour tout $g \in G$ parce que la distance est invariante et donc $d(x, yG) = \sup_{g\in G} d(xg, yG)$. Par des arguments similaires

$$d(x, yG) = \inf_{h\in G} d(x, yh) = \inf_{h\in G} d(xh^{-1}, y) = d(xG, y) = \sup_{h\in G} d(xG, yh)$$

de sorte que $\delta(< x >, < y >) = d(x, yG)$. Cette dernière inégalité montre aussi que $\delta(< x >, < y >)$ est fini. C'est une distance parce que les ensembles $< x >$ sont fermés dans E. Pour prouver que δ définit la topologie de E montrons que $\pi\left(B_d(x, r)\right) = B_\delta(\pi(x), r)$. D'une part $d(x, y) < r$ implique $d(x, yG) < r$ donc $\delta(< x >, < y >) < r$ et ceci prouve l'inclusion $\subset$, d'autre part, si $\delta(< x >, < y >) < r$ on a $d(x, yG) < r$ et donc $d(x, z) < r$ pour un $z \in yG$ ce qui prouve l'inclusion $\supset$. Ainsi les boules ouvertes de E pour d sont transformées par π en les boules ouvertes de E/G pour δ et tout ceci prouve la première assertion.

Passons à la seconde : soit $g_k \in G$ tel que $\delta(< x_k >, < x >) = d(x, x_k G) = d(x, x_k g_k)$. Un tel g_k existe du fait de la compacité des orbites. On a $d(x, x_k g_k) \to 0$ dès que $\delta(< x_k >, < x >) \to 0$. La réciproque provient de la continuité de π.

Pour prouver que E/F est complet nous partons d'une suite de Cauchy $< x_k > \in E/G$. Nous allons montrer qu'elle contient une sous-suite $< x_{N_k} >$ telle que, pour une suite $g_k \in G$, la suite $x_{N_k} g_k$ soit de Cauchy dans E. Cette dernière converge puisque E est complet et par continuité de π la suite $(< x_k >)$ possède une valeur d'adhérence. C'est donc une suite convergente puisqu'elle est de Cauchy. Nous savons que pour tout $\epsilon > 0$, il existe un entier N_ϵ tel que pour tout $p, q \geq N_\epsilon$ on ait $\delta(< x_p >, < x_q >) < \epsilon$ c'est-à-dire $d(x_p, x_q G) < \epsilon$. Prenons $\epsilon = 1/2^k$ et notons $N_k = N_{1/2^k}$. On suppose que cette suite est strictement croissante, on peut toujours s'y ramener. Pour tout $k \geq 0$ et $q \geq N_k$ on a $d(x_{N_k}, x_q G) < 1/2^k$ de sorte que, pour $q = N_{k+1}$, il existe $g_{k+1} \in G$ tel que $d(x_{N_k}, x_{N_{k+1}} g_{k+1}) < 1/2^k$. Définissons $y_k = x_{N_k} g_k g_{k-1} \dots g_0$. On a $y_k \in x_k G$ et $d(y_k, y_{k+1}) < 1/2^k$ pour tout $k \geq 0$ ce qui prouve bien que (y_k) est de Cauchy. □

2.9 Exemple : valeurs propres et méthode de la puissance

Nous allons étudier la méthode de la puissance pour le calcul de la valeur propre dominante d'une matrice. Soit A une matrice $n \times n$ à coefficients complexes inversible. La méthode de la puissance consiste à calculer la suite des itérés $x_k = A^k x$ où x est un vecteur non nul dans $\mathbb{C}^n$. A cette suite on associe une suite de droites vectorielles qui sont $\overline{x_k} = \mathbb{C}x_k$ et on constate que, en général, cette suite de droites converge vers la direction propre correspondant à la valeur propre de A de plus grand module. Une fois calculée cette direction propre il est facile d'en déduire la valeur propre correspondante.

Nous allons voir que, dans un cadre géométrique adéquat, il s'agit d'un exemple de la méthode des approximations successives.

Définition 51. *On appelle espace projectif $\mathbb{P}_{n-1}(\mathbb{C})$ l'ensemble des droites vectorielles (c'est-à-dire issues de l'origine) contenues dans $\mathbb{C}^n$. La droite passant par $x \neq 0$ est notée*

$$\bar{x} = \{\alpha x : \alpha \in \mathbb{C}\}.$$

Nous définissons une structure métrique sur $\mathbb{P}_{n-1}(\mathbb{C})$ de la façon suivante :

Définition 52. *Pour $\bar{x}$ et $\bar{y} \in \mathbb{P}_{n-1}(\mathbb{C})$ on pose*

$$d(\bar{x}, \bar{y}) = \min_{\lambda \in \mathbb{C}} \frac{\|x - \lambda y\|}{\|x\|}.$$

Cette définition est consistante : si l'on change x et y par des multiples scalaires la valeur du minimum reste inchangée.

Proposition 53. *Les propriétés de d sont les suivantes :*

1. $d(\bar{x}, \bar{y}) = \left(1 - \dfrac{|\langle x, y\rangle|^2}{\|x\|^2 \|y\|^2}\right)^{1/2}$,
2. $0 \leq d(\bar{x}, \bar{y}) \leq 1$,
3. $d(\bar{x}, \bar{y}) = 1$ *si et seulement si les droites $\bar{x}$ et $\bar{y}$ sont orthogonales,*
4. *d est une distance.*

Preuve La première propriété provient du fait que le minimum, dans la définition de d est égal à la distance de x à sa projection orthogonale sur la droite $\bar{y}$ c'est-à-dire $\|x - \frac{\langle x,y\rangle}{\langle y,y\rangle} y\|$. Les seconde et troisième propriétés sont évidentes. Pour prouver que d est une distance la seule difficulté est d'établir l'inégalité du triangle : $d(x, z) \leq d(x, y) + d(y, z)$ où x, y et z sont trois vecteurs non nuls que l'on peut supposer de norme 1. Par une transformation unitaire on se ramène au cas de trois points pris sur sur sphère unité de $\mathbb{R}^3$. Puisque $d(x, y) = d(x, -y)$ on peut toujours supposer que nos trois points sont dans une même hémisphère et enfin supposer que leurs coordonnnées sont $x = (\cos a, 0, \sin a)$, $y = (0, 0, 1)$, $z = (\cos b \cos c, \cos b \sin c, \sin b)$ et que $0 \leq a, b \leq \pi/2$. Il faut alors prouver que

$$\sqrt{1-(\cos a\cos b\cos c+\sin a\sin b)^2}\leq \cos a+\cos b.$$

Comme la plus grande valeur possible pour la racine carrée est obtenue lorsque $\cos c=\pm 1$ il suffit de prouver que

$$1-\cos^2 a\cos^2 b-\sin^2 a\sin^2 b\pm 2\cos a\cos b\sin a\sin b\leq \cos^2 a+\cos^2 b+2\cos a\cos b.$$

On prouve cette dernière inégalité en la scindant en

$$\pm\cos a\cos b\sin a\sin b\leq \cos a\cos b \quad \text{et}$$

$$1-\cos^2 a\cos^2 b-\sin^2 a\sin^2 b\leq \cos^2 a+\cos^2 b. \ \square$$

La propriété suivante, donnée ici sans démonstration, est une conséquence de la Proposition 80.

Proposition 54. *L'espace projectif $\mathbb{P}_{n-1}(\mathbb{C})$ est compact (donc complet).*

Définition 55. *Soit A une matrice $n\times n$ à coefficients complexes inversible. On définit*

$$\bar{A}:\mathbb{P}_{n-1}(\mathbb{C})\to\mathbb{P}_{n-1}(\mathbb{C}) \quad \textit{par} \quad \bar{A}\bar{x}=\overline{Ax}.$$

Cette définition a un sens pour deux raisons. La première est que l'image d'une droite par une application linéaire inversible est une droite. Si l'on voulait considérer des applications non inversibles il faudrait restreindre $\bar{A}$ à une partie de $\mathbb{P}_{n-1}(\mathbb{C})$. La seconde raison est que $\overline{Ax}$ ne dépend pas de x mais bien de $\bar{x}$. La proposition suivante est évidente :

Proposition 56. *Soit A une matrice $n\times n$ à coefficients complexes inversible. Il est équivalent de dire :*

1. *x est un vecteur propre de A,*
2. *$\bar{x}$ est un point fixe de $\bar{A}$.*

Nous allons calculer ces points fixes, c'est-à-dire les vecteurs propres de A, en utilisant la méthode des approximations successives.

Théorème 57. *Supposons que A ait ses valeurs propres de modules distincts. Notons les par module décroissant*

$$|\lambda_1|>|\lambda_2|>\ldots>|\lambda_n|>0$$

et soit $v_1,v_2,\ldots,v_n$ une base de vecteurs propres. Pour tout $x\in\mathbb{C}^n$, $x\neq 0$, la suite des itérés $\overline{x_k}=\bar{A}^k\bar{x}$ converge vers un point fixe de $\bar{A}$ c'est-à-dire une direction propre de A. Le bassin d'attraction de $\bar{v}_i$ pour $\bar{A}$ (respectivement, pour $(\bar{A})^{-1}$) est l'ensemble des

$$x=\sum_{k=i}^{n}\alpha_k v_k \quad (\textit{respectivement,} \quad x=\sum_{k=1}^{i}\alpha_k v_k)$$

avec $\alpha_k\in\mathbb{C}$ et $\alpha_i\neq 0$ et $\bar{A}$ (respectivement, $(\bar{A})^{-1}$) laisse cet ensemble invariant. Le bassin d'attraction de $\bar{v}_1$ est un ouvert dense.

Preuve Soit $\bar{x} \in \mathbb{P}_{n-1}(\mathbb{C})$. Il existe i, $1 \leq i \leq n$, pour lequel

$$x = \alpha_i v_i + \ldots + \alpha_n v_n$$

avec $\alpha_i \neq 0$. On a

$$A^k x = \lambda_i^k \alpha_i v_i + \ldots + \lambda_n^k \alpha_n v_n = \lambda_i^k \alpha_i \left(v_i + \sum_{j=i+1}^{n} \left(\frac{\lambda_j}{\lambda_i} \right)^k \frac{\alpha_j}{\alpha_i} v_j \right) = \lambda_i^k \alpha_i (v_i + w_k)$$

et $w_k \to 0$ lorsque $k \to \infty$. On en déduit que

$$d(A^k x, v_i) = d(\lambda_i^k \alpha_i (v_i + w_k), v_i) = d(v_i + w_k, v_i) \to d(v_i, v_i) = 0$$

lorsque $k \to \infty$. Ainsi $\bar{x}$ est dans le bassin d'attraction de $\bar{v}_i$ et $\mathbb{P}_{n-1}(\mathbb{C})$ est l'union disjointe des différents bassins. Il est clair que ces bassins sont invariants par $\bar{A}$. La propriété relative à $(\bar{A})^{-1}$ se prouve en transposant les rôles de A et de A^{-1}.

Montrons enfin que le bassin d'attraction de $\bar{v}_1$ est un ouvert dense. Ce bassin est l'image par la surjection canonique $x \in \mathbb{C}^n \to \bar{x} \in \mathbb{P}_{n-1}(\mathbb{C})$ de l'ensemble des vecteurs $x \in \mathbb{C}^n$ qui s'écrivent $x = \alpha_1 v_1 + \ldots + \alpha_n v_n$ avec $a_1 \neq 0$. Cet ensemble est ouvert et dense dans $\mathbb{C}^n$ donc aussi son image. □

Ce théorème décrit les variétés stables et instables associées à $\bar{A}$ et aux différents points fixes. $\bar{v}_1$ est un point fixe attractif, $\bar{v}_n$ est répulsif, les autres sont « hyperboliques ». Notons que ce concept, que nous n'avons introduit que dans le cadre des espaces vectoriels, s'étend aux variétés différentiables comme ici l'espace projectif $\mathbb{P}_{n-1}(\mathbb{C})$.

L'implémentation de cette méthode se réalise dans $\mathbb{C}^n$. Pour éviter overflow ou underflow on normalise les vecteurs à chaque étape ce qui conduit à poser

$$x_{k+1} = \frac{Ax_k}{\|Ax_k\|}, \quad x_0 = x$$

où $x \in \mathbb{C}^n$ est donné. Cette suite converge, pour presque tout x, vers le vecteur propre v_1. La vitesse de convergence est linéaire et mesurée par le rapport $\left| \frac{\lambda_2}{\lambda_1} \right|$. Attention ! La convergence a lieu dans $\mathbb{P}_{n-1}(\mathbb{C})$ et pas nécessairement dans $\mathbb{C}^n$. Toutefois, dans le cas réel, normaliser Ax_k revient à choisir entre deux points antipodaux pris sur une sphère et l'on peut récupérer la convergence dans $\mathbb{R}^n$ de la suite (x_k) en contrôlant les signes des coordonnées.

2.10 Exemple : calcul simultané des valeurs propres par l'algorithme QR

Donnons nous une matrice A, $n \times n$, réelle ou complexe, inversible. Dans cette section nous allons analyser trois algorithmes de calcul des valeurs propres

d'une matrice A : QR, LR et la méthode de Cholesky lorsque A est définie positive. Nous allons voir qu'ils sont trois réalisations différentes d'un même algorithme géométrique.

Dans ce qui suit nous notons $\mathbb{H}_n^+$ l'ensemble des matrices définies positives, $\mathbb{R}_n$ le groupe des matrices inversibles et triangulaires supérieures, $\mathbb{R}_n^+$ l'ensemble des matrices triangulaires supérieures à diagonale positive, $\mathbb{U}_n$ le groupe unitaire, $\mathbb{T}_n$ le groupe des matrices unitaires et diagonales (leurs termes diagonaux sont 1 ou -1 dans le cas réel et des nombres complexes de module 1 dans le cas complexe).

2.10.1 Les décompositions QR et de Cholesky

Définition 58. *On appelle décomposition QR de A une identité $A = QR$ avec Q orthogonale dans le cas réel, unitaire dans le cas complexe et R triangulaire supérieure et inversible.*

Proposition 59. *Toute matrice A inversible possède une décomposition QR. Il n'y a pas unicité d'une telle décomposition : $Q_1R_1 = Q_2R_2$ si et seulement s'il existe $T \in \mathbb{T}_n$ telle que $Q_2 = Q_1T^*$ et $R_2 = TR_1$. Il existe une unique décomposition $A = QR$ telle que R ait des termes diagonaux strictement positifs.*

Preuve Une telle décomposition peut s'obtenir par la méthode d'orthonormalisation de Gram-Schmidt appliquée aux colonnes de la matrice A. Ceci prouve l'existence de la décomposition. Si $Q_1R_1 = Q_2R_2$ alors $Q_2^{-1}Q_1 = R_2R_1^{-1}$ qui est une matrice à la fois triangulaire supérieure et unitaire. Une telle matrice est nécessairement diagonale. Enfin, il faut noter qu'il n'y a qu'une seule manière de rendre positif un nombre complexe non nul en le multipliant par un nombre complexe de module 1. Ceci prouve la dernière assertion. □

On peut relier la décomposition QR de A à la décomposition de Cholesky de A^TA que nous allons décrire :

Définition 60. *Soit B une matrice $n \times n$ définie positive. On appelle décomposition de Choleski de B une identité $B = R^*R$ avec R triangulaire supérieure à termes diagonaux strictement positifs.*

Proposition 61. *Toute matrice B définie positive possède une décomposition de Choleski. Cette décomposition est unique.*

Preuve Si $B = R_1^*R_1 = R_2^*R_2$ avec $(R_1)_{ii}$ et $(R_2)_{ii} > 0$ alors $R_2^{-*}R_1^* = R_2R_1^{-1}$. Cette matrice étant à la fois triangulaire inférieure et triangulaire supérieure est diagonale. Les termes diagonaux sont égaux à $(R_1)_{ii}/(R_2)_{ii} = (R_2)_{ii}/(R_1)_{ii}$ et sont positifs donc égaux à 1. Ainsi $R_2R_1^{-1} = I_n$ et ceci

prouve que $R_2 = R_1$. L'existence se prouve par récurrence. Le résultat est évident pour $n = 1$. Ecrivons

$$B = \begin{pmatrix} B_{n-1} & b \\ b^* & \beta \end{pmatrix} = \begin{pmatrix} R_{n-1}^* & 0 \\ r^* & \gamma \end{pmatrix} \begin{pmatrix} R_{n-1} & r \\ 0 & \gamma \end{pmatrix}$$

où $R_{n-1}^* R_{n-1}$ est la décomposition de Choleski de B_{n-1} ; cette matrice, obtenue en supprimant de B la dernière ligne et la dernière colonne, est définie positive et, par l'hypothèse de récurrence, elle possède une décomposition de Choleski. L'égalité précédente suppose que $R_{n-1}^* r = b$ et $r^* r + \gamma^2 = \beta$ ce qui détermine $r = R_{n-1}^{-*} b$ et $\gamma^2 = \beta - r^* r$. Il reste à prouver que l'on peut prendre $\gamma > 0$. Cela résulte de l'équation

$$0 < \det B = \det R_{n-1}^* \det R_{n-1} \ \gamma^2 = |\det R_{n-1}|^2 \gamma^2$$

qui prouve que $\gamma^2 > 0$. On peut donc prendre $\gamma > 0$. □

Proposition 62. *La décomposition de Choleski est une application bijective, de classe C^∞ ainsi que son inverse.*

Preuve Notons $Ch : \mathbb{H}_n^+ \to \mathbb{R}_n^+$ l'application qui à la matrice $B \in \mathbb{H}_n^+$ associe $R \in \mathbb{R}_n^+$ telle que $B = R^* R$. Nous venons de voir que Ch est bijective, la bijection réciproque est $Ch^{-1}(R) = R^* R$. Notons que Ch^{-1} est de classe C^∞. La dérivée de Ch^{-1} est donnée par

$$DCh^{-1}(R) : \mathbb{R}_n \to \mathbb{H}_n, \quad DCh^{-1}(R)(S) = S^* R + R^* S.$$

Nous allons prouver que $DCh^{-1}(R)$ est un isomorphisme. On en déduira, par application du théorème d'inversion locale 185, que Ch^{-1} possède un inverse C^∞. Comme les espaces $\mathbb{R}_n$ et $\mathbb{H}_n$ ont même dimension, il suffit de prouver que le noyau de $DCh^{-1}(R)$ est nul. Si $DCh^{-1}(R)(S) = 0$ on a $S^* R + R^* S = 0$ c'est-à-dire que $R^* S = -(R^* S)^*$. Cette matrice est donc antihermitienne. Les entrées de sa première colonne sont $(R^* S)_{i1} = R_{i1}^* S_{11}$. Le premier de la liste est $(R^* S)_{11} = R_{11} S_{11}$ qui est nul puisque $R^* S$ est anti-hermitienne. Donc $S_{11} = 0$ et par suite $(R^* S)_{i1} = 0$ pour tout i. Comme $R^* S$ est anti-hermitienne la première ligne est aussi nulle ; en continuant ainsi avec les autres colonnes et lignes on prouve que $R^* S = 0$ et donc $S = 0$ puisque R est inversible. □

Nous pouvons maintenant relier les décompositions de Choleski et QR. Le résultat suivant est une conséquence immédiate de la proposition précédente, nous le donnons sans démonstration.

Proposition 63. *Notons $A = Q_A R_A$ la décomposition QR de A telle que $R_A \in \mathbb{R}_n^+$. R_A est la décomposition de Cholesky de $A^* A$. Les applications $A \in \mathbb{GL}_n \to \mathbb{R}_A \in \mathbb{R}_n^+$ et $A \in \mathbb{GL}_n \to \mathbb{Q}_A \in \mathbb{U}_n$ sont de classe C^∞.*

2.10.2 La décomposition de Schur

Définition 64. *On appelle décomposition de Schur de A une identité $A = QRQ^*$ avec Q orthogonale dans le cas réel, unitaire dans le cas complexe et R triangulaire supérieure et inversible.*

Proposition 65. *Toute matrice A possède une décomposition de Schur.*

Preuve Cela se prouve par récurrence sur la taille de la matrice. Pour $n = 1$ il n'y a rien à démontrer. Le passage de $n-1$ à n se fait comme suit : on se donne une valeur propre et un vecteur propre associé : $Av = \lambda v$ ainsi qu'une matrice unitaire Q dont la première colonne est v : $Q = (\, v \; P \,)$. On a alors

$$Q^*AQ = \begin{pmatrix} v^*Av & v^*AP \\ P^*Av & P^*AP \end{pmatrix}.$$

Notons que $P^*Av = \lambda P^*v = 0$ puisque Q est unitaire. Si l'on introduit une décomposition de Schur de $P^*AP = Q_1R_1Q_1^*$ on obtient

$$Q^*AQ = \begin{pmatrix} v^*Av & v^*AP \\ 0 & Q_1R_1Q_1^* \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & Q_1 \end{pmatrix} \begin{pmatrix} v^*Av & v^*APQ_1 \\ 0 & R_1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & Q_1^* \end{pmatrix}.$$

Notons que la matrice $\begin{pmatrix} 1 & 0 \\ 0 & Q_1 \end{pmatrix}$ est unitaire et que $\begin{pmatrix} v^*Av & v^*APQ_1 \\ 0 & R_1 \end{pmatrix}$ est triangulaire supérieure d'où la conclusion. □

Pour une matrice A réelle, les matrices Q et R d'une décomposition de Schur ne seront réelles que si les valeurs propres de A sont elles-mêmes réelles.

2.10.3 La variété des drapeaux

Nous allons maintenant relier cette décomposition au concept géométrique de drapeau.

Définition 66. *Un drapeau F est un $n+1-$uplet de sous-espaces vectoriels de $\mathbb{C}^n$, $F = F_0 \subset F_1 \subset \ldots \subset F_n$, avec $\dim F_i = i$. L'espace des drapeaux est noté $\mathbb{F}_n$.*

Un drapeau peut être décrit à l'aide d'une matrice $X \in \mathbb{GL}_n$: $F_0 = \{0\}$ et F_i est le sous-espace vectoriel de $\mathbb{C}^n$ engendré par les colonnes $X_1, \ldots, X_i$ de X. Deux matrices X et Y décrivent le même drapeau si et seulement s'il existe une matrice triangulaire supérieure et inversible R telle que $Y = XR$. Nous résumons cela dans la proposition suivante :

Proposition 67. *La variété des drapeaux $\mathbb{F}_n$ s'identifie à l'espace des orbites de l'action suivante du groupe $\mathbb{R}_n$ sur le groupe linéaire :*

$$\mathbb{GL}_n \times \mathbb{R}_n \to \mathbb{GL}_n, \quad (X, R) \to XR$$

c'est-à-dire au quotient $\mathbb{GL}_n/\mathbb{R}_n$ *de* $\mathbb{GL}_n$ *par la relation d'équivalence*

$$X \equiv Y \quad \textit{si et seulement si} \quad \exists R \in \mathbb{R}_n \quad Y = XR.$$

La classe de la matrice X *est l'ensemble* $< X >= X\mathbb{R}_n$. *Il y a toujours dans cette classe une matrice unitaire donnée par une décomposition* QR *de* X.

On peut aussi décrire un drapeau F par une matrice unitaire (orthogonale dans le cas réel) dont les colonnes constituent une base orthonormée du drapeau. Deux telles matrices U et V donnent le même drapeau si et seulement s'il existe une matrice triangulaire supérieure et inversible T telle que $V = UT$. Cette matrice est nécessairement diagonale et unitaire donc $T \in \mathbb{T}_n$. On vient de prouver que

Proposition 68. *La variété des drapeaux* $\mathbb{F}_n$ *s'identifie au quotient* $\mathbb{U}_n/\mathbb{T}_n$ *du groupe unitaire* $\mathbb{U}_n$ *par la relation d'équivalence*

$$U \equiv V \quad \textit{si et seulement si} \quad \exists T \in \mathbb{T}_n \quad V = UT.$$

La classe de la matrice U *est l'ensemble* $< U >= U\mathbb{T}_n$.

2.10.4 La structure topologique de la variété des drapeaux

Cette structure topologique est déduite de sa description d'espace quotient : $\mathbb{F}_n = \mathbb{GL}_n/\mathbb{R}_n = \mathbb{U}_n/\mathbb{T}_n$.

Lemme 69. *Les deux structures quotient* $\mathbb{GL}_n/\mathbb{R}_n$ *et* $\mathbb{U}_n/\mathbb{T}_n$ *définissent sur* $\mathbb{F}_n$ *la même topologie.*

Preuve C'est une conséquence du lemme 49. On y prend $E = \mathbb{GL}_n$, $F = \mathbb{U}_n$ et pour $f : E \to F$ l'application $X \in \mathbb{GL}_n \to Q_X \in \mathbb{U}_n$ donnée par la décomposition QR : $X = Q_X R_X$. Cette application est continue par la Proposition 63. □

Puisque $\mathbb{F}_n$ est muni de cette topologie quotient, par le Lemme 48 et puisque $\mathbb{U}_n$ est compact on a :

Proposition 70. $\mathbb{F}_n$ *est un espace compact, l'application* $X \in \mathbb{GL}_n \to < X > \in \mathbb{F}_n$ *est continue. De plus, l'image d'un ouvert de* $\mathbb{GL}_n$ *par cette application est un ouvert de* $\mathbb{F}_n$.

Enfin, par le Lemme 49 et puisque $\mathbb{U}_n$ est compact on a :

Proposition 71. *Notons* (F_k) *une suite de drapeaux et* P_k *une matrice unitaire telle que* $F_k =< P_k >$. *Soient* $F \in \mathbb{F}_n$ *et* $P \in \mathbb{U}_n$ *avec* $F =< P >$. *Une condition nécessaire et suffisante pour que* $F_k \to F$ *est qu'il existe des matrices* $T_k \in \mathbb{T}_n$ *telles que* $P_k T_k \to P$.

2.10.5 L'action de A sur la variété des drapeaux

L'opérateur $A \in \mathbb{GL}_n$ définit une action $A_\sharp : \mathbb{F}_n \to \mathbb{F}_n$ sur cet espace de la façon suivante : à tout drapeau $F = F_0 \subset F_1 \subset \ldots \subset F_n$ on associe le drapeau image $A_\sharp(F) = A(F_0) \subset A(F_1) \subset \ldots \subset A(F_n)$. Du point de vue matriciel $A_\sharp(< X >) =< AX >$.

L'intérêt d'introduire ce nouvel opérateur réside est décrit dans la proposition suivante :

Proposition 72. *Un drapeau $F =< Q >$ avec Q unitaire est un point fixe de $A_\sharp$ si et seulement si on peut écrire $A = QRQ^*$ avec R triangulaire supérieure. Autrement dit, les points fixes de $A_\sharp$ sont associés à ses décompositions de Schur.*

Preuve La condition de point fixe $A_\sharp < Q >=< Q >$ signifie que les matrices AQ et Q définissent le même drapeau. Donc il existe R triangulaire supérieure et inversible telle que $AQ = QR$ c'est à dire $A = QRQ^*$. □

Nous reprenons l'idée de calculer de tels points fixes par la méthode des approximations successives. On va prouver le théorème suivant dû à Shub et Vasquez 1987 [47] dont nous suivons la démonstration.

Théorème 73. *(Shub-Vasquez) Supposons que A ait des valeurs propres de modules distincts. Alors $A_\sharp$ possède $n!$ points fixes. Pour tout drapeau F, la suite*

$$F_0 = F,\ F_{k+1} = A_\sharp(F_k)$$

converge. Le bassin d'attraction de l'un de ces points fixes est ouvert et dense dans $\mathbb{F}_n$.

Preuve Puisque les valeurs propres de A ont des modules distincts, A est diagonalisable : $A = MDM^{-1}$ avec $D = \mathrm{Diag}(\lambda_1, \ldots, \lambda_n)$ les valeurs propres étant rangées par module décroissant :

$$|\lambda_1| > \ldots > |\lambda_n| > 0.$$

De plus,

$$A_\sharp = (MDM^{-1})_\sharp = M_\sharp D_\sharp (M^{-1})_\sharp = M_\sharp D_\sharp (M_\sharp)^{-1}$$

de sorte que la dynamique de $A_\sharp$ se déduit de celle de $D_\sharp$ pour laquelle nous allons établir le théorème.

Soit Σ le sous-groupe de $\mathbb{GL}_n$ constitué par les $n!$ matrices de permutation. L'ensemble

$$\{< P >: P \in \Sigma\} \subset \mathbb{F}_n$$

est constitué de $n!$ éléments distincts dans $\mathbb{F}_n$. En effet $< P_1 >=< P_2 >$ si et seulement si $P_1 = P_2 R$ avec R triangulaire supérieure. Comme $\mathbb{R}_n \cap \Sigma = \{I_n\}$ on a $P_1 = P_2$.

Pour voir que $< P >$ pour $P \in \Sigma$ est un point fixe on note que

$$D_\sharp(< P >) =< DP >=< P(P^{-1}DP) >= P(P^{-1}DP)\mathbb{R}_n = P\mathbb{R}_n =< P >$$

parce que $P^{-1}DP$ est une matrice diagonale.

Notons

$$W^s(< P >) = \{< X >: \lim_{k\to\infty} D_\sharp^k(< X >) =< P >\}$$

et $\mathbb{L}_n$ l'ensemble des matrices triangulaires inférieures et inversibles. Nous allons prouver que

$$\mathbb{L}_n < P >= \{< LP >: L \in \mathbb{L}_n\} \subset W^s(< P >).$$

En effet

$$D_\sharp(< LP >) = DLP\mathbb{R}_n = (DLD^{-1})DP\mathbb{R}_n = (DLD^{-1})P\mathbb{R}_n =< DLD^{-1}P >$$

et par récurrence

$$D_\sharp^k(< LP >) =< D^kLD^{-k}P > .$$

Mais $(D^kLD^{-k})_{ij} = (\lambda_i/\lambda_j)^k L_{ij}$ de sorte que

1. $(D^kLD^{-k})_{ij} = 0$ lorsque $j > i$ puisque L est triangulaire inférieure,
2. $(D^kLD^{-k})_{ij} \to 0$ lorsque $j < i$ parce que $|\lambda_i/\lambda_j| < 1$,
3. $(D^kLD^{-k})_{ii} = L_{ii}$.

Ceci prouve, en utilisant la Proposition 70 que $\lim_{k\to\infty} D_\sharp^k(< LP >) =< D'P >$ avec $D' = \mathrm{Diag}(L_{ii})$. En conséquence

$$\lim_{k\to\infty} D_\sharp^k(< LP >) = P(P^{-1}D'P)\mathbb{R}_n = P\mathbb{R}_n =< P >$$

de sorte que $< LP > \in W^s(< P >)$.

Nous allons démontrer maintenant que

$$\mathbb{F}_n = \bigcup_{P\in\Sigma} \mathbb{L}_n < P > .$$

Cela résulte de l'égalité

$$\mathbb{GL}_n = \mathbb{L}_n \Sigma \mathbb{R}_n :$$

toute matrice inversible peut sécrire LPR avec $L \in \mathbb{L}_n$, $P \in \Sigma$ et $R \in \mathbb{R}_n$. Prouver ce résultat demande un peu d'attention. Soit $B \in \mathbb{GL}_n$. Supposons que les lignes L_i et L_j de B^{-1}, $i < j$, se terminent par le même nombre de zéros. Alors, en additionnant à L_i un multiple convenable de L_j on peut augmenter le nombre de zéros terminaux de L_i d'au moins une unité. Cette opération revient à mutiplier B^{-1} à gauche par une matrice triangulaire supérieure à diagonale unité convenable. Si l'on répète cette opération autant que faire se peut, on arrive à une matrice C dont les lignes ont des

nombres différents de zéros terminaux. Une telle matrice est du type $C = PL$ avec P matrice de permutation et L triangulaire inférieure. Ainsi il existe U triangulaire supérieure à diagonale unité telle que $UB^{-1} = PL$ c'est à dire

$$B = L^{-1}P^{-1}U \in \mathbb{L}_n \Sigma \mathbb{R}_n.$$

On a obtenu

$$\mathbb{F}_n = \bigcup_{P \in \Sigma} W^s(< P >),$$

cette union est disjointe et

$$W^s(< P >) = \mathbb{L}_n < P >.$$

Pour conclure il faut prouver que l'un de ces ensembles est ouvert et dense dans $\mathbb{F}_n$. Ceci provient du fait que, dans la plupart des cas, la construction de U, P et L peut se faire avec $P = I_n$ de sorte que, $\mathbb{L}_n\mathbb{R}_n$ est ouvert et dense dans $\mathbb{GL}_n$. Cette affirmation résulte par exemple du lemme suivant : *$A \in \mathbb{GL}_n$ possède une décomposition LU, c'est-à-dire $A = LU \in \mathbb{L}_n\mathbb{R}_n$, si et seulement si ses mineurs principaux sont non nuls* voir [55] Proposition. L'ensemble défini dans $\mathbb{GL}_n$ par des conditions de nullité sur les mineurs est fermé et son complémentaire dense. En utilisant la Proposition 70 on déduit que $\mathbb{L}_n < I_n >= W^s(< I_n >)$ est ouvert et dense dans $\mathbb{F}_n$. Le point fixe qui correspond à $P = I_n$ est le drapeau $F = F_0 \subset \ldots \subset F_n$ où F_k est le sous-espace engendré par les vecteurs propres associés aux valeurs propres $\lambda_1, \ldots, \lambda_k$ c'est-à-dire les k valeurs propres de plus grand module. □

2.10.6 L'algorithme QR de Francis

Soit $A \in \mathbb{GL}_n$. L'algorithme QR, pour le calcul de toutes les valeurs propres de A, est dû à Francis 1961 [19] et Kublanovskaya 1961 [32]. Il est défini de la façon suivante : on construit deux suites $Q_k \in \mathbb{U}_n$ et $R_k \in \mathbb{R}_n$ par :

$$A = Q_1R_1 \text{ et } A_{k+1} = R_kQ_k = Q_{k+1}R_{k+1},$$

à chaque étape on calcule la décomposition QR de la matrice définie par $A_{k+1} = R_kQ_k$.

Proposition 74. *Soit $A \in \mathbb{GL}_n$ dont les valeurs propres ont des modules distincts. Notons (A_k) la suite produite par la méthode QR. Lorsque $k \to \infty$ la diagonale de A_k converge vers l'ensemble des valeurs propres de A et les éléments de la partie triangulaire inférieure stricte convergent vers* 0. *Il n'y a pas nécessairement convergence des éléments de la partie triangulaire supérieure stricte : leurs modules convergent mais pas leurs arguments.*

Preuve Comme les matrices Q_k sont unitaires on obtient

$$A_{k+1} = R_k Q_k = Q_k^* Q_k R_k Q_k = Q_k^* A_k Q_k = \ldots = P_k^* A P_k$$

avec

$$P_k = Q_1 \ldots Q_k.$$

Ceci prouve que l'algorithme QR produit une suite de matrices A_k qui sont unitairement semblables à la matrice A. De plus

$$\begin{aligned} AP_k &= (Q_1 R_1) Q_1 Q_2 \ldots Q_k = Q_1 (R_1 Q_1) Q_2 \ldots Q_k = Q_1 (Q_2 R_2) Q_2 \ldots Q_k \\ &= Q_1 Q_2 \ldots Q_k (R_k Q_k) = Q_1 Q_2 \ldots Q_k Q_{k+1} R_{k+1} = P_{k+1} R_{k+1} \end{aligned}$$

de sorte que $< P_{k+1} >= A_\sharp(< P_k >)$, autrement dit $(< P_k >)$ est la suite des approximations successives associée à l'opérateur $A_\sharp$ et au point initial $< P_0 >=< I_n >$. Nous avons vu qu'une telle suite converge : il existe $P \in \mathbb{U}_n$ telle que $\lim < P_k >=< P >$ dans l'espace $\mathbb{F}_n$. Par la Proposition 71 il existe une suite (T_k) dans $\mathbb{T}_n$ telle que $\lim P_k T_k = P$. Revenons maintenant à la suite (A_k). Notons $S_k = (P_k T_k)^* A (P_k T_k)$. On a

$$A_{k+1} = P_k^* A P_k = T_k (P_k T_k)^* A (P_k T_k) T_k^* = T_k S_k T_k^*.$$

Comme $< P >$ est un point fixe de $A_\sharp$ il existe $R \in \mathbb{R}_n$ tel que $AP = PR$ de sorte que

$$\lim (P_k T_k)^* A (P_k T_k) = P^* (AP) = P^* (PR) = R.$$

Il est maintenant facile d'étudier le comportement limite de $A_{k+1} = T_k S_k T_k^*$: les termes diagonaux convergent vers ceux de R et les termes de la partie triangulaire inférieure stricte vers 0. Quant à ceux de la partie supérieure ils ne convergent pas nécessairement : leurs modules convergent (puisque les entrées des matrices T_k sont de module 1) mais pas leurs arguments. □

2.10.7 L'algorithme LR de Rutishauser

Cet algorithme, dû à Rutishauser 1955 [40], est conçu comme l'algorithme QR mais, au lieu de la décomposition QR, il utilise la décomposition LU : $A = LU$ avec L triangulaire inférieure à diagonale unité et U triangulaire supérieure. Une telle décomposition s'obtient par la méthode d'élimination de Gauss sans pivotage. Elle existe si et seulement si les mineurs principaux de A sont non nuls, ce qui est le cas générique.

L'algorithme LR est le suivant :

$$A_1 = A = L_1 U_1, \;\; A_{k+1} = U_k L_k = L_{k+1} U_{k+1}.$$

En fait cette méthode est un avatar de la méthode des approximations successives associée à l'action de A sur la variété des drapeaux. Par des arguments identiques à ceux développés quant à QR on a :

$$A_{k+1} = M_k^{-1} A M_k \text{ avec } M_k = L_1 L_2 \dots L_k.$$

De plus

$$AM_k = M_{k+1} U_{k+1}$$

ce qui prouve que

$$A_\sharp < M_k >=< M_{k+1} > \text{ et } < M_0 >=< I_n > .$$

Par une analyse similaire à celle faite pour QR et sous l'hypothèse que A est inversible avec des valeurs propres de modules différents, on montre que la suite (A_k) devient triangulaire supérieure avec une diagonale constituée par les valeurs propres de A.

2.10.8 L'algorithme Cholesky de Wilkinson

Cet algorithme, décrit par Wilkinson 1965 dans [55], a été conçu pour calculer les valeurs propres d'une matrice A. L'algorithme Cholesky est donné par

$$A_1 = A = R_1^* R_1, \; A_{k+1} = R_k R_k^* = R_{k+1}^* R_{k+1}$$

avec $R_k \in \mathbb{R}_n^+$. C'est encore un avatar de la méthode des approximations successives associée à l'action de A sur la variété des drapeaux :

$$A_{k+1} = R_k R_k^* = R_k^{-*} R_k^* R_k R_k^* = R_k^{-*} A_k R_k^* = S_k^{-1} A S_k$$

avec $S_k = R_1^* R_2^* \dots R_k^*$. Comme précédemment on voit que

$$AS_k = S_{k+1} R_{k+1}$$

de sorte que la suite des $(< S_k >)$ vérifie

$$< S_0 >=< I_n > \text{ et } A_\sharp < S_k >=< S_{k+1} > .$$

Encore une fois nous retrouvons la suite des approximations successives de l'action de A sur $\mathbb{F}$ et associée au point initial $< I_n >$.

2.11 Exemple : calcul de sous-espaces invariants

Le problème des vecteurs propres est un cas particulier de celui plus général des sous-espaces invariants. Soit A une matrice $n \times n$ inversible. Un sous-espace vectoriel $F \subset \mathbb{C}^n$ est invariant lorsque $A(F) \subset F$ ou, ce qui revient au même, $A(F) = F$. Cette définition montre que F est un point fixe pour l'action de A sur l'ensemble des sous-espaces vectoriels de $\mathbb{C}^n$. Commençons par décrire ce cadre d'étude.

2.11.1 La variété de Grassmann

Définition 75. *On appelle grassmannienne $\mathbb{G}_{n,p}$ l'ensemble des sous-espaces vectoriels de dimension p contenus dans $\mathbb{C}^n$.*

Nous allons représenter un tel sous-espace par une matrice $n \times p$ de rang p dont les colonnes en constituent une base. Une telle représentation n'est pas unique ce qui conduit à décrire $\mathbb{G}_{n,p}$ comme un espace quotient.

On note $\mathbb{GL}_{n,p}$ l'espace des matrices $n \times p$ de rang p et $\mathbb{S}_{n,p}$ la variété de Stiefel : matrices $n \times p$ dont les p colonnes sont normées et orthogonales deux à deux.

Proposition 76. *La grassmannienne $\mathbb{G}_{n,p}$ s'identifie à l'espace des orbites de l'action suivante de $\mathbb{GL}_p$ sur $\mathbb{GL}_{n,p}$*

$$\mathbb{GL}_{n,p} \times \mathbb{GL}_p \to \mathbb{GL}_{n,p}, \quad (X, L) \to XL.$$

C'est le quotient $\mathbb{GL}_{n,p}/\mathbb{GL}_p$ de $\mathbb{GL}_{n,p}$ pour la relation d'équivalence

$$X \equiv Y \quad \textit{si et seulement si} \quad \exists L \in \mathbb{GL}_p \quad Y = XL.$$

La classe de X pour cette relation est $< X >= X\mathbb{GL}_p$.
$\mathbb{G}_{n,p}$ s'identifie aussi au quotient $\mathbb{G}_{n,p} = \mathbb{S}_{n,p}/\mathbb{U}_p$ de la variété de Stiefel $\mathbb{S}_{n,p}$ pour la relation d'équivalence

$$U \equiv V \quad \textit{si et seulement si} \quad \exists L \in \mathbb{U}_p \quad V = UL.$$

La classe de U pour cette relation est $< U >= U\mathbb{U}_p$.

Preuve Soient $F \in \mathbb{G}_{n,p}$ et X une matrice de taille $n \times p$, de rang $p = \dim F$, dont les colonnes engendrent le sous-espace vectoriel F. On a ainsi représenté F à l'aide d'une matrice $X \in \mathbb{GL}_{n,p}$ mais cette représentation n'est pas unique : deux matrices X et $Y \in \mathbb{GL}_{n,p}$, décrivent le même sous-espace si et seulement s'il existe une matrice $L \in \mathbb{GL}_p$ telle que $X = YL$. D'où la représentation $\mathbb{G}_{n,p} = \mathbb{GL}_{n,p}/\mathbb{GL}_p$. Pour obtenir la seconde structure quotient il suffit de ne considérer que des bases orthonormées. Les p colonnes d'une telle base constituent une matrice $U \in \mathbb{S}_{n,p}$ et si la relation $V = UL$ a lieu entre deux telles matrices on a nécessairement $L \in \mathbb{U}_p$. □

2.11.2 La grassmannienne en tant qu'espace topologique

Munissons $\mathbb{G}_{n,p}$ des topologies quotient déduites des représentations

$$\mathbb{G}_{n,p} = \mathbb{GL}_{n,p}/\mathbb{GL}_p = \mathbb{S}_{n,p}/\mathbb{U}_p.$$

On a :

Proposition 77. *Ces deux topologies quotient sur $\mathbb{G}_{n,p}$ sont identiques.*

Preuve On utilise le Lemme 49 avec $E = \mathbb{GL}_{n,p}$, $F = \mathbb{S}_{n,p}$ et $f : E \to F$ est l'application suivante. A toute matrice $X \in \mathbb{GL}_{n,p}$ on associe une décomposition $X = Q_X R_X$ avec $Q_X \in St$ et $R_X \in \mathbb{R}_p^+$. Cette décomposition est définie par $R_X = Cholesky(X^*X)$ et $Q_X = XR_X^{-1}$. On a

$$Q_X^* Q_X = R_X^{-*} X^* X R_X^{-1} = R_X^{-*} R_X^* R_X R_X^{-1} = I_p$$

ce qui prouve que $Q_X \in \mathbb{S}_{n,p}$. On prend $f(X) = Q_X$. Cette application est continue par la Proposition 62. □

Il existe une troisième manière de faire de $\mathbb{G}_{n,p}$ un espace quotient : en projetant la composante de dimension p d'un drapeau. Notons

$$\Pi_{G,F} : \mathbb{F}_n \to \mathbb{G}_{n,p}, \quad \Pi(F_0 \subset F_1 \subset \ldots \subset F_n) = F_p.$$

Il est clair que cette application est surjective. On a donc

Proposition 78. *$\mathbb{G}_{n,p}$ est le quotient de $\mathbb{F}_n$ par la relation d'équivalence*

$$F \mathcal{R} G \quad \textit{si et seulement si} \quad \Pi_{G,F}(F) = \Pi_{G,F}(G).$$

On peut maintenant munir $\mathbb{G}_{n,p}$ de la topologie quotient déduite de cette nouvelle structure.

Proposition 79. *Les topologies sur $\mathbb{G}_{n,p}$ associées à $\mathbb{GL}_{n,p}/\mathbb{GL}_p$, $\mathbb{S}_{n,p}/\mathbb{U}_p$ et $\mathbb{F}_n/\mathcal{R}$ sont identiques.*

Preuve Notons $\Pi_{S,U} : \mathbb{U}_n \to \mathbb{S}_{n,p}$ l'opérateur qui à une matrice $U \in Un$ associe la matrice obtenue en supprimant de U les $n-p$ dernières colonnes. Alors $\mathbb{S}_{n,p} = \mathbb{U}_n/\Pi_{S,U}$ et l'identification est ensembliste et topologique (nous laissons au lecteur le soin de le justifier). Notons aussi $\Pi_{G,S} : \mathbb{S}_{n,p} \to \mathbb{G}_{n,p}$ la surjection canonique associée au quotient $\mathbb{G}_{n,p} = \mathbb{S}_{n,p}/\mathbb{U}_p$ et $\Pi_{F,U} : \mathbb{U}_n \to \mathbb{F}_n$ la surjection canonique associée au quotient $\mathbb{F}_n = \mathbb{U}_n/\mathbb{T}_n$. Pour prouver que les deux topologies considérées (notées $\mathcal{T}_F$ et $\mathcal{T}_S$) sont identiques il suffit de prouver qu'elles donnent les mêmes fonctions continues $f : \mathbb{G}_{n,p} \to H$ où H est un espace topologique arbitraire. Pour ce faire on remarque qu'un tel f est continu pour $\mathcal{T}_F$ si et seulement si $f \circ \Pi_{G,F}$ est continu pour la topologie de $\mathbb{F}_n$ (Lemme 48), si et seulement si $f \circ \Pi_{G,F} \circ \Pi_{F,U}$ est continu pour la topologie de $\mathbb{U}_n$. On a

$$\Pi_{G,F} \circ \Pi_{F,U} = \Pi_{G,S} \circ \Pi_{S,U}$$

de sorte que la condition sur f devient $f \circ \Pi_{G,S} \circ \Pi_{S,U}$ continu pour la topologie de $\mathbb{U}_n$ qui devient $f \circ \Pi_{G,S}$ continu pour la topologie de $\mathbb{S}_{n,p}$ et donc f continu pour $\mathcal{T}_S$. □

On a obtenu une seule topologie sur $\mathbb{G}_{n,p}$ à l'aide de trois descriptions différentes. Il en résulte, via les Lemmes 48 et 49 le résultat suivant :

Proposition 80. *$\mathbb{G}_{n,p}$ est un espace compact, l'application $\Pi_{G,F} : \mathbb{F}_n \to \mathbb{G}_{n,p}$ est surjective, continue et l'image d'un ouvert de $\mathbb{F}_n$ est un ouvert de $\mathbb{G}_{n,p}$.*

2.11.3 L'action de A sur la grassmannienne

On définit cette action de A par

$$A_\spadesuit : \mathbb{G}_{n,p} \to \mathbb{G}_{n,p},\ A_\spadesuit(< X >)) =< AX >$$

pour tout $X \in \mathbb{GL}_{n,p}$. Les points fixes de $A_\spadesuit$ sont les sous-espaces de dimension p de $\mathbb{C}^n$ qui sont invariants par A. La méthode des approximations successives pour calculer ces points fixes est définie par

$$< X_{k+1} >= A_\spadesuit(< X_k >), X_0 \in \mathbb{GL}_{n,p} \text{ donné.}$$

Les implémentations de cette méthode sont dans $\mathbb{GL}_{n,p}$: la plus simple, qui généralise la méthode de la puissance, consiste à poser

$$P_0 = X_0 \ \text{ et } \ P_{k+1} = \alpha_k A P_k$$

où α_k est un scalaire non nul, un facteur de normalisation. On a $< X_k >$ $=< P_k >$ pour tout k.

Une seconde possibilité consiste à utiliser la décomposition LU d'une matrice $X \in \mathbb{GL}_{n,p}$. On entend par là une identité $X = LU$ où $L \in \mathbb{GL}_{n,p}$ est triangulaire inférieure à diagonale unité et $U \in \mathbb{GL}_n$ est triangulaire supérieure ; pour une matrice rectangulaire telle que L, triangulaire inférieure à diagonale unité signifie que $L_{ii} = 1$ pour tout $1 \leq i \leq p$ et que $L_{ij} = 0$ pour tout $1 \leq i < j \leq p$. Une telle décomposition peut s'obtenir via la méthode du pivot de Gauss. Notons que lorsque $X = LU$ on a $< X >=< L >$.

La méthode de Treppen construit une suite (L_k) de matrices triangulaires inférieures à diagonale unité telle que $< X_k >=< L_k >$. Cette suite est construite via la décomposition LU

$$X_0 = L_0 U_0 \ \text{ et } \ AL_k = L_{k+1} U_{k+1}.$$

Une troisième implémentation utilise la décomposition QR d'une matrice $X \in \mathbb{GL}_{n,p}$: $X = QR$ avec $Q \in \mathbb{S}_{n,p}$ et $R \in \mathbb{GL}_p$ triangulaire supérieure. Une telle décomposition peut s'obtenir par le procédé d'orthonormalisation de Gram-Schmidt appliqué aux colonnes de X. On construit une suite $Q_k \in \mathbb{S}_{n,p}$ telle que $< X_k >=< Q_k >$ en posant

$$X_0 = Q_0 R_0 \ \text{ et } \ AQ_k = Q_{k+1} R_{k+1}.$$

Dans le théorême qui suit on analyse la convergence de cette méthode lorsque les modules des valeurs propres sont distincts.

Théorème 81. *Supposons que A ait des valeurs propres de modules distincts. Alors, $A_\spadesuit$ possède $\binom{n}{p}$ points fixes. Ceux-ci sont les sous-espaces engendrés par p vecteurs propres indépendants de A. Pour tout $F \in \mathbb{G}_{n,p}$, la suite*

$$F^0 = F,\ F^{k+1} = A_\spadesuit(F^k)$$

converge vers l'un de ces sous-espaces. Le bassin d'attraction du sous-espace engendré par les p vecteurs correspondant aux p valeurs propres de plus grand module est ouvert et dense dans $\mathbb{G}_{n,p}$.

Preuve La preuve de ce théorème utilise le Théorème 73 dont nous allons utiliser les notations. Notons pour simplifier

$$\Pi : \mathbb{F}_n \to \mathbb{G}_{n,p},\quad \Pi(F_0 \subset F_1 \subset \ldots \subset F_n) = F_p.$$

Cette application est surjective, continue et transforme les ouverts de $\mathbb{F}_n$ en ouverts de $\mathbb{G}_{n,p}$ (Proposition 80). L'action de A sur la grassmannienne se déduit de celle de A sur la variété des drapeaux puisque

$$A_\spadesuit \circ \Pi = \Pi \circ A_\sharp.$$

Les sous-espaces engendrés par p vecteurs propres indépendants constituent autant de points fixes de $A_\spadesuit$. Nous allons voir que les suites des approximations successives $(A^k_\spadesuit(F_p))$ convergent vers ces points fixes pour tout $F_p \in \mathbb{G}_{n,p}$ donné. Ceci prouvera qu'il n'y en a pas d'autres. Pour ce faire on écrit $F_p = \Pi(F) = \Pi(F_0 \subset F_1 \subset \ldots \subset F_n)$ et on applique le Théorème 73 à la suite des itérés $F^k = A^k_\sharp(F)$ dans $\mathbb{F}_n$. Elle converge vers un point fixe de $A_\sharp$. Un tel point fixe est un drapeau dont les composantes sont des sous-espaces engendrés par des vecteurs propres de A. On projette cette situation par Π et l'on obtient le résultat souhaité.

Rangeons les valeurs propres de A par module décroissant : $|\lambda_1| > \ldots > |\lambda_n|$. Nous avons vu, à la fin de la preuve du Théorème 73, que le point fixe de $A_\sharp$, $F = F_0 \subset \ldots \subset F_n$, où F_k est le sous-espace engendré par les vecteurs propres associés aux valeurs propres $\lambda_1, \ldots, \lambda_k$, c'est-à-dire les k valeurs propres de plus grand module, possède un bassin d'attraction ouvert et dense dans $\mathbb{F}_n$. En projettant cette situation par Π on prouve que le bassin d'attraction du sous-espace engendré par les p vecteurs correspondant aux p valeurs propres de plus grand module pour $A_\spadesuit$ est ouvert et dense dans $\mathbb{G}_{n,p}$. □

2.12 Angles entre sous-espaces d'un espace hermitien

Dans les lignes qui suivent nous décrivons une distance sur la grassmannienne à l'aide du concept d'angle entre sous-espaces. Notons $\mathbb{E}$ un espace hermitien complexe ou bien euclidien réel. Pour mesurer la distance entre deux sous-espaces vectoriels V et W de $\mathbb{E}$ nous considérons la quantité :

Définition 82.

$$d(V,W) = \max_{v\in V,\ v\neq 0}\ \min_{w\in W}\ \frac{\|v-w\|}{\|v\|} = \max_{v\in V,\ \|v\|=1}\ \min_{w\in W}\ \|v-w\|.$$

Ce nombre est le maximum du sinus de l'angle fait par un vecteur $v \in V$ avec sa projection orthogonale w sur W. Lorsque V et W sont des droites vectorielles on retrouve la Définition 52.

Soit X un sous-espace vectoriel de $\mathbb{E}$. Nous notons Π_X la projection orthogonale sur X. La proposition suivante donne les principales propriétés de $d(V, W)$.

Proposition 83.

1. $d(V, W) = \|\Pi_{W^\perp} \circ \Pi_V\|$,
2. $d(V, W) = d(W^\perp, V^\perp)$,
3. $d(V, W) = d(V \cap (V \cap W)^\perp, W \cap (V \cap W)^\perp)$,
4. $0 \leq d(V, W) \leq 1$,
5. $d(V, W) = 0$ *si et seulement si* $V \subset W$,
6. $d(V, W) < 1$ *si et seulement si* $V \cap W^\perp = \{0\}$,
7. $d(V_1, V_3) \leq d(V_1, V_2) + d(V_2, V_3)$,
8. *Si* $V_1 \subset V_2$ *alors* $d(V_1, W) \leq d(V_2, W)$ *et si* $W_1 \subset W_2$ *alors* $d(V, W_2) \leq d(V, W_1)$,
9. $d(V, W_1 + W_2) \leq \min(d(V, W_1), d(V, W_2))$,
10. *Si* V_1 *alors* V_2 *sont orthogonaux* $d(V_1 \oplus V_2, W) \leq d(V_1, W) + d(V_2, W)$ *et*
11. $d(V_1 \oplus V_2, W) \leq \sqrt{2} \max(d(V_1, W), d(V_2, W))$,
12. *Si* $dim V = dim W$ *alors* $d(V, W) = d(W, V)$,
13. $d(V, W)$ *est une distance sur l'ensemble* $\mathbb{G}(\mathbb{E}, p)$ *des sous-espaces vectoriels de* $\mathbb{E}$ *de dimension* p.

Preuve 1 est une conséquence de

$$\begin{aligned} d(V, W) &= \max_{v \in V,\ \|v\|=1} \|(id - \Pi_W)v\| = \max_{v \in V,\ \|v\| \leq 1} \|\Pi_{W^\perp} v\| \\ &= \max_{\|v\|=1} \|\Pi_{W^\perp} \Pi_V v\| = \|\Pi_{W^\perp} \Pi_V\|. \end{aligned}$$

2 est une consequence de 1 parce que les normes d'un operateur et de son transposé sont égales.

Pour 3, soit $v \in V$ décomposé en

$$v = v_1 + v_2 \in (V \cap W) \oplus (V \cap (V \cap W)^\perp).$$

On a

$$\Pi_W v = w_1 + w_2 \in (V \cap W) \oplus (W \cap (V \cap W)^\perp)$$

avec $w_1 = v_1$ et $w_2 = \Pi_{W \cap (V \cap W)^\perp}(v_2)$.

Les propriétés 4 à 10 sont faciles.

Prouvons 11. Si v_1 et v_2 sont orthogonaux alors $\|v_1\| + \|v_2\| \leq \sqrt{2}\|v_1 + v_2\|$. Donc, si V_1 et V_2 sont orthogonaux

$$\begin{aligned} d(V_1 \oplus V_2, W) &= \|\Pi_{W^\perp}(v_1+v_2)\| \le \|\Pi_{W^\perp} v_1\| + \|\Pi_{W^\perp} v_2\| \\ &\le d(V_1,W)\|v_1\| + d(V_2,W)\|v_2\| \\ &\le \max(d(V_1,W), d(V_2,W))(\|v_1\| + \|v_2\|) \\ &\le \sqrt{2}\max(d(V_1,W), d(V_2,W))\|v_1+v_2\|. \end{aligned}$$

Pour prouver 12 remarquons que $d(V,W)$ est la plus grande valeur singulière de $\Pi_{W^\perp}\Pi_V = (id - \Pi_W)\Pi_V = \Pi_V - \Pi_W\Pi_V$ et, de la même manière, $d(W,V)$ est la plus grande valeur singulière de $\Pi_W - \Pi_V\Pi_W$. Soit Q une transformation unitaire dans $\mathbb{E}$ qui vérifie $Q^2 = \mathrm{id}_{\mathbb{E}}$ and $QV = W$. L'existence d'une telle involution unitaire sera prouvée au lemme 84. On a $\Pi_W = Q\Pi_V Q$, donc

$$\Pi_{W^\perp}\Pi_V = \Pi_V - \Pi_W\Pi_V = \Pi_V - Q\Pi_V Q\Pi_V$$

et de même

$$\Pi_{V^\perp}\Pi_W = Q(\Pi_V - Q\Pi_V Q\Pi_V)Q.$$

Ainsi $\Pi_{W^\perp}\Pi_V$ and $\Pi_{V^\perp}\Pi_W$ ont les mêmes valeurs singulières de sorte que $d(V,W) = d(W,V)$.

L'assertion 13 est une conséquence de 5, 7 et 12. □

Lemme 84. *Soient V et W deux sous-espaces vectoriels de $\mathbb{E}$ de même dimension p. Il existe un endomorphisme de $\mathbb{E}$ qui soit involutif ($Q\circ Q = id_{\mathbb{E}}$), unitaire ($Q^*\circ Q = Q\circ Q^* = id_{\mathbb{E}}$) et tel que $Q(V) = W$.*

Preuve Nous donnons ici une preuve élégante et concise due à A. J. Hoffman. On considère le cas $\mathbb{E} = \mathbb{C}^{2p}$, $V \cap W = \{0\}$ et $V \oplus W = \mathbb{C}^{2p}$. Le cas général s'y ramène. On suppose aussi que V est engendré par les p premiers vecteurs de la base canonique de $\mathbb{C}^{2p}$. Introduisons les matrices $2p \times p$ suivantes :

$$S = \begin{pmatrix} I_p \\ O \end{pmatrix} \quad \text{et} \quad T = \begin{pmatrix} A \\ C \end{pmatrix}$$

de sorte que les colonnes de T constituent une base orthonormée de W ; nous dirons que c'est une matrice de Stiefel. Une telle matrice vérifie $T^*T = I_p$. Notons que les colonnes de S engendrent V. Ecrivons $AU = H$ la décomposition polaire de A : U est unitaire et H est semi-définie positive. Les colonnes de $TU = \begin{pmatrix} H \\ B^\star \end{pmatrix}$ engendrent aussi W. Remarquons que $B^\star$ est inversible : si $B^\star x = 0$ alors $TUx = \begin{pmatrix} Hx \\ 0 \end{pmatrix}$ de sorte que $TUx \in V \cap W = \{0\}$. Il en résulte que $x = 0$ puisque U est unitaire et que T est de Stiefel. B est aussi inversible. Considérons maintenant la matrice $2p \times 2p$

$$Q = \begin{pmatrix} H & B \\ B^\star & -B^{-1}HB \end{pmatrix}.$$

On a

$$H^2 + BB^\star = \begin{pmatrix} H & B \end{pmatrix} \begin{pmatrix} H \\ B^\star \end{pmatrix} = U^\star T^\star T U = I_p$$

de sorte que

$$HBB^\star = H(I_p - H^2) = (I_p - H^2)H = BB^\star H.$$

Nous en déduisons que $B^{-1}HB = B^\star H B^{-\star}$ de sorte que Q est hermitienne. En utilisant le même argument on voit que $Q^2 = I_{2p}$ de sorte que Q est une involution. Pour terminer cette démonstration notons que $QS = \begin{pmatrix} H \\ B^\star \end{pmatrix} = TU$ engendre W. □

3

La méthode de Newton

3.1 Introduction

L'itération de Newton est une méthode numérique classique de recherche des zéros d'un système d'équations

$$f : \mathbb{E} \to \mathbb{F}$$

où $\mathbb{E}$ et $\mathbb{F}$ sont des espaces de Banach réels ou complexes. Si x est une approximation d'un zéro de ce système, la méthode de Newton raffine cette approximation en prenant pour nouvelle valeur la solution y de l'équation linéarisée au voisinage de x :

$$f(x) + Df(x)(y - x) = 0.$$

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

$$y = x - Df(x)^{-1} f(x).$$

On appelle opérateur de Newton l'expression ainsi définie : $N_f(x) = x - Df(x)^{-1} f(x)$. Il est défini sur $\mathbb{E} \setminus \Sigma_f$, l'ensemble des points réguliers pour f, c'est-à-dire de dérivée inversible.

L'idée d'améliorer la qualité d'une approximation par ajout d'un terme correctif (à x on ajoute ici $-Df(x)^{-1} f(x)$) est fort ancienne. La méthode que nous présentons apparait dans un contexte déjà très général dans *De analysi per aequationes numero terminorum infinitas* de 1669, où Newton considère des équations polynomiales et utilise une technique de linéarisation. Le cas de l'équation de Kepler $x - esin(x) = M$, une équation qui n'est pas polynomiale, est décrit dans *Philosophiae Naturalis Principia Mathematica* publié en 1687. La méthode y trouve toute sa force puisqu'il n'est plus possible de linéariser par des techniques algébriques, comme cela peut se faire pour des équations polynomiales. Deux 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. Il s'agit de la même méthode que celle décrite dans *De analysi . . .* mais présentée différemment. Puis vient Simpson, qui dans son essai *Essays in Mathematicks*, 1740, introduit « une nouvelle méthode de résolution des équations » utilisant « la méthode des fluxions » c'est-à-dire les dérivées. 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. On doit l'étude des systèmes d'équations à L. Runge et H. Koenig, 1924, et le point de vue « moderne » à L. Kantorovich et A. Ostrowski. Le dernier des grands noms associés à la méthode de Newton est S. Smale qui a introduit le point de vue appelé « théorie alpha » que nous décrivons dans les lignes qui suivent. L'histoire de la méthode de Newton est décrite par Ypma [57] où nous renvoyons le lecteur.

La méthode de Newton est fondée sur l'étude de la suite

$$x_{k+1} = N_f(x_k) = x_k - Df(x_k)^{-1} f(x_k)$$

où x_0 est donné et dont on cherche les points fixes. Si la suite (x_k) converge vers $\zeta \notin \Sigma_f$ alors $f(\zeta) = 0$: les zéros non-singuliers de f correspondent aux points fixes de N_f. De plus, la dérivée de l'opérateur de Newton est donnée par

$$DN_f(x) = Df(x)^{-1} D^2 f(x) Df(x)^{-1} f(x)$$

qui est donc nulle en un point fixe. En vertu du Théorème 7 ces points fixes sont super-attractifs : la convergence de la suite (x_k) est quadratique.

A l'opérateur de Newton est associée l'équation différentielle (équation de Newton)

$$x' = -Df(x)^{-1} f(x).$$

Il est bon de voir la suite de Newton comme la solution approchée de cette équation donnée par la méthode d'Euler :

$$\frac{x_{k+1} - x_k}{t_{k+1} - t_k} = -Df(x_k)^{-1} f(x_k)$$

où x_k est l'approximation de la solution $x(t)$ de l'équation correspondant à l'état initial $x(t_0) = x_0$ et à l'instant t_k. On obtient très exactement la suite de Newton en normalisant à 1 les périodes de temps $t_{k+1} - t_k$.

Quelles sont les propriétés de convergence de la suite de Newton ? Comment faut-il choisir le point initial x_0 pour être assuré que la suite converge ? Quelle vitesse de convergence peut-on obtenir ? Nous aborderons ces questions sous deux angles. Le premier, que l'on qualifie de « théorie de Kantorovitch » privilégie des systèmes $f(x) = 0$ de classe C^2 et l'étude de la suite de Newton $x_{k+1} = N_f(x_k)$ se fait à partir du comportement de ce système au voisinage du point initial x_0.

Le second point de vue, « la théorie alpha de Smale », suppose que le système $f(x) = 0$ est analytique, donc beaucoup plus régulier que pour la théorie de Kantorovitch, mais les hypothèses faites sont plus faibles et portent

uniquement sur le comportement du système au point initial x_0, non plus au voisinage de ce point.

Deux types de théorèmes vont être formulés. L'un décrit le bassin d'attraction quadratique d'un zéro donné du système, l'autre donne un critère au point initial x_0 pour que la suite de Newton converge vers un zéro du système, dont par là même on prouve l'existence.

3.2 La théorie de Kantorovitch

Le contexte que nous utilisons est le suivant : $\mathbb{E}$ et $\mathbb{F}$ sont des espaces de Banach réels ou complexes, U est un ouvert de $\mathbb{E}$ et $f : U \to \mathbb{F}$ est de classe C^2 sur U. On note $\bar{B}(x, r)$ la boule fermée de centre x et de rayon r et $B(x, r)$ la boule ouverte. Le premier résultat que nous donnons est une reformulation du Théorème 7.

Théorème 85. *Soit $\zeta \in U$ tel que $f(\zeta) = 0$ et que $Df(\zeta)$ soit inversible. Soit $r > 0$ tel que $\bar{B}(\zeta, r) \subset U$. Notons*

$$K(f, \zeta, r) = \sup_{\|x-\zeta\| \leq r} \|Df(\zeta)^{-1} D^2 f(x)\|.$$

Si $2K(f, \zeta, r)r \leq 1$ alors, pour tout $x_0 \in \bar{B}(\zeta, r)$, la suite de Newton $x_{k+1} = N_f(x_k)$ est définie et converge vers ζ. De plus

$$\|x_k - \zeta\| \leq \left(\frac{1}{2}\right)^{2^k - 1} \|x_0 - \zeta\|.$$

Nous utiliserons le lemme suivant :

Lemme 86. *Soit $L : \mathbb{E} \to \mathbb{E}$ un opérateur linéaire et continu. Si*

$$\|L\| = \sup_{\|x\|=1} \|L(x)\| < 1$$

alors $id_{\mathbb{E}} - L$ est inversible. Son inverse est donné par la somme de la série absolument convergente :

$$(id_{\mathbb{E}} - L)^{-1} = \sum_{k=0}^{\infty} L^k.$$

De plus

$$\|(id_{\mathbb{E}} - L)^{-1}\| \leq \frac{1}{1 - \|L\|}.$$

Preuve du Lemme 86. La série ci-dessus est absolument convergente puisqu'on peut majorer la série des normes par la série géométrique de raison $\|L\| < 1$. Comme l'espace des endomorphismes continus de $\mathbb{E}$ est complet, la série converge et sa somme est un endomorphisme continu M de $\mathbb{E}$. On a

$$(\mathrm{id}_{\mathbb{E}} - L)\sum_{k=0}^{p} L^k = \mathrm{id}_{\mathbb{E}} - L^{p+1}.$$

Lorsque $p \to \infty$ on a $L^{p+1} \to 0$ puisque la série converge et d'autre part

$$(\mathrm{id}_{\mathbb{E}} - L)\sum_{k=0}^{p} L^k \to (\mathrm{id}_{\mathbb{E}} - L)M.$$

On obtient donc à la limite

$$(\mathrm{id}_{\mathbb{E}} - L)M = (\mathrm{id}_{\mathbb{E}} - L)\sum_{k=0}^{\infty} L^k = \mathrm{id}_{\mathbb{E}}$$

autrement dit

$$(\mathrm{id}_{\mathbb{E}} - L)^{-1} = \sum_{k=0}^{\infty} L^k.$$

L'inégalité sur les normes s'en déduit. □

Preuve du Théorème 85. Commençons par prouver que $Df(x)$ est inversible pour tout x tel que $\|x - \zeta\| \leq r$. La formule de Taylor, donnée en appendice, appliquée à la fonction $Df(\zeta)^{-1}Df(x)$ donne

$$Df(\zeta)^{-1}Df(x) = \mathrm{id}_{\mathbb{E}} + \int_0^1 Df(\zeta)^{-1}D^2f(\zeta + t(x-\zeta))(x-\zeta)dt$$

de sorte que

$$\left\|\mathrm{id}_{\mathbb{E}} - Df(\zeta)^{-1}Df(x)\right\| = \left\|\int_0^1 Df(\zeta)^{-1}D^2f(\zeta + t(x-\zeta))(x-\zeta)dt\right\|$$

$$\leq \int_0^1 \left\|Df(\zeta)^{-1}D^2f(\zeta + t(x-\zeta))\right\| \ \|(x-\zeta)\|\, dt \leq rK(f,\zeta,r) \leq \frac{1}{2}.$$

Nous en déduisons, par le Lemme 86, que $Df(\zeta)^{-1}Df(x) = id_{\mathbb{E}} - (\mathrm{id}_{\mathbb{E}} - Df(\zeta)^{-1}Df(x))$ est inversible et aussi que

$$\|Df(x)^{-1}Df(\zeta)\| \leq 2.$$

Ainsi, l'opérateur de Newton est défini sur $\bar{B}(\zeta, r)$. Par la formule de Taylor, appliquée à $Df(\zeta)^{-1}f(x)$ on a

$$0 = Df(\zeta)^{-1}f(\zeta) = Df(\zeta)^{-1}f(x) + Df(\zeta)^{-1}Df(x)(\zeta - x)$$
$$+ \int_0^1 (1-t)Df(\zeta)^{-1}D^2f(x + t(\zeta - x))(\zeta - x)^2dt$$

d'où l'on déduit, en composant à gauche par $Df(x)^{-1}Df(\zeta)$, que

$$N_f(x) - \zeta = Df(x)^{-1}Df(\zeta) \int_0^1 (1-t)Df(\zeta)^{-1}D^2f(x+t(\zeta-x))(\zeta-x)^2 dt.$$

Compte tenu des estimations précédentes, on a

$$\begin{aligned}\|N_f(x) - \zeta\| &\leq \|Df(x)^{-1}Df(\zeta)\| \\ &\quad \times \int_0^1 (1-t)\ \|Df(\zeta)^{-1}D^2f(x+t(\zeta-x))\|\ \|\zeta-x\|^2 dt \\ &\leq K(f,\zeta,r)\|\zeta-x\|^2.\end{aligned}$$

On prouve alors par récurrence sur k que $x_k \in \bar{B}(\zeta, r)$ et l'estimation

$$\|x_k - \zeta\| \leq \left(\frac{1}{2}\right)^{2^k-1} \|x_0 - \zeta\|$$

en suivant les lignes de la preuve du Théorème 7. □

Nous allons maintenant établir un critère de convergence, pour une suite de Newton, qui ne fasse pas appel, à priori, à la connaissance d'un zéro du système.

Définition 87. *Définissons*

$$\beta(f, x_0) = \|Df(x_0)^{-1}f(x_0)\|$$

si $Df(x_0)$ est un isomorphisme et $\beta(f, x_0) = \infty$ sinon.

Théorème 88. *Soient $x_0 \in U$ et $r > 0$ tels que $\bar{B}(x_0, r) \subset U$. Si les conditions suivantes sont satisfaites,*
- *$Df(x_0)$ est un isomorphisme,*
- *$2\beta(f, x_0) \leq r$,*
- *$2\beta(f, x_0)K(f, x_0, r) \leq 1$,*

alors il existe un unique $\zeta \in \bar{B}(x_0, r)$ tel que
- *$f(\zeta) = 0$,*
- *$Df(\zeta)$ est un isomorphisme,*
- *$\|x_0 - \zeta\| \leq 1.63281\ldots\beta(f, x_0)$.*

De plus, la suite de Newton $x_{k+1} = N_f(x_k)$ est définie, converge vers ζ et

$$\|x_k - \zeta\| \leq 1.63281\ldots\left(\frac{1}{2}\right)^{2^k-1} \beta(f, x_0)$$

avec

$$1.63281\ldots = \sum_{k=0}^{\infty} \frac{1}{2^{2^k-1}}.$$

Preuve du Théorème 88. Considérons la suite de Newton

$$x_{k+1} = x_k - Df(x_k)^{-1} f(x_k).$$

On a :

$$\|x_1 - x_0\| = \|Df(x_0)^{-1} f(x_0)\| = \beta(f, x_0) \leq \frac{r}{2}.$$

De plus, par la formule de Taylor,

$$id_{\mathbb{E}} - Df(x_0)^{-1} Df(x_1) = Df(x_0)^{-1} Df(x_0) - Df(x_0)^{-1} Df(x_1)$$

$$= -\int_0^1 Df(x_0)^{-1} D^2 f(x_0 + t(x_1 - x_0))(x_1 - x_0) dt$$

dont la norme est majorée par

$$\|id_{\mathbb{E}} - Df(x_0)^{-1} Df(x_1)\| \leq \|x_1 - x_0\| K(f, x_0, r) = \beta(f, x_0) K(f, x_0, r) \leq \frac{1}{2}.$$

Par le Lemme 86 $Df(x_0)^{-1} Df(x_1)$ est inversible et son inverse vérifie

$$\|Df(x_1)^{-1} Df(x_0)\| \leq 2.$$

En conséquence

$$\beta(f, x_1) = \|x_2 - x_1\| = \|Df(x_1)^{-1} f(x_1)\|$$

$$= \|Df(x_1)^{-1} Df(x_0) Df(x_0)^{-1} (f(x_1) - f(x_0) - Df(x_0)(x_1 - x_0))\|$$

$$\leq \|Df(x_1)^{-1} Df(x_0)\| \, \| \int_0^1 (1-t) Df(x_0)^{-1} D^2 f(x_0 + t(x_1 - x_0))(x_1 - x_0)^2 dt\|$$

$$\leq \|x_1 - x_0\|^2 K(f, x_0, r) = \beta(f, x_0)^2 K(f, x_0, r) \leq \frac{\beta(f, x_0)}{2} \leq \frac{r}{4}.$$

Notons que, $\bar{B}(x_1, r/2) \subset \bar{B}(x_0, r)$ de sorte que

$$K\left(f, x_1, \frac{r}{2}\right) \leq \|Df(x_1)^{-1} Df(x_0)\| K(f, x_0, r) \leq 2K(f, x_0, r)$$

et

$$2\beta(f, x_1) K\left(f, x_1, \frac{r}{2}\right) \leq 2 \frac{\beta(f, x_0)}{2} 2K(f, x_0, r) \leq 1.$$

Nous pouvons donc appliquer à $(x_1, r/2)$ un argument similaire et par récurrence on vérifie que

$$\|x_{k+1} - x_k\| \leq \frac{\beta(f, x_0)}{2^{2^k - 1}}.$$

Cette suite est de Cauchy, notons ζ sa limite. Il est clair que

$$\|\zeta - x_0\| \leq \sum_{k=0}^{\infty} \frac{\beta(f, x_0)}{2^{2^k - 1}} \leq 1.63281 \ldots \beta(f, x_0) \leq r$$

et aussi que

$$\|\zeta - x_p\| \leq \sum_{k=p}^{\infty} \frac{\beta(f,x_0)}{2^{2^k-1}} \leq \frac{1}{2^{2^p-1}} \sum_{k=0}^{\infty} \frac{\beta(f,x_0)}{2^{2^k-1}} \leq 1.63281\ldots \frac{1}{2^{2^p-1}}\beta(f,x_0).$$

Montrons maintenant que ζ est un zéro non-singulier de f. Comme précedemment pour x_1, nous prouvons que

$$\begin{aligned}\|id_{\mathbb{E}} - Df(x_0)^{-1}Df(\zeta)\| &\leq \|\zeta - x_0\|K(f,x_0,r)\\ &\leq 1.63281\ldots\beta(f,x_0)K(f,x_0,r) < 0.85 < 1,\end{aligned}$$

et par conséquent $Df(x_0)^{-1}Df(\zeta) = \mathrm{id}_{\mathbb{E}} - (id_{\mathbb{E}} - Df(x_0)^{-1}Df(\zeta))$ est un isomorphisme.

Pour prouver que ζ est un zéro nous avons besoin d'une borne sur $\|Df(x)\|$. Elle est obtenue via la formule de Taylor. Pour $x \in \bar{B}(x_0,r)$ on a

$$Df(x) = Df(x_0) + \int_0^1 Df(x_0)Df(x_0)^{-1}D^2f(x_0 + t(x-x_0))(x-x_0)\ dt$$

de sorte que

$$\begin{aligned}\|Df(x)\| &= \|Df(x_0)\|\left(1 + \int_0^1 \|Df(x_0)^{-1}D^2f(x_0+t(x-x_0))\|\ \|x-x_0\|dt\right)\\ &\leq \|Df(x_0)\|\ (1 + rK(f,x_0,r)).\end{aligned}$$

On a alors

$$\begin{aligned}\|f(x_k)\| &\leq \|Df(x_k)\|\|Df(x_k)^{-1}f(x_k)\| \leq \|Df(x_0)\|\ (1+rK(f,x_0,r))\beta(f,x_k)\\ &\leq \|Df(x_0)\|\ (1+rK(f,x_0,r))\frac{\beta(f,x_0)}{2^{2^k-1}}\end{aligned}$$

et cette expression a pour limite 0 lorsque $k \to \infty$. Ceci prouve que $f(\zeta) = 0$.

Pour finir cette démonstration, nous devons montrer qu'un seul zéro satisfait ces critères. Soit ζ' tel que $f(\zeta') = 0$ et $\|\zeta' - x_0\| \leq 1.63281\ldots\beta(f,x_0)$. La suite (x_k) définie précédemment vérifie $\|x_k - x_0\| \leq 2\beta(f,x_0) \leq r$ et $f(x_k) + Df(x_k)(x_{k+1} - x_k) = 0$ de sorte que

$$Df(x_k)(x_{k+1} - \zeta') = f(\zeta') - f(x_k) - Df(x_k)(\zeta' - x_k).$$

Ainsi

$$\begin{aligned}x_{k+1} - \zeta' &= Df(x_k)^{-1}(f(\zeta') - f(x_k) - Df(x_k)(\zeta' - x_k))\\ &= Df(x_k)^{-1}Df(x_0)Df(x_0)^{-1}(f(\zeta') - f(x_k) - Df(x_k)(\zeta' - x_k)).\end{aligned}$$

Par l'argument déjà utilisé (formule de Taylor à l'ordre 2 et $\|Df(x_k)^{-1}Df(x_0)\| \leq 2$) on obtient

$$\|x_{k+1} - \zeta'\| \leq \|x_k - \zeta'\|^2 K(f,x_0,r),$$

puis par récurrence que

$$\|x_k - \zeta'\| \leq \frac{1}{2^{2^k-1}}\beta(f,x_0).$$

Il suffit alors de passer à la limite lorsque $k \to \infty$ pour obtenir $\zeta = \zeta'$. □

3.3 La théorie alpha de Smale

Le contexte que nous utilisons tout au long de cette section est le suivant : $\mathbb{E}$ et $\mathbb{F}$ sont des espaces de Banach réels ou complexes, U est un ouvert de $\mathbb{E}$ et $f : U \to \mathbb{F}$ est analytique sur U. On note $\bar{B}(x, r)$ la boule fermée de centre x et de rayon r et $B(x, r)$ la boule ouverte. Puisque f est analytique, elle est développable en série de Taylor au voisinage de x :

$$f(y) = f(x) + \sum_{k=1}^{\infty} \frac{D^k f(x)}{k!}(y - x)^k$$

dont le rayon de convergence $R(f, x) > 0$ est donné par

$$R(f, x)^{-1} = \limsup_{k\to\infty} \left\| \frac{D^k f(x)}{k!} \right\|^{\frac{1}{k}}.$$

Nous ferons l'hypothèse que, pour tout $x \in U$,

$$B\left(x, \left(1 - \frac{\sqrt{2}}{2}\right) R(f, x)\right) \subset U.$$

Cette hypothèse est toujours satisfaite lorsque $U = \mathbb{E}$ ou bien, plus généralement, lorsque U est le domaine d'analyticité de la fonction f.

Le nombre suivant va jouer un grand rôle dans l'étude des propriétés de convergence des suites de Newton.

Définition 89. *Pour tout $x \in U$ tel que $Df(x)$ soit un isomorphisme on pose*

$$\gamma(f, x) = \sup_{k\geq 2} \left\| Df(x)^{-1} \frac{D^k f(x)}{k!} \right\|^{\frac{1}{k-1}}$$

et $\gamma(f, x) = \infty$ sinon.

La définition de $\gamma(f, x)$ est à rapprocher de

$$K(f, x, r) = \sup_{\|x-y\|\leq r} \|Df(x)^{-1} D^2 f(y)\|.$$

introduit dans l'énoncé du Théorème 85. Lorsque f est quadratique, donc de dérivée seconde constante, on a $K(f, x, r) = 2\gamma(f, x)$.

Nous allons voir que $1/\gamma(f, x)$ minore le rayon de convergence de cette série.

Proposition 90. $R(f,x)^{-1} \leq \gamma(f,x)$.

Preuve On a

$$\begin{aligned} R(f,x)^{-1} &\leq \limsup_{k\to\infty} \|Df(x)\|^{\frac{1}{k}} \left\| Df(x)^{-1}\frac{D^k f(x)}{k!}\right\|^{\frac{1}{k}} \\ &= \limsup_{k\to\infty} \left\| Df(x)^{-1}\frac{D^k f(x)}{k!}\right\|^{\frac{1}{k}} = \limsup_{k\to\infty} \left\| Df(x)^{-1}\frac{D^k f(x)}{k!}\right\|^{\frac{1}{k-1}} \\ &\leq \sup_{k\geq 2} \left\| Df(x)^{-1}\frac{D^k f(x)}{k!}\right\|^{\frac{1}{k-1}} = \gamma(f,x). \end{aligned}$$

Le théorème suivant décrit le rayon d'une boule contenue dans le bassin d'attraction quadratique d'un zéro de f :

Théorème 91. *(Théorème gamma) Soit $\zeta \in U$ tel que $f(\zeta) = 0$ et que $Df(\zeta)$ soit inversible. Soit $x_0 \in U$ tel que*

$$\|x_0 - \zeta\|\gamma(f,\zeta) \leq \frac{3-\sqrt{7}}{2} = 0.17712\ldots.$$

Alors, la suite de Newton $x_{k+1} = N_f(x_k)$ est définie et converge vers ζ. De plus

$$\|x_k - \zeta\| \leq \left(\frac{1}{2}\right)^{2^k-1} \|x_0 - \zeta\|.$$

La démonstration de ce théorème repose sur les trois lemmes suivants :

Lemme 92. *La fonction*

$$\psi(u) = 1 - 4u + 2u^2$$

décroit de 1 à 0 sur l'intervalle $0 \leq u \leq 1 - \dfrac{\sqrt{2}}{2} = 0.29289\ldots$.

Lemme 93. *Soient $x,\ x_1 \in U$ avec*

$$u = \|x_1 - x\|\gamma(f,x) < 1 - \frac{\sqrt{2}}{2}.$$

Alors $Df(x)^{-1}Df(x_1)$ est inversible et

$$\|Df(x_1)^{-1}Df(x)\| \leq \frac{(1-u)^2}{\psi(u)}.$$

Preuve Remarquons que l'hypothèse $\|x_1 - x\|\gamma(f,x) < 1 - \sqrt{2}/2$ fait que $Df(x)$ est un isomorphisme (lorsque $Df(x)$ n'est pas un isomorphisme $\gamma(f,x)$ est égal à ∞ par définition) et entraîne que $x_1 \in U$ par la Proposition 90 et l'hypothèse $B\left(x, \left(1 - \frac{\sqrt{2}}{2}\right) R(f,x)\right) \subset U$. Ce lemme est une conséquence du Lemme 86. Le développement de Taylor de $Df(x_1)$ au voisinage de x est donné par

$$Df(x_1) = Df(x) + \sum_{k=1}^{\infty} \frac{D^{k+1}f(x)}{k!}(x_1 - x)^k$$

de sorte que

$$Df(x)^{-1}Df(x_1) = \mathrm{id}_{\mathbb{E}} + \sum_{k=1}^{\infty} Df(x)^{-1}\frac{D^{k+1}f(x)}{k!}(x_1 - x)^k.$$

En passant aux normes, on obtient :

$$\|Df(x)^{-1}Df(x_1) - \mathrm{id}_{\mathbb{E}}\| \leq \sum_{k=1}^{\infty}(k+1)\|Df(x)^{-1}\frac{D^{k+1}f(x)}{(k+1)!}\|\ \|(x_1 - x)\|^k$$

et, compte tenu des définitions de $\gamma(f,x)$ et u,

$$\|Df(x)^{-1}Df(x_1) - \mathrm{id}_{\mathbb{E}}\| \leq \sum_{k=1}^{\infty}(k+1)u^k = \frac{1}{(1-u)^2} - 1.$$

Cette dernière quantité est < 1 parce que $u < 1 - \sqrt{2}/2$; ainsi le Lemme 86 s'applique, prouve que $Df(x)^{-1}Df(x_1)$ est un isomorphisme et donne l'estimation voulue

$$\|Df(x_1)^{-1}Df(x)\| \leq \frac{1}{1 - \left(\frac{1}{(1-u)^2} - 1\right)} = \frac{(1-u)^2}{\psi(u)}. \ \square$$

Lemme 94. *Soit $\zeta \in U$ tel que $f(\zeta) = 0$ et que $Df(\zeta)$ soit inversible. Soit $x \in U$ tel que*

$$u = \|x - \zeta\|\gamma(f,\zeta) < \frac{5 - \sqrt{17}}{4} = 0.21922\ldots.$$

Alors, pour tout $k \geq 0$,

$$\|N_f^k(x) - \zeta\| \leq \left(\frac{u}{\psi(u)}\right)^{2^k - 1}\|x - \zeta\|.$$

Preuve Elle consiste à écrire le développement de Taylor de $f(x)$ et de $Df(x)$ au point ζ puis de celui de

$$Df(x)(x-\zeta) - f(x) = \sum_{k=1}^{\infty}(k-1)\frac{D^k f(\zeta)}{k!}(x-\zeta)^k.$$

L'hypothèse faite et le lemme précédent prouvent que $Df(\zeta)^{-1}Df(x)$ est un isomorphisme. On en déduit que

$$\begin{aligned} N_f(x) - \zeta &= Df(x)^{-1}Df(\zeta)Df(\zeta)^{-1}(Df(x)(x-\zeta) - f(x)) \\ &= Df(x)^{-1}Df(\zeta)\sum_{k=1}^{\infty}(k-1)Df(\zeta)^{-1}\frac{D^k f(\zeta)}{k!}(x-\zeta)^k. \end{aligned}$$

On majore la norme de cette expression en utilisant le Lemme 93 et la définition de $\gamma(f,\zeta)$:

$$\begin{aligned} \|N_f(x) - \zeta\| &\leq \|Df(x)^{-1}Df(\zeta)\| \sum_{k=1}^{\infty}(k-1)\ \|Df(\zeta)^{-1}\frac{D^k f(\zeta)}{k!}\|\ \|x-\zeta\|^k \\ &\leq \frac{(1-u)^2}{\psi(u)}\sum_{k=1}^{\infty}(k-1)u^{k-1}\|z-\zeta\| \\ &= \frac{(1-u)^2}{\psi(u)}\left(\frac{1}{(1-u)^2} - \frac{1}{1-u}\right)\|x-\zeta\| \\ &= \frac{u}{\psi(u)}\|x-\zeta\|. \end{aligned}$$

Nous terminons la preuve de ce lemme en raisonnant par récurrence sur k : il faut donc vérifier que

$$\|N_f(x) - \zeta\|\gamma(f,\zeta) < \frac{5-\sqrt{17}}{4}.$$

De l'inégalité $u < (5-\sqrt{17})/4 < 1 - \sqrt{2}/2$ on déduit que $u/\psi(u) < 1$ de sorte que

$$\|N_f(x) - \zeta\|\gamma(f,\zeta) \leq \frac{u}{\psi(u)}\|x-\zeta\|\gamma(f,\zeta) < u < \frac{5-\sqrt{17}}{4}$$

et le tour est joué. □

Preuve du Théorème 91. C'est une conséquence immédiate du Lemme 94 et de l'inégalité $\dfrac{u}{\psi(u)} \leq \dfrac{1}{2}$ lorsque $u \leq \dfrac{3-\sqrt{7}}{2}$. □

Nous allons maintenant prouver, dans le cadre de la théorie alpha, l'équivalent du Théorème 88. Ce théorème provient de Kim [29], [30], qui traite le cas des polynômes d'une variable complexe et Smale [50] pour le cas général.

Définition 95. *Notons*

$$\beta(f,x) = \|Df(x)^{-1}f(x)\|$$

la longueur de la correction de Newton et

$$\alpha(f,x) = \beta(f,x)\gamma(f,x) = \|Df(x)^{-1}f(x)\| \sup_{k\geq 2} \left\| Df(x)^{-1}\frac{D^k f(x)}{k!}\right\|^{\frac{1}{k-1}}.$$

Théorème 96. *(Théorème alpha de Smale.) Il existe une constante $\alpha_0 > 0$ ayant la propriété suivante. Pour tout $x \in U$ qui vérifie $\alpha(f,x) < \alpha_0$ il existe un zéro ζ de f tel que*

$$\|\zeta - x\| \leq 1.63281\ldots\beta(f,x)$$

et

$$1.63281\ldots = \sum_{k=0}^{\infty} \frac{1}{2^{2^k-1}}.$$

De plus, la suite de Newton $x_{k+1} = N_f(x_k)$ avec $x_0 = x$ est définie et vérifie

$$\|x_k - \zeta\| \leq \left(\frac{1}{2}\right)^{2^k-1} \|x_0 - \zeta\|$$

pour tout $k \geq 0$.

La preuve que nous donnons ici de ce théorème (ce n'est pas la seule possible) repose sur trois arguments. Le premier est une borne sur la norme de la dérivée de l'opérateur de Newton :

$$\|DN_f(y)\| \leq 2\alpha(f,y),$$

le second est une estimation de $\alpha(f,y)$ en termes de $\alpha(f,x)$ et $r > 0$ pour tout $y \in \bar{B}(x,r)$, qui permet de donner une constante de contraction pour N_f sur cette boule et le troisième est l'application du théorème des approximations successives à cette situation.

Lemme 97. *Soient $x,\ x_1 \in U$ avec $u = \|x - x_1\|\gamma(f,x) < 1 - (\sqrt{2}/2)$. Alors, pour tout $k \geq 2$,*

- $\left\| Df(x_1)^{-1}\frac{D^k f(x_1)}{k!}\right\| \leq \frac{1}{\psi(u)}\left(\frac{\gamma(f,x)}{1-u}\right)^{k-1}$,
- $\|Df(x)^{-1}f(x_1)\| \leq \beta(f,x) + \frac{\|x_1 - x\|}{1-u}$.

Preuve Pour prouver la première assertion nous utilisons un développement de Taylor en x pour $D^k f(x_1)$ et nous le composons à gauche par $Df(x_1)^{-1}$.

Cela donne

$$Df(x_1)^{-1}\frac{D^k f(x_1)}{k!} = Df(x_1)^{-1}Df(x)\sum_{l=0}^{\infty} Df(x)^{-1}\frac{D^{k+l}f(x)}{k!l!}(x_1-x)^l.$$

En passant aux normes, on a

$$\left\|Df(x_1)^{-1}\frac{D^k f(x_1)}{k!}\right\| \le \|Df(x_1)^{-1}Df(x)\| \times \sum_{l=0}^{\infty}\frac{(k+l)!}{k!l!}\left\|Df(x)^{-1}\frac{D^{k+l}f(x)}{(k+l)!}\right\| \|x_1-x\|^l$$

et, à l'aide du Lemme 93, on obtient

$$\begin{aligned}\left\|Df(x_1)^{-1}\frac{D^k f(x_1)}{k!}\right\| &\le \frac{(1-u)^2}{\psi(u)}\sum_{l=0}^{\infty}\frac{(k+l)!}{k!l!}\gamma(f,x)^{k+l-1}\|x_1-x\|^l \\ &= \frac{(1-u)^2}{\psi(u)}\gamma(f,x)^{k-1}\frac{1}{(1-u)^{k+1}}\end{aligned}$$

ce qui prouve la première assertion. Pour la seconde, par un argument désormais familier,

$$Df(x)^{-1}f(x_1) = Df(x)^{-1}f(x) + (x_1-x) + \sum_{k=2}^{\infty} Df(x)^{-1}\frac{D^k f(x)}{k!}(x_1-x)^k,$$

ce qui donne l'estimation suivante

$$\begin{aligned}\|Df(x)^{-1}f(x_1)\| &\le \|Df(x)^{-1}f(x)\| + \|x_1-x\| \\ &\quad + \sum_{k=2}^{\infty}\left\|Df(x)^{-1}\frac{D^k f(x)}{k!}\right\| \|x_1-x\|^k \\ &\le \beta(f,x) + \|x_1-x\| + \sum_{k=2}^{\infty}\gamma(f,z)^{k-1}\ \|x_1-x\|^k \\ &= \beta(f,x) + \|x_1-x\|\left(1+\left(\frac{1}{1-u}-1\right)\right) \\ &= \beta(f,x) + \frac{\|x_1-x\|}{1-u}.\ \square\end{aligned}$$

Lemme 98. *Soient $x,\ x_1 \in U$ avec $u = \|x-x_1\|\gamma(f,x) < 1-(\sqrt{2}/2)$. Alors, pour tout $k \ge 2$,*

- $\beta(f,x_1) \le \dfrac{1-u}{\psi(u)}((1-u)\beta(f,x) + \|x_1-x\|)$,
- $\gamma(f,x_1) \le \dfrac{\gamma(f,x)}{(1-u)\psi(u)}$,
- $\alpha(f,x_1) \le \dfrac{(1-u)\alpha(f,x)+u}{\psi(u)^2}$.

Preuve Pour β on utilise les Lemmes 93, 97 et l'estimation suivante :

$$\begin{aligned}\beta(f,x_1) &= \|Df(x_1)^{-1}f(x_1)\| \le \|Df(x_1)^{-1}Df(x)\|\ \|Df(x)^{-1}f(x_1)\| \\ &\le \frac{(1-u)^2}{\psi(u)}\left(\beta(f,x)+\frac{\|x_1-x\|}{1-u}\right).\end{aligned}$$

L'estimation sur γ est une conséquence du Lemme 97 :

$$\begin{aligned}\gamma(f,x_1) &= \sup_{k\ge 2}\left\|Df(x_1)^{-1}\frac{D^k f(x_1)}{k!}\right\|^{\frac{1}{k-1}} \\ &\le \sup_{k\ge 2}\left(\frac{1}{\psi(u)}\right)^{\frac{1}{k-1}}\frac{\gamma(f,x)}{1-u} = \frac{\gamma(f,x)}{(1-u)\psi(u)}.\end{aligned}$$

En effet, pour $u < 1-\sqrt{2}/2$ on a $\psi(u) < 1$ et ce sup est atteint pour $k=2$.

La troisième inégalité est obtenue en multipliant les deux premières entre-elles. □

Lemme 99. *Pour tout $x \in U$, $\|DN_f(x)\| \le 2\alpha(f,x)$.*

Preuve La dérivée de l'opérateur de Newton est donnée par

$$\begin{aligned}DN_f(x) &= D(x) - D(Df(x)^{-1})f(x) - Df(x)^{-1}Df(x) \\ &= \mathrm{id}_{\mathbb{E}} + Df(x)^{-1}D^2f(x)Df(x)^{-1}f(x) - \mathrm{id}_{\mathbb{E}}\end{aligned}$$

d'où

$$\begin{aligned}\|DN_f(x)\| &= \|Df(x)^{-1}D^2f(x)Df(x)^{-1}f(x)\| \\ &\le \|Df(x)^{-1}D^2f(x)\|\ \|Df(x)^{-1}f(x)\| \\ &\le 2\gamma\beta = 2\alpha.\ \square\end{aligned}$$

Théorème 100. *Soient $r>0$, α_0 et $x \in U$ qui vérifient les conditions suivantes :*

- $u_0 = r\gamma(f,x) < 1 - \dfrac{\sqrt{2}}{2}$,
- $\alpha(f,x) \le \alpha_0$,
- $\lambda = 2\dfrac{(1-u_0)\alpha_0 + u_0}{\psi(u_0)^2} < 1$,
- $\alpha_0 + \lambda u_0 \le u_0$.

Alors N_f est une contraction de $\bar{B}(x,r)$ dans elle-même, de constante de contraction λ. Il existe donc un unique zéro ζ de f dans cette boule et pour tout $x_0 \in \bar{B}(x,r)$ la suite de Newton $x_{k+1} = N_f(x_k)$ initialisée en x_0 converge vers ζ.

Preuve C'est une conséquence du Corollaire 5 dont nous allons vérifier les hypothèses. D'une part, pour tout $x_1 \in \bar{B}(x,r)$, puisque $u = \|x - x_1\|\gamma(f,x) \leq r\gamma(f,x) < 1 - (\sqrt{2}/2)$, par les lemmes 98 et 99

$$\|DN_f(x_1)\| \leq 2\alpha(f,x_1) \leq 2\frac{(1-u)\alpha(f,x)+u}{\psi(u)^2} = \leq 2\frac{(1-u_0)\alpha_0+u_0}{\psi(u_0)^2} = \lambda < 1$$

et donc N_f est une contraction de constante λ. On aura $N_f(\bar{B}(x,r)) \subset \bar{B}(x,r)$ si $\lambda r + \|x - N_f(x)\| \leq r$ c'est à dire si $\lambda u_0 + \|x - N_f(x)\|\gamma(f,x) \leq u_0$, donnée par $\lambda u_0 + \alpha_0 \leq u_0$ qui est notre hypothèse. □

Les valeurs numériques $u_0 = 0.06$ et $\alpha_0 = 0.04$ conduisent à la valeur $\lambda = 0.33163\ldots < 1/2$. De plus, pour tout $x_0 \in \bar{B}(x, u_0/\gamma(f,x))$, et pour le zéro ζ de f contenu dans cette boule, on a

$$\|x_0 - \zeta\| \leq \|x_0 - x\| + \|x - \zeta\| \leq \frac{2u_0}{\gamma(f,x)}.$$

On déduit de cette inégalité et du Lemme 98 la suivante :

$$\|x_0 - \zeta\|\gamma(f,\zeta) \leq \frac{2u_0\gamma(f,\zeta)}{\gamma(f,x)} \leq \frac{2u_0}{\psi(u_0)(1-u_0)} = 0.16639\ldots < \frac{3-\sqrt{7}}{2}.$$

Autrement dit, la boule $\bar{B}(x, u_0/\gamma(f,x))$ est contenue dans $\bar{B}(\zeta, (3-\sqrt{7})/2\gamma(f,\zeta))$. Dans cette boule, l'opérateur de Newton est une contraction de constante $\leq 1/2$ comme nous l'avons vu au Théorème 91. Nous venons de prouver le théorème suivant :

Théorème 101. *(Théorème alpha robuste) Il existe des constantes positives u_0 et α_0 telles que : si $x \in U$ vérifie $\alpha(f,x) \leq \alpha_0$ alors, il existe un unique zéro ζ de f vérifiant $\|\zeta - x\| \leq u_0/\gamma(f,x)$. De plus*

$$\bar{B}\left(x, \frac{u_0}{\gamma(f,x)}\right) \subset \bar{B}\left(\zeta, \frac{3-\sqrt{7}}{2\gamma(f,\zeta)}\right)$$

et N_f est une contraction de $\bar{B}(x, u_0/\gamma(f,x))$ de constante de contraction au plus $1/2$.

Preuve du Théorème 96. On applique le théorème précédent au centre x de la boule. □

Peut-on préciser les constantes u_0 et α_0 du Théorème alpha robuste ? Cette question a été étudiée par Wang et Han dans [54] qui donnent la réponse suivante

Théorème 102. *(Wang-Han) Pour tout* $\alpha \in \left[0, 3-2\sqrt{2}\right]$, *la quantité* $(1+\alpha)^2 - 8\alpha$ *décroît de* 1 *à* 0. *Posons*

$$q = \frac{1-\alpha-\sqrt{(1+\alpha)^2-8\alpha}}{1-\alpha+\sqrt{(1+\alpha)^2-8\alpha}}.$$

On a

$$\begin{array}{ll} 0 \le q < 1 & si \;\; 0 \le \alpha < 3-2\sqrt{2}, \\ q = 1 & si \;\; 0 \le \alpha = 3-2\sqrt{2}. \end{array}$$

Pour tout $x \in U$ *tel que* $\alpha = \alpha(f,x) \le 3-2\sqrt{2}$, *il existe un et un seul zéro* ζ *de* f *tel que*

$$\|\zeta - x\| \le \frac{1+\alpha-\sqrt{(1+\alpha)^2-8\alpha}}{4\gamma(f,x)}.$$

De plus, la suite de Newton $x_{k+1} = N_f(x_k)$, $x_0 = x$, *est définie et vérifie*

$$\begin{array}{ll} \|\zeta - x_k\| \le \dfrac{1+\alpha-\sqrt{(1+\alpha)^2-8\alpha}}{4\gamma(f,x)} q^{2^k-1} & si \;\; 0 \le \alpha < 3-2\sqrt{2}, \\ \|\zeta - x_k\| \le \dfrac{2-\sqrt{2}}{2\gamma(f,x)} \left(\dfrac{1}{2}\right)^k & si \;\; \alpha = 3-2\sqrt{2}, \end{array}$$

pour tout $k \ge 0$.

Ce théorème admet les deux corollaires suivants :

Corollaire 103. *Pour tout* $x \in U$ *tel que* $\alpha = \alpha(f,x) \le 3-2\sqrt{2}$, *il existe un et un seul zéro* ζ *de* f *tel que*

$$\|\zeta - x\| \le \frac{2-\sqrt{2}}{2\gamma(f,x)}.$$

De plus, la suite de Newton $x_{k+1} = N_f(x_k)$, $x_0 = x$, *est définie et converge vers* ζ.

Preuve $(2-\sqrt{2})/2$ est le maximum de $(1+\alpha-\sqrt{(1+\alpha)^2-8\alpha})/4$ lorsque $\alpha \in \left[0, 3-2\sqrt{2}\right]$. □

Corollaire 104. *Pour tout* $x \in U$ *tel que* $\alpha = \alpha(f,x) \le \dfrac{13-3\sqrt{17}}{4}$, *la suite de Newton* $x_{k+1} = N_f(x_k)$, $x_0 = x$, *converge vers un zéro* ζ *de* f *et de plus*

$$\|\zeta - x_k\| \le \frac{5-\sqrt{17}}{4\gamma(f,x)} \left(\frac{1}{2}\right)^{2^k-1}$$

pour tout $k \ge 0$.

Preuve C'est une conséquence du théorème précédent obtenue en prenant $q = 1/2$. Ceci impose la condition $\alpha \leq 13 - 3\sqrt{17}/4$. L'expression $1 + \alpha - \sqrt{(1+\alpha)^2 - 8\alpha}$ est alors majorée par $5 - \sqrt{17}$. □

Remarque 3. La preuve du théorème de Wang-Han n'est pas donnée ici. Elle repose sur une technique très astucieuse de suites majorantes. Nous renvoyons le lecteur intéressé à l'article original.
Le Corollaire 104 est à comparer au Théorème 101.
Nous donnons ci-dessous une démonstration du Corollaire 103 : voir celle du Théorème 115.

3.4 Exemples

3.4.1 Calcul des racines carrées

Le procédé suivant, pour le calcul du nombre $\sqrt{a}$, consiste en l'itération définie par

$$x_{k+1} = \frac{1}{2}\left(x_k + \frac{a}{x_k}\right).$$

Cette formule est attribuée à Héron d'Alexandrie, grec du premier siècle, mais elle était déjà connue des babyloniens 300 à 400 années avant.

Il s'agit de la méthode de Newton appliquée à $f(x) = x^2 - a$:

$$x_{k+1} = x_k - \frac{x_k^2 - a}{2x_k}.$$

Notons que cette suite possède trois points fixes qui sont $\pm\sqrt{a}$ et l'infini qui est un point fixe répulsif. L'étude de ce dernier point fixe se fait en 0 via le changement de variable $X = 1/x$.

Quelles sont les propriétés de convergence de la suite (x_k) ? Lorsque x est un grand nombre positif, la quantité $\frac{1}{2}(x + a/x)$ est approximativement égale à $x/2$: la suite (x_k) qui démarre en un grand x_0 se comporte comme une suite géométrique de raison $1/2$. Il y a donc convergence linéaire et non pas quadratique. Lorsque x est proche de $\sqrt{a}$ l'approximation ci-dessus n'est plus valide. Le Théorème 85 décrit un intervalle centré en $\sqrt{a}$ et contenu dans le bassin de convergence quadratique qui est ici égal à :

$$\left[\frac{\sqrt{a}}{2}, \frac{3\sqrt{a}}{2}\right].$$

On voit donc qu'il faut nuancer l'affirmation « la méthode de Newton a une convergence quadratique » et bien distinguer le bassin d'attraction de $\sqrt{a}$ qui est défini par

$$BA(\sqrt{a}) = \{x_0 \ : \ (x_k) \to \sqrt{a}\},$$

ici égal à l'intervalle $]0, \infty[$, du bassin de convergence quadratique que l'on peut définir par

$$BAQ(\sqrt{a}) = \{x_0 \ : \ \|x_k - \sqrt{a}\| \leq \left(\frac{1}{2}\right)^{2^k - 1} \|x_0 - \sqrt{a}\|\}$$

et qui est contenu dans le précédent.

3.4.2 Equations du second degré

Posons

$$f(z) = az^2 + bz + c$$

où $a \neq 0$, b, c sont des nombres complexes et $z \in \mathbb{C}$. Lorsque $\Delta = b^2 - 4ac \neq 0$ cette équation possède deux racines distinctes que l'on note r_1 et r_2. Nous allons étudier la méthode de Newton appliquée à ce cas.

Une première réduction consiste à prendre $a = 1$. Elle ne change rien à l'affaire puisque $N_{\lambda f} = N_f$ pour tout scalaire $\lambda \neq 0$. Ainsi $f(z) = (z - r_1)(z - r_2)$. Soit $g(z)$ la transformation homographique suivante

$$g(z) = \frac{z - r_1}{z - r_2}, \quad g^{-1}(z) = \frac{zr_2 - r_1}{z - 1},$$

qui est prolongée sur la sphère de Riemann par $g(r_2) = \infty$ et $g(\infty) = 1$. De façon similaire,

$$N_f(z) = \frac{1}{2} \frac{x^2 - r_1 r_2}{x - \frac{r_1 + r_2}{2}}$$

est prolongé à cette sphère par : $N_f((r_1 + r_2)/2) = \infty$ et $N_f(\infty) = \infty$.

Par ce changement de variable, l'opérateur de Newton devient l'élévation au carré :

$$g \circ N_f \circ g^{-1}(z) = z^2.$$

Posons $g^{-1}(z_k) = x_k$. Puisque $x_k = N_f(x_{k-1}) = N_f^k(x_0)$ nous obtenons pour la suite (z_k) :

$$z_k = g \circ N_f \circ g^{-1}(z_{k-1}) = g \circ N_f^k \circ g^{-1}(z_0) = z_0^{2^k}.$$

Cette suite converge vers 0 si et seulement si $|z_0| < 1$, circule sur le cercle unité si $|z_0| = 1$ et converge vers l'infini si $|z_0| > 1$. Revenons par g^{-1} à la suite de Newton : l'image de 0 est r_1, celle de ∞ est r_2, le cercle unité est transformé en la médiatrice $\mathcal{M}$ du segment $[r_1, r_2]$, l'intérieur du cercle en le demi-plan qui contient r_1 et enfin l'extérieur du cercle en le demi-plan qui contient r_2. Nous en déduisons le résultat suivant :

- Si $x_0 \in \mathcal{M}$ la suite de Newton $x_k = N_f(x_{k-1})$ reste enfermée dans $\mathcal{M}$,
- Si $x_0 \in \mathcal{M}(r_1)$ (resp. $x_0 \in \mathcal{M}(r_2)$), le demi-plan ouvert délimité par $\mathcal{M}$ qui contient r_1 (resp. r_2), la suite (x_k) converge vers r_1 (resp. r_2).

Pour en finir avec cet exemple, il faut noter que le disque donné par le Théorème 85 et contenu dans le bassin de convergence quadratique de r_1 a pour rayon $\frac{|r_1 - r_2|}{4}$, c'est-à-dire la moitié de la distance de r_1 à la médiatrice $\mathcal{M}$.

3.4.3 Equations du troisième degré

Nous avons vu que pour les équations du second degré, sauf pour un ensemble de conditions initiales de mesure nulle (la médiatrice du segment qui relie les deux racines), les suites de Newton sont toujours convergentes. Ce résultat n'est pas général et dès le degré trois on trouve des polynômes pour lesquels il existe un ensemble ouvert $U \subset \mathbb{C}$ tel que les suites $(N_p^k(x))_k$ ne convergent pas quelque soit $x \in U$. Un exemple est donné par

$$p(x) = x^3 - 2x + 2$$

pour lequel l'opérateur de Newton

$$N_p(x) = x - \frac{x^3 - 2x + 2}{3x^2 - 2}$$

possède le cycle de période 2 : $N_p(0) = 1$, $N_p(1) = 0$. Ce cycle est super-attractif puisque 0 est un point fixe super-attractif de $N_p^2 = N_p \circ N_p$. Ainsi, pour tout x dans un voisinage de 0, la suite de Newton $(N_p^k(x))_k$ est captée par le cycle et ne peut donc converger vers une des racines.

3.4.4 Comment calculer toutes les racines d'un polynôme ?

La méthode que nous allons présenter ici a pour but le calcul de toutes les racines d'un polynôme $p(x)$ de degré d à coefficients complexes. Cette méthode est due à Hubbard, Schleicher et Sutherland [25] et consiste à construire un nombre fini de points dans le plan complexe tels que les suites de Newton partant de ces points convergent vers toutes les racines de $p(x)$. Autrement dit, ces points sont suffisament bien répartis dans le plan complexe pour que les bassins d'attraction des racines en contiennent au moins un. De plus, la construction de cet ensemble de points est indépendante du polynôme $p(x)$, elle ne dépend que de d.

Notons $\mathcal{P}_d$ l'ensemble des polynômes unitaires et de degré d

$$p(x) = x^d + a_{d-1}x^{d-1} + \ldots + a_1 x + a_0$$

et dont toutes les racines sont dans le disque unité $|r| \leq 1$. On peut toujours se ramener à ce cas par un changement d'échelle. La borne de Cauchy pour

le maximum des modules des racines de $p(x)$ est :

$$|r| \leq 1 + \max_{1 \leq k \leq d} |a_k|,$$

et celle de Montel :

$$|r| \leq \left(1 + \sum_{k=1}^{d} |a_k|^2\right)^{1/2}.$$

On peut donc par une homothétie ramener les racines dans le disque unité.

Théorème 105. *Pour tout $d \geq 2$ il existe un ensemble $\mathcal{S}_d$ qui consiste en au plus $1.11d(\log d)^2$ points de $\mathbb{C}$ avec la propriété suivante : pour tout polynôme $p(x) \in \mathcal{P}_d$ et pour toute racine r de ce polynôme, il existe un point $x \in \mathcal{S}_d$ pour lequel la suite de Newton $(N_f^k(x))$ converge vers r. Pour les polynômes dont toutes les racines sont réelles, il y a un ensemble analogue avec au plus $1.3d$ points.*

Le facteur multiplicatif $1.11(\log d)^2$ entre le nombre maximum de racines et le nombre de suites considérées n'est pas très grand. On ignore s'il peut être abaissé à $C \log d$ pour une constante C convenable.

Construction de $\mathcal{S}_d$. C'est une grille constituée de $s = \lceil 0.26632 \log d \rceil$ cercles centrés en 0 et de $n = \lceil 8.32547 d \log d \rceil$ points sur chacun de ces cercles ($\lceil x \rceil$ est le plus petit entier $\geq x$). Posons

$$r_k = \left(1 + \sqrt{2}\right)\left(\frac{d-1}{d}\right)^{\frac{2k-1}{4s}} \quad \text{et} \quad \theta_j = \frac{2\pi j}{n},$$

avec $1 \leq k \leq s$ et $0 \leq j \leq n-1$. La grille $\mathcal{S}_d$ consiste en les points $r_k \exp(i\theta_j)$.

Cette construction est fondée sur le fait remarquable suivant : les bassins d'attraction des racines d'un polynôme pour la méthode de Newton sont tous adhérents au point à l'infini, qui est lui un point fixe répulsif. Ce sont des « canaux », qui ne peuvent pas être partout trop minces et qui vont des racines à l'infini. Un cercle de rayon assez grand va tous les couper et si l'on prend assez de points sur un tel cercle il y en aura un dans chaque « canal ». Le procédé est raffiné en prenant plusieurs cercles et moins de points sur chacun d'eux. Le nombre de cercles est égal à 1 jusqu'au degré ≤ 42, 2 cercles pour $43 \leq d \leq 1825$ puis 3 cercles pour $d \leq 78015$.

3.4.5 La méthode de Weierstrass pour le calcul simultané des racines d'un polynôme

Rappelons tout d'abord la définition des fonctions symétriques : étant donné un vecteur $r \in \mathbb{C}^d$, $d \geq 0$, et un entier $k \geq 0$ on définit la fonction symétrique

$\sigma_k(r)$ par $\sigma_0(r) = 1$, $\sigma_k(r) = 0$ si $k > d$ et

$$\sigma_k(r) = \sum_{1 \leq i_1 < \ldots < i_k \leq d} r_{i_1} \ldots r_{i_k}$$

lorsque $1 \leq k \leq d$.

Considérons le polynôme à coefficients complexes

$$p(z) = z^d - a_1 z^{d-1} + a_2 z^{d-2} \cdots + (-1)^k a_k z^{d-k} + \cdots + (-1)^d a_d.$$

Notons aussi r_k, $1 \leq k \leq d$, ses racines, chacune comptée autant de fois que sa multiplicité. Comme on a aussi

$$p(z) = \prod_{k=1}^{d} (z - r_k)$$

les coefficients de $p(z)$ sont reliés aux racines via leurs fonctions symétriques :

$$\sigma_k(r) = a_k, \quad 1 \leq k \leq d.$$

On utilisera les notations suivantes : r est le vecteur colonne dont les entrées sont les r_k, $\sigma_0(r) = 1$, $\sigma_k(\hat{r}_l)$ est la fonction symétrique relative au vecteur r privé de sa l-ième composante :

$$\sigma_k(\hat{r}_l) = \sigma_k(r_1, \ldots, r_{l-1}, r_{l+1}, \ldots, r_d),$$

et enfin $\Sigma(r)$ (resp. A) est le vecteur colonne dont les entrées sont les $\sigma_k(r)$ (resp. a_k).

La recherche de toutes les racines de $p(z)$ revient à résoudre le système $\Sigma(r) = A$. Ce système possède $d!$ solutions qui sont toutes obtenues par permutation des coordonnées du vecteur r. La méthode de Weierstrass consiste à calculer une solution du système $\Sigma(r) = A$ en utilisant la méthode de Newton. Ceci définit un nouvel opérateur

$$W(r) = N_{\Sigma - A}(r)$$

dont les coordonnées seront notées $W_i(r)$. Un petit miracle se produit : on peut donner une expression analytique pour cet opérateur :

Proposition 106. *L'opérateur $W(r)$ est défini pour tout r dont les coordonnées sont deux à deux distinctes. Dans ce cas*

$$W_i(r) = r_i - \frac{p(r_i)}{\displaystyle\prod_{\substack{1 \leq k \leq d \\ k \neq i}} (r_i - r_k)}.$$

L'algorithme de Weierstrass consiste, à partir d'un vecteur initial $w^0 \in \mathbb{C}^d$ dont toutes les coordonnées sont distinctes, à calculer la suite de vecteurs $w^k = W(w^{k-1})$. Cette méthode est facile à implémenter et donne de bon résultats numériques au moins pour des polynômes dont les racines sont bien séparées.

La preuve de cette proposition repose sur la méthode d'interpolation de Lagrange. Nous en donnons ici les grandes lignes.

Première étape : Calcul de la dérivée de Σ. Le résultat est le suivant

$$D\Sigma(r) = \begin{pmatrix} D\sigma_1(r) \\ D\sigma_2(r) \\ \vdots \\ D\sigma_d(r) \end{pmatrix} = \begin{pmatrix} \sigma_0(\hat{r}_1) & \sigma_0(\hat{r}_2) & \dots & \sigma_0(\hat{r}_d) \\ \sigma_1(\hat{r}_1) & \sigma_1(\hat{r}_2) & \dots & \sigma_1(\hat{r}_d) \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{d-1}(\hat{r}_1) & \sigma_{d-1}(\hat{r}_2) & \dots & \sigma_{d-1}(\hat{r}_d) \end{pmatrix}.$$

Cela se démontre en utilisant la formule suivante :

$$\sigma_k(r) = \sigma_k(\hat{r}_k) + r_k \sigma_{k-1}(\hat{r}_k).$$

Deuxième étape : $D\Sigma(r)$ est-elle inversible ? Nous allons prouver que

$$D(d,r) = \det(D\Sigma(r)) = \prod_{1 \le i < j \le d} (r_i - r_j)$$

de sorte que $D\Sigma(r)$ est inversible, c'est-à-dire $W(r)$ définie, si et seulement si les r_i sont deux à deux distincts. Pour prouver cette formule, on note que la première ligne de $D\Sigma(r)$ est constituée de $\sigma_0(\hat{r}_i) = 1$ et les $d-1$ autres lignes par des polynômes en les r_i dont les degrés partiels sont

$$\deg(\sigma_i(\hat{r}_j), r_k) \le 1.$$

Considérons $D(d,r)$ comme un polynôme en la variable r_1. On vient de voir que son degré est $\le d-1$. On voit aussi que si l'on donne à r_1 les valeurs $r_2, \dots, r_d$, deux colonnes du déterminant sont égales et donc ce déterminant est nul. On obtient, en factorisant,

$$D(d,r) = E(r_2, \dots, r_d) \prod_{2 \le j \le d} (r_1 - r_j).$$

Le même raisonnement est appliqué à r_2 et $E(r_2, \dots, r_d)$, puis $r_3 \dots$ et ainsi de suite, pour obtenir

$$D(d,r) = C_d \prod_{1 \le i < j \le d} (r_i - r_j)$$

où C_d est une constante ne dépendant que de d. Nous allons prouver par récurrence qu'elle est égale à 1. C'est vrai pour $d = 1$. Pour passer de $d-1$ à

d, on écrit

$$D(d,r) = D(d,r_1,\ldots,r_{d-1},r_d) = C_d \prod_{1\le i\le d-1}(r_i - r_d) \prod_{1\le i<j\le d-1}(r_i - r_j)$$
$$= C_d \prod_{1\le i\le d-1}(r_i - r_d) D(d-1,r_1,\ldots,r_{d-1}).$$

Pour $r_d = 0$ on obtient

$$D(d,r_1,\ldots,r_{d-1},0) = C_d\ \sigma_{d-1}(r_1,\ldots,r_{d-1})\ D(d-1,r_1,\ldots,r_{d-1}).$$

D'autre part

$$D\Sigma(r_1,\ldots,r_{d-1},0) = \begin{pmatrix} D\Sigma(r_1,\ldots,r_{d-1}) & * \\ 0 & \sigma_{d-1}(r_1,\ldots,r_{d-1}) \end{pmatrix}$$

ce qui prouve que

$$D(d,r_1,\ldots,r_{d-1},0) = D(d-1,r_1,\ldots,r_{d-1})\ \sigma_{d-1}(r_1,\ldots,r_{d-1}).$$

L'hypothèse de récurrence assure que $D(d-1,r_1,\ldots,r_{d-1})\ \sigma_{d-1}(r_1,\ldots,r_{d-1}) \neq 0$, d'où l'égalité $C_d = 1$.

Troisième étape : calcul de l'inverse de $D\Sigma(r)$.

Supposons que les racines r_i de $p(z)$ soient simples. Considérons le polynôme

$$L_j(z) = \frac{\sum_{k=1}^{d}(-1)^{k-1}\sigma_{k-1}(\hat{r}_j)z^{d-k}}{\prod_{k\neq j}(r_j - r_k)}.$$

Il est de degré $d-1$ et, par construction de ses coefficients, il vérifie

$$L_j(r_i) = \frac{\sum_{k=1}^{d}(-1)^{k-1}\sigma_{k-1}(\hat{r}_j)r_i^{d-k}}{\prod_{k\neq j}(r_j - r_k)} = \delta_{ij}$$

pour tout $1 \le i,j \le d$. C'est un polynôme d'interpolation de Lagrange associé aux noeuds d'interpolation r_i. Cette dernière égalité peut être vue comme le terme général du produit de matrices suivant :

$$W\ D\Sigma(r)\ \Delta = I_d$$

où $W_{ij} = (-1)^{j-1}r_i^{d-j}$, Δ est la matrice diagonale dont les entrées sont

$$\Delta_{jj} = \left(\prod_{k\neq j}(r_j - r_k)\right)^{-1}$$

et I_d la matrice identité. Ceci prouve que $D\Sigma(r) = W^{-1}\Delta^{-1}$ ou bien que $D\Sigma(r)^{-1} = \Delta W$, c'est-à-dire

$$D\Sigma(r)^{-1}_{ij} = \frac{(-1)^{j-1} r_i^{d-j}}{\prod_{k\neq i}(r_i - r_k)}.$$

Quatrième étape : calcul de l'opérateur $N_{\Sigma-A}$.

Par définition, $N_{\Sigma-A}(r) = r - D\Sigma(r)^{-1}(\Sigma(r) - A)$ dont la $i-$ème composante est

$$N_{\Sigma-A}(r)_i = r_i - \sum_{j=1}^{d} \frac{(-1)^{j-1} r_i^{d-j}(\sigma_j(r) - a_j)}{\prod_{k\neq i}(r_i - r_k)}.$$

Pour simplifier cette expression, on note que

$$r_i^d + \sum_{j=1}^{d} (-1)^j \sigma_j(r) r_i^{d-j} = 0$$

par définition des fonctions symétriques des racines, de sorte que

$$\sum_{j=1}^{d} (-1)^{j-1} r_i^{d-j} \sigma_j(r) = r_i^d$$

et donc

$$N_{\Sigma-A}(r)_i = r_i - \frac{p(r_i)}{\prod_{k\neq i}(r_i - r_k)}$$

qui est le résultat cherché.

3.4.6 Le problème symétrique des valeurs propres

Soit A une matrice réelle, symétrique et de taille $n \times n$. Ses valeurs propres sont réelles, nous les notons $\lambda_1, \ldots, \lambda_n$ et il existe une base orthonormée de vecteurs propres correspondants notés $v_1, \ldots, v_n$. Si V désigne la matrice orthogonale dont les colonnes sont les v_i, on a $A = VDV^T$ où $D = \mathrm{Diag}(\lambda_1, \ldots, \lambda_n)$.

Le problème symétrique des valeurs propres consiste à calculer les λ_i ainsi que les v_i c'est-à-dire résoudre le système d'équations $Av = \lambda v$. Ce système contient $n+1$ inconnues et n équations, il est donc sous-déterminé. On rajoute l'équation manquante en normalisant le vecteur propre ce qui conduit au système

$$\begin{aligned} (\lambda I_n - A)v &= 0, \\ \frac{1}{2}(\|v\|^2 - 1) &= 0. \end{aligned}$$

que nous noterons $F\begin{pmatrix} v \\ \lambda \end{pmatrix} = 0$. La présence du facteur $1/2$ est purement cosmétique et a pour but de disparaitre dans les dérivations futures. Lorsque les valeurs propres de A sont distinctes, les vecteurs propres correspondants sont n droites vectorielles distinctes qui coupent la sphère unité en $2n$ points. Ceci prouve que le système $F\begin{pmatrix} v \\ \lambda \end{pmatrix} = 0$ possède $2n$ solutions distinctes dans ce cas. Le cas d'une matrice non symétrique, où la structure complexe risque d'intervenir, est plus délicat puisque l'intersection d'une droite vectorielle contenue dans $\mathbb{C}^n$ avec la sphère unité ne produit plus deux points diamétralement opposés comme dans le cas réel mais un grand cercle sur cette sphère. Une statégie possible consiste plutôt à remplacer l'équation $\|v\| = 1$ par l'équation linéaire $\langle v, a\rangle = 1$ où a est un vecteur pris au hasard dans $\mathbb{C}^n$.

Notre objectif, dans ce paragraphe, est l'étude de la méthode de Newton dans ce contexte. Nous allons donc commencer par élucider le calcul de DF^{-1}.

Proposition 107. *Pour tout $v \in \mathbb{R}^n$ et $\lambda \in \mathbb{R}$ la dérivée de F est donnée par*

$$DF\begin{pmatrix} v \\ \lambda \end{pmatrix} = \begin{pmatrix} \lambda I_n - A & v \\ v^T & 0 \end{pmatrix}.$$

Lorsque v est un vecteur propre associé à la valeur propre λ, $DF\begin{pmatrix} v \\ \lambda \end{pmatrix}$ est inversible si et seulement si cette valeur propre est simple. Dans ce cas et si de plus $\|v\| = 1$, notons $v^\perp$ le sous-espace de $\mathbb{R}^n$ orthogonal à v, $\Pi_{v^\perp}$ la projection orthogonale sur ce sous-espace et $(\lambda I_n - A)|_{v^\perp}$ la restriction de $\lambda I_n - A$ à ce sous-espace. On a

$$\left\| DF\begin{pmatrix} v \\ \lambda \end{pmatrix}^{-1} \right\| = \max\left(1, \left\| \left(\Pi_{v^\perp} \circ (\lambda I_n - A)|_{v^\perp}\right)^{-1} \right\|\right).$$

Preuve Calculons la dérivée de F :

$$DF\begin{pmatrix} v \\ \lambda \end{pmatrix}\begin{pmatrix} \dot{v} \\ \dot{\lambda} \end{pmatrix} = \begin{pmatrix} (\lambda I_n - A)\dot{v} + \dot{\lambda}x \\ v^T\dot{v} \end{pmatrix} = \begin{pmatrix} \lambda I_n - A & v \\ v^T & 0 \end{pmatrix}\begin{pmatrix} \dot{v} \\ \dot{\lambda} \end{pmatrix}.$$

Passons à la seconde assertion. Supposons que $Av_1 = \lambda_1 v_1$ et $Ve_1 = v_1$ où $e_1, \ldots, e_n$ désigne la base canonique de $\mathbb{R}^n$. Utilisant la décomposition $A = VDV^T$ on a

$$\begin{pmatrix} \lambda_1 I_n - A & v_1 \\ v_1^T & 0 \end{pmatrix} = \begin{pmatrix} V & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} \lambda_1 I_n - D & e_1 \\ e_1^T & 0 \end{pmatrix}\begin{pmatrix} V^T & 0 \\ 0 & 1 \end{pmatrix}.$$

Il est alors facile de voir que le déterminant de cette matrice est égal à $-(\lambda_1 - \lambda_2)\ldots(\lambda_1 - \lambda_n)$. Il est non nul si et seulement si λ_1 est simple. Considérons

maintenant l'équation

$$\begin{pmatrix} \lambda_1 I_n - A & v_1 \\ v_1^T & 0 \end{pmatrix} \begin{pmatrix} y \\ \mu \end{pmatrix} = \begin{pmatrix} z \\ \nu \end{pmatrix}.$$

Posons $y = \alpha v_1 + y_1$ et $z = \beta v_1 + z_1$ où y_1 et z_1 sont orthogonaux à v_1. Rappelons que $\|v_1\| = 1$. Cette équation matricielle s'écrit

$$(\lambda_1 I_n - A)y + \mu v_1 = z,$$

$$\langle y, v_1 \rangle = \nu$$

ou encore

$$(\lambda_1 I_n - A)(\alpha v_1 + y_1) + \mu v_1 = \beta v_1 + z_1,$$

$$\langle \alpha v_1 + y_1, v_1 \rangle = \nu$$

de sorte que $(\lambda_1 I_n - A)y_1 = z_1$, $\mu = \beta$ et $\alpha = \nu$. Comme y_1 et $z_1 \in v_1^\perp$ on obtient $y_1 = \left(\Pi_{v_1^\perp} \circ (\lambda_1 I_n - A)|_{v_1^\perp} \right)^{-1} z_1$. Il est alors facile de prouver la dernière assertion. □

Théorème 108. *Soit $\lambda \in \mathbb{R}$ une valeur propre simple de A associée au vecteur propre $v \in \mathbb{R}^n$ tel que $\|v\| = 1$. La suite de Newton associée à F et aux données initiales $\lambda_0 \in \mathbb{R}$ et $v_0 \in \mathbb{R}^n$:*

$$\begin{pmatrix} v_{k+1} \\ \lambda_{k+1} \end{pmatrix} = \begin{pmatrix} v_k \\ \lambda_k \end{pmatrix} - DF \begin{pmatrix} v_k \\ \lambda_k \end{pmatrix}^{-1} F \begin{pmatrix} v_k \\ \lambda_k \end{pmatrix}$$

converge quadratiquement vers $\begin{pmatrix} v \\ \lambda \end{pmatrix}$ au sens du Théorème 91 dès lors que

$$\left(\|v - v_0\|^2 + |\lambda - \lambda_0|^2 \right)^{1/2} \leq \frac{3 - \sqrt{7}}{\sqrt{2} \max \left(1, \left\| (\Pi_{v^\perp} \circ (\lambda I_n - A)|_{v^\perp})^{-1} \right\| \right)}.$$

Preuve C'est une conséquence du Théorème 91 et de l'estination

$$\gamma \left(F, \begin{pmatrix} v \\ \lambda \end{pmatrix} \right) \leq \frac{\sqrt{2}}{2} \max \left(1, \left\| (\Pi_{v^\perp} \circ (\lambda I_n - A)|_{v^\perp})^{-1} \right\| \right)$$

que nous devons établir. Puisque F est polynomiale de degré 2, on a

$$\begin{aligned} \gamma \left(F, \begin{pmatrix} v \\ \lambda \end{pmatrix} \right) &= \frac{1}{2} \left\| DF \begin{pmatrix} v \\ \lambda \end{pmatrix}^{-1} D^2 F \begin{pmatrix} v \\ \lambda \end{pmatrix} \right\| \\ &\leq \frac{1}{2} \left\| DF \begin{pmatrix} v \\ \lambda \end{pmatrix}^{-1} \right\| \left\| D^2 F \begin{pmatrix} v \\ \lambda \end{pmatrix} \right\|. \end{aligned}$$

Cette dérivée seconde vaut

$$D^2F\begin{pmatrix} v \\ \lambda \end{pmatrix}\begin{pmatrix} x \\ \alpha \end{pmatrix}\begin{pmatrix} y \\ \beta \end{pmatrix} = \begin{pmatrix} \alpha y + \beta x \\ \langle x, y \rangle \end{pmatrix}$$

de sorte que

$$\begin{aligned}\left\| D^2F\begin{pmatrix} v \\ \lambda \end{pmatrix}\begin{pmatrix} x \\ \alpha \end{pmatrix}\begin{pmatrix} y \\ \beta \end{pmatrix}\right\|^2 &= \|\alpha y + \beta x\|^2 + |\langle x, y \rangle|^2 \\ &\leq (|\alpha|\|y\| + |\beta|\|x\|)^2 + \|x\|^2\|y\|^2 \\ &\leq (\|x\|^2 + |\alpha|^2)(\|y\|^2 + |\beta|^2) + \|x\|^2\|y\|^2 \\ &\leq 2(\|x\|^2 + |\alpha|^2)(\|y\|^2 + |\beta|^2).\end{aligned}$$

Ce calcul prouve que

$$\left\| D^2F\begin{pmatrix} v \\ \lambda \end{pmatrix}\right\| = \max_{\substack{\|(x,\alpha)\| = 1 \\ \|(y,\beta)\| = 1}} \left\| D^2F\begin{pmatrix} v \\ \lambda \end{pmatrix}\begin{pmatrix} x \\ \alpha \end{pmatrix}\begin{pmatrix} y \\ \beta \end{pmatrix}\right\| \leq \sqrt{2}$$

d'où le résultat. □

3.4.7 L'équation de Riccati algébrique

L'équation de Riccati, que nous considérons ici, est donnée par

$$R(X) = A^TX + X^TA - XBB^TX + Q = 0$$

où A, Q et X sont des matrices $n \times n$ réelles, Q et X symétriques et B est une matrice $n \times p$ réelle et de rang p. Cette équation provient de problèmes de contrôle optimal et, sous certaines hypothèses relatives à l'observabilité et la contrôlabilité du problème, il possède une unique solution X définie positive.

Nous envisageons ici le calcul de cette solution par la méthode de Newton. Notons $\mathcal{S}_n$ l'espace des matrices $n \times n$ réelles et symétriques que l'on munit de la structure euclidienne canonique. La dérivée de $R : \mathcal{S}_n \to \mathcal{S}_n$ est donnée par $DR(X) : \mathcal{S}_n \to \mathcal{S}_n$,

$$DR(X)(S) = (A - BB^TX)^TS + S(A - BB^TX).$$

C'est un opérateur de Liapunov associé à la matrice $A - BB^TX$. Plus généralement on pose

$$\mathcal{L}_M(S) = M^TS + SM.$$

L'opérateur $\mathcal{L}_M$ est inversible si et seulement si la matrice M est elle même inversible, voir Stewart-Sun [51] Chap. V, sect. 1. 2.

Afin d'utiliser le Théorème 91 qui décrit le bassin quadratique d'attraction pour la méthode de Newton, nous devons estimer l'invariant $\gamma(R, X)$ qui vaut ici

$$\gamma(R, X) = \frac{1}{2} \left\| DR(X)^{-1} D^2 R(X) \right\|$$

puisque R est une application polynomiale de degré 2. Nous allons estimer séparément les quantités $\|DR(X)^{-1}\|$ et $\left\|D^2 R(X)\right\|$.

Lemme 109. *Soit M une matrice $n \times n$, réelle et stable c'est-à-dire dont les valeurs propres ont une partie réelle négative. Pour toute matrice B l'équation*

$$\mathcal{L}_M(X) = M^T X + XM = B$$

a pour solution

$$X = \int_0^\infty \exp(M^T t) B \exp(Mt) dt.$$

Preuve Cette intégrale est convergente parce que M est stable. Ce point sera rendu plus clair au cours de la démonstration du prochain lemme. De plus

$$\begin{aligned} M^T X + XM &= M^T \left(\int_0^\infty \exp(M^T t) B \exp(Mt) dt \right) \\ &\quad + \left(\int_0^\infty \exp(M^T t) B \exp(Mt) dt \right) M \\ &= \int_0^\infty M^T \exp(M^T t) B \exp(Mt) + \exp(M^T t) B \exp(Mt) M dt \\ &= \int_0^\infty \frac{d}{dt} \left(\exp(M^T t) B \exp(Mt) \right) dt = B. \ \square \end{aligned}$$

Lemme 110. *Sous l'hypothèse du lemme précédent*

$$\|\mathcal{L}_M^{-1}\| \leq \int_0^\infty \|\exp(Mt)\|^2 dt.$$

Si de plus M est diagonalisable et $M = PDP^{-1}$ avec D diagonale alors

$$\|\mathcal{L}_M^{-1}\| \leq \frac{\kappa(P)^2}{2 \min\limits_{1 \leq k \leq n} |\Re(\lambda_k)|}$$

où $\kappa(P)$ est le conditionnement de la matrice P, λ_k, $1 \leq k \leq n$, les valeurs propres de M et $\Re(\lambda_k)$ la partie réelle.

Preuve On a $\mathcal{L}_M^{-1}(B) = X$, dont l'expression est donnée au lemme précédent, d'où

$$\|\mathcal{L}_M^{-1}(B)\| = \|X\| = \left\| \int_0^\infty \exp(M^T t) B \exp(Mt) dt \right\| \leq \|B\| \int_0^\infty \|\exp(Mt)\|^2 dt$$

ce qui établit la première inégalité. Lorsque M est diagonalisable on a

$$X = \int_0^\infty P^{-T} \exp(Dt) P^T B P \exp(Dt) P^{-1} dt$$

de sorte que

$$\|X\| \leq \kappa(P)^2 \|B\| \int_0^\infty \|\exp(Dt)\|^2 dt.$$

Par ailleurs, puisque $D = \mathrm{Diag}(\lambda_k)$,

$$\|\exp(Dt)\| = \max_{1\leq k\leq n} |\exp(\lambda_k t)| = \max_{1\leq k\leq n} \exp(\Re(\lambda_k) t).$$

Comme ces parties réelles sont négatives, les intégrales correspondantes sont convergentes et

$$\int_0^\infty \|\exp(Dt)\|^2 dt = \max_{1\leq k\leq n} \int_0^\infty \exp(2\Re(\lambda_k) t) dt = \max_{1\leq k\leq n} \frac{1}{2|\Re(\lambda_k)|}. \ \square$$

Lemme 111. $\left\|D^2 R(X)\right\| = \|BB^T\| \leq \|B\|^2$.

Preuve C'est une conséquence immédiate de l'égalité

$$D^2 R(X)(S,T) = -SBB^T T. \ \square$$

A partir de ces calculs le Théorème 91 se transcrit de la façon suivante :

Théorème 112. *Soit $X \in \mathcal{S}_n$ solution de l'équation de Riccati*

$$R(X) = A^T X + X^T A - XBB^T X + Q = 0.$$

Supposons que $A - BB^T X$ soit stable. Alors, pour toute matrice $X_0 \in \mathcal{S}_n$ telle que

$$\|X_0 - X\|_F \leq \frac{3 - \sqrt{7}}{\|BB^T\| \int_0^\infty \|\exp(Mt)\|^2 dt}$$

la suite de Newton $X_{k+1} = N_R(X_k)$ est définie et converge vers X. Elle s'obtient en résolvant l'équation de Liapunov

$$(A - BB^T X_k)^T X_{k+1} + X_{k+1}(A - BB^T X_k) + X_k BB^T X_k + Q = 0.$$

De plus

$$\|X_k - X\| \leq \left(\frac{1}{2}\right)^{2^k - 1} \|X_0 - X\|.$$

Les résultats classiques concernant l'équation de Riccati sont présentés dans le livre de Bittanti, Laub et Willems [5]. Les auteurs décrivent la résolution numérique de cette équation via la méthode de Newton d'un point de vue différent de celui que nous adoptons ici.

L'hypothèse de stabilité faite sur la matrice $A - BB^T X$ est réaliste dans le contexte des problèmes de contrôle. De façon plus précise on a le résultat suivant que nous empruntons au livre cité ci-dessus :

Théorème 113. *On dit que (A, B) est stabilisable s'il existe une matrice F telle que $A + BF$ soit stable. On dit que (C, A) est détectable s'il existe une matrice L telle que $A+LC$ soit stable. Lorsque (A, B) est stabilisable et (C, A) détectable, l'équation de Riccati*

$$R(X) = A^T X + X^T A - XBB^T X + Q = 0$$

admet une unique solution semi-définie positive X. De plus $A - BB^T X$ est une matrice stable.

3.4.8 Sur la séparation des racines d'un système

Un nombre qui intervient souvent dans l'analyse de certains algorithmes est le nombre de séparation. Etant donné une application analytique entre deux espaces de Banach,

$$f : \mathbb{E} \to \mathbb{F}$$

et une solution de ce système : $f(\zeta) = 0$, on note

$$\text{sep } (f, \zeta) = \inf_{f(\zeta')=0,\ \zeta' \neq \zeta} \|\zeta' - \zeta\| .$$

Nous pouvons estimer ce nombre de séparation à l'aide de l'invariant γ :

Théorème 114. *Lorsque $f(\zeta) = 0$ et que $Df(\zeta)$ est un isomorphisme on a*

$$sep\ (f, \zeta) \geq \frac{1}{2\gamma(f, \zeta)}.$$

Preuve Raisonnons par l'absurde et supposons qu'il existe $\zeta' \neq \zeta$ tel que $f(\zeta') = 0$ et

$$\|\zeta' - \zeta\| < \frac{1}{2\gamma(f, \zeta)}.$$

Par la proposition 90 la série de Taylor de f au point ζ converge en ζ' de sorte que

$$f(\zeta') = f(\zeta) + Df(\zeta)(\zeta' - \zeta) + \sum_{k=2}^{\infty} \frac{D^k f(\zeta)}{k!}(\zeta' - \zeta)^k.$$

Comme $f(\zeta) = f(\zeta') = 0$ on obtient

$$\zeta' - \zeta = -\sum_{k=2}^{\infty} Df(\zeta)^{-1} \frac{D^k f(\zeta)}{k!} (\zeta' - \zeta)^k$$

d'où

$$1 \leq \sum_{k=2}^{\infty} \left\| Df(\zeta)^{-1} \frac{D^k f(\zeta)}{k!} \right\| \|\zeta' - \zeta\|^{k-1} \leq \sum_{k=2}^{\infty} \gamma(f,\zeta)^{k-1} \|\zeta' - \zeta\|^{k-1}$$

$$= \frac{\|\zeta' - \zeta\| \gamma(f,\zeta)}{1 - \|\zeta' - \zeta\| \gamma(f,\zeta)}.$$

Puisque, par hypothèse, $\|\zeta' - \zeta\| \gamma(f,\zeta) < 1/2$, on obtient

$$1 < \frac{1/2}{1 - 1/2} = 1$$

ce qui est absurde. □

L'application du Théorème gamma (Théorème 91) donne immédiatement

$$\text{sep } (f,\zeta) > \frac{3 - \sqrt{7}}{2\gamma(f,\zeta)}.$$

La constante $1/2$ donnée ici est bien meilleure.

3.4.9 Séparation des racines via le théorème de Rouché

Nous allons démontrer, par une méthode de séries majorantes, l'énoncé suivant que nous avons déjà rencontré au Corollaire 103.

Théorème 115. *Soient $f : \mathbb{C}^n \to \mathbb{C}^n$ une fonction analytique, $x_0 \in \mathbb{C}^n$ tel que $Df(x_0)$ soit un isomorphisme. Si $\alpha(f, x_0) < 3 - 2\sqrt{2}$, il existe un et seul zéro ζ de f avec*

$$\|x_0 - \zeta\| < \frac{2 - \sqrt{2}}{2\gamma(f, x_0)}.$$

Nous suivons ici une démonstration de Jean-Claude Yakoubsohn basée sur le théorème de Rouché dont voici un énoncé (voir [11] Chap. IV-18, Théorème 2) :

Théorème 116. *Donnons nous un domaine borné $D \subset \mathbb{C}^n$ de frontière de Jordan S et deux applications analytiques f et g définies sur un voisinage ouvert de D et à valeurs dans $\mathbb{C}^n$. Si pour tout $z \in S$ on a*

$$\|f(z)\| > \|g(z)\|$$

alors $f + g$ a autant de zéros (comptés avec multiplicités) que f dans D.

Preuve du Théorème 115. Définissons $g(x) = f(x) - f(x_0)$. Notons que $g(x_0) = 0$, que $Dg(x_0) = Df(x_0)$ est un isomorphisme et que $\gamma(g, x_0) = \gamma(f, x_0)$. Par le Théorème 114, x_0 est le seul zéro de g dans la boule ouverte

$$B\left(x_0, \frac{1}{2\gamma(f, x_0)}\right).$$

Notons enfin que, dans cette boule, la série de Taylor de g est convergente (Proposition 90). Elle est donnée par

$$g(x) = Df(x_0)(x - x_0) + \sum_{k\geq 2} \frac{D^k f(x_0)}{k!}(x - x_0)^k.$$

Nous allons prouver que pour un certain r, $0 < r < 1/2$, on a

$$\left\|Df(x_0)^{-1} f(x) - Df(x_0)^{-1} g(x)\right\| < \left\|Df(x_0)^{-1} g(x)\right\|$$

pour tout x avec

$$\|x - x_0\| = \frac{r}{\gamma(f, x_0)}.$$

Par le théorème de Rouché, $Df(x_0)^{-1} f(x)$ et $Df(x_0)^{-1} g(x)$ auront le même nombre de zéros dans cette boule : 1. Notons que

$$\left\|Df(x_0)^{-1} f(x) - Df(x_0)^{-1} g(x)\right\| = \left\|Df(x_0)^{-1} f(x_0)\right\| = \beta(f, x_0)$$

et que

$$x - x_0 = Df(x_0)^{-1} g(x) - \sum_{k\geq 2} Df(x_0)^{-1} \frac{D^k f(x_0)}{k!}(x - x_0)^k$$

de sorte que

$$\frac{r}{\gamma(f, x_0)} \leq \left\|Df(x_0)^{-1} g(x)\right\| + \sum_{k\geq 2} \gamma(f, x_0)^{k-1} \left(\frac{r}{\gamma(f, x_0)}\right)^k$$

c'est à dire

$$\frac{r}{\gamma(f, x_0)} - \frac{1}{\gamma(f, x_0)} \frac{r^2}{1 - r} \leq \left\|Df(x_0)^{-1} g(x)\right\|.$$

L'hypothèse du théorème de Rouché sera satisfaite si

$$\beta(f, x_0) < \frac{r}{\gamma(f, x_0)} - \frac{1}{\gamma(f, x_0)} \frac{r^2}{1 - r}$$

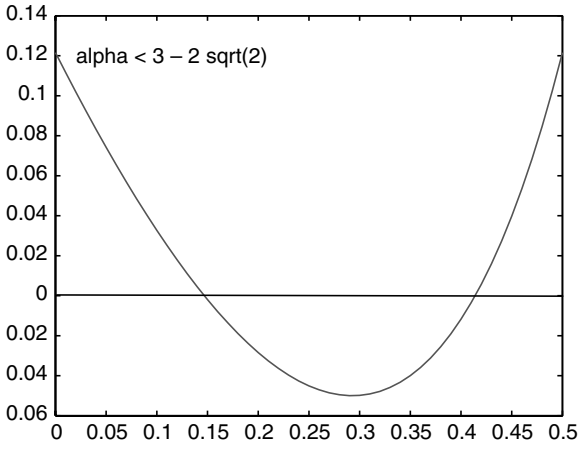


Fig. 3.1. La fonction h.

autrement dit

$$\alpha(f, x_0) < r - \frac{r^2}{1-r}.$$

Supposons que $0 < \alpha < 3 - 2\sqrt{2}$. La fonction

$$h(r) = \alpha - r + \frac{r^2}{1-r}$$

est convexe sur l'intervalle $[0, 1[$, elle y possède deux zéros qui sont

$$0 < \frac{\alpha + 1 - \sqrt{\alpha^2 - 6\alpha + 1}}{4} < \frac{\alpha + 1 + \sqrt{\alpha^2 - 6\alpha + 1}}{4} < \frac{1}{2}.$$

On aura

$$\alpha < r - \frac{r^2}{1-r}$$

pour tout r dans l'intervalle ouvert défini par ces deux racines. Donc, si $\alpha(f, x_0) < 3 - 2\sqrt{2}$, pour tout α tel que

$$\alpha(f, x_0) < \alpha < 3 - 2\sqrt{2}$$

et pour tout r avec

$$\frac{\alpha + 1 - \sqrt{\alpha^2 - 6\alpha + 1}}{4} < r < \frac{\alpha + 1 + \sqrt{\alpha^2 - 6\alpha + 1}}{4}$$

f possède un unique zéro ζ qui vérifie

$$\|x_0 - \zeta\| < \frac{r}{\gamma(f, x_0)}.$$

On obtient

$$\|x_0 - \zeta\| \leq \frac{\alpha + 1 - \sqrt{\alpha^2 - 6\alpha + 1}}{4\gamma(f, x_0)}$$

pour tout $\alpha < 3 - 2\sqrt{2}$ d'où, en passant à la limite,

$$\|x_0 - \zeta\| < \frac{2 - \sqrt{2}}{2\gamma(f, x_0)}. \square$$

3.4.10 Une version quantitative du théorème des fonctions implicites

Le théorème des fonctions implicites est l'énoncé suivant :

Théorème 117. *Soient $\mathbb{E}$, $\mathbb{F}$, $\mathbb{G}$ trois espaces de Banach et soit $F : \mathbb{E}\times\mathbb{F} \to \mathbb{G}$ une application de classe $\mathbb{C}^k$, $1 \leq k \leq \infty$, ou bien analytique. Soient $x_0 \in \mathbb{E}$ et $y_0 \in \mathbb{F}$ tels que $F(x_0, y_0) = 0$ et que $D_2F(x_0, y_0) : \mathbb{F} \to \mathbb{G}$ soit un isomorphisme. Il existe un voisinage ouvert V de x_0 et une unique fonction $f : V \to \mathbb{F}$ de classe $\mathbb{C}^k$ ou analytique qui satisfasse*

$$f(x_0) = y_0 \quad et \quad F(x, f(x)) = 0$$

pour tout $x \in V$. De plus

$$Df(x_0) = -D_2F(x_0, y_0)^{-1} D_1F(x_0, y_0).$$

La fonction f est appelée la fonction implicite associée à F et au point (x_0, y_0).

Nous allons voir qu'en utilisant le Corollaire 103 ou bien le Théorème 115 on peut un peu mieux préciser les choses : donner une estimation du voisinage V dont l'existence est affirmée ci-dessus et décrire un procédé itératif pour calculer la fonction implicite. Plus précisément, avec les notations du théorème précédent, on a :

Théorème 118. *Supposons que F soit analytique. Soient $x_0 \in \mathbb{E}$ et $y_0 \in \mathbb{F}$ tels que $F(x_0, y_0) = 0$ et que $D_2F(x_0, y_0) : \mathbb{F} \to \mathbb{G}$ soit un isomorphisme. Notons*

$$\gamma_2(F, x_0, y_0) = \sup_{k\geq 2} \left\| D_2F(x_0, y_0)^{-1} \frac{D^k F(x_0, y_0)}{k!} \right\|^{\frac{1}{k-1}}.$$

La fonction implicite associée à F est définie pour tout $x \in \mathbb{E}$ tel que

$$\|x - x_0\| \leq \frac{3 - 2\sqrt{2}}{(1 + \|D_2F(x_0, y_0)^{-1} D_1F(x_0, y_0)\|^2)^{1/2}\ \gamma_2(F, x_0, y_0)}.$$

Elle vérifie

$$\|f(x) - f(x_0)\| \leq \frac{2-\sqrt{2}}{2\gamma_2(F, x_0, y_0)}$$

pour tout x dans cette boule. De façon plus précise, $(x, f(x))$ est l'unique solution du système

$$\mathcal{F}(u, y) = \begin{pmatrix} u - x \\ F(u, y) \end{pmatrix} = 0$$

dans la boule fermée

$$\bar{B}\left((x_0, y_0), \frac{2-\sqrt{2}}{2\gamma_2(F, x_0, y_0)}\right).$$

De plus, la suite de Newton $(u_{k+1}, y_{k+1}) = N_{\mathcal{F}}(u_k, y_k)$ avec $(u_0, y_0) = (x_0, y_0)$ converge vers $(x, f(x))$.

Preuve Notons que $\mathcal{F}(u, y) = 0$ si et seulement si $u = x$ et $F(x, y) = 0$. L'existence d'une telle solution est donnée par le Corollaire 103 et, pour ce faire, nous calculons les invariants $\alpha(\mathcal{F}, u, y)$, $\beta(\mathcal{F}, u, y)$ et $\gamma(\mathcal{F}, u, y)$. Nous avons

$$D\mathcal{F}(u, y) = \begin{pmatrix} \mathrm{id}_{\mathbb{E}} & 0 \\ D_1F(u, y) & D_2F(u, y) \end{pmatrix}$$

et

$$D\mathcal{F}(u, y)^{-1} = \begin{pmatrix} \mathrm{id}_{\mathbb{E}} & 0 \\ -D_2F(u, y)^{-1}D_1F(u, y) & D_2F(u, y)^{-1} \end{pmatrix}$$

si $D_2F(u, y)$ est inversible. On en déduit l'estimation suivante :

$$\begin{aligned}
\beta(\mathcal{F}, x_0, y_0) &= \|D\mathcal{F}(x_0, y_0)^{-1}\mathcal{F}(x_0, y_0)\| \\
&= \left\| \begin{pmatrix} x_0 - x \\ -D_2F(x_0, y_0)^{-1}D_1F(x_0, y_0)(x_0 - x) \end{pmatrix} \right\| \\
&= \left(\|x - x_0\|^2 + \left\| D_2F(x_0, y_0)^{-1}D_1F(x_0, y_0)(x - x_0) \right\|^2 \right)^{1/2} \\
&\leq \left(1 + \left\| D_2F(x_0, y_0)^{-1}D_1F(x_0, y_0) \right\|^2 \right)^{1/2} \|x - x_0\|.
\end{aligned}$$

De plus, pour tout $k \geq 2$,

$$\begin{aligned}
&D\mathcal{F}(u, y)^{-1}D^k\mathcal{F}(u, y) \\
&\quad = \begin{pmatrix} \mathrm{id}_{\mathbb{E}} & 0 \\ -D_2F(u, y)^{-1}D_1F(u, y) & D_2F(u, y)^{-1} \end{pmatrix} \begin{pmatrix} 0 \\ D^kF(u, y) \end{pmatrix}
\end{aligned}$$

de sorte que

$$\gamma(\mathcal{F}, x_0, y_0) = \sup_{k\geq 2} \left\| D_2F(x_0,y_0)^{-1} \frac{D^kF(x_0,y_0)}{k!} \right\|^{1/k-1} = \gamma_2(F, x_0, y_0).$$

La condition $\alpha(\mathcal{F}, x_0, y_0) \leq 3 - 2\sqrt{2}$, qui assure l'existence d'un unique zéro pour $\mathcal{F}$ proche de (x_0, y_0), est satisfaite dès que

$$\|x - x_0\| \left(1 + \left\|D_2F(x_0,y_0)^{-1} D_1F(x_0,y_0)\right\|^2\right)^{1/2} \gamma_2(F, x_0, y_0) \leq 3 - 2\sqrt{2}$$

qui est précisément l'hypothèse du théorème. Nous savons, par le Corollaire 103, qu'il existe alors un unique zéro du système $\mathcal{F}(u, y) = 0$. Par unicité, ce zéro est $(x, f(x))$ et il vérifie

$$\|(x, f(x)) - (x_0, y_0)\| \leq \frac{2-\sqrt{2}}{2\gamma(\mathcal{F}, x_0, y_0)} = \frac{2-\sqrt{2}}{2\gamma_2(F, x_0, y_0)}.$$

Par ce même corollaire, la suite de Newton associée au système $\mathcal{F}(u, y) = 0$ et initialisée en (x_0, y_0) converge vers $(x, f(x))$. □

4

La méthode de Newton pour des systèmes sous-déterminés

4.1 Introduction

Dans les chapitres précédents nous nous sommes attachés au calcul de points, donnés comme points fixes ou bien solutions de systèmes d'équations. Nous souhaitons faire de même avec des objets plus compliqués dont la dimension n'est plus 0 comme c'est le cas des points mais positive comme celle de courbes et de surfaces. Un tel ensemble est décrit ici par une équation

$$V = f^{-1}(0), \quad f : \mathbb{E} \to \mathbb{F},$$

où f est une application de classe C^p entre deux espaces de Hilbert.

Dans un premier temps nous allons prouver que, pour $x \in V$ dont la dérivée $Df(x)$ est surjective, on peut paramétrer V au voisinage de ce point : V est localement le graphe d'une fonction définie dans un voisinage de x contenu dans $x + \ker Df(x)$, l'espace tangent à V en x, et qui prend ses valeurs dans un voisinage de x dans $x + \ker Df(x)^{\perp}$, l'espace normal à V à x. Notons le changement de point de vue : l'équation définissant V est exprimée dans les coordonnées de $\mathbb{E}$ alors que la paramétrisation est décrite dans le repère $x + \big(\ker Df(x) \times \ker Df(x)^{\perp}\big)$. Nous donnons ensuite des estimations métriques de cette paramétrisation en termes de l'invariant $\gamma(f, x)$, défini ci-dessous, qui est une mesure de la courbure de V au point x et nous montrons comment calculer numériquement cette paramétrisation.

Le second problème abordé est celui du calcul de V. Nous allons décrire un opérateur de Newton dont l'ensemble des points fixes est V. Pour définir cet opérateur, reprenons l'équation linéarisée associée à f et à un point $x \in \mathbb{E}$:

$$0 = f(y) \approx f(x) + Df(x)(y - x).$$

Lorsque $Df(x)$ est surjective l'équation $f(x) + Df(x)(y - x) = 0$ possède des solutions : on choisit celle de norme minimale c'est à dire

$$N_f(x) = x - Df(x)^{\dagger} f(x)$$

où l'on désigne par $Df(x)^\dagger$ l'inverse généralisé de $Df(x)$. Cette stratégie a été introduite pour la première fois par Ben-Israel [3] 1966 pour des systèmes d'équations sous-déterminés, puis par Allgower et Georg [1] 1990, Beyn [2] 1993, Shub et Smale [45] 1996.

La troisième partie de ce chapitre est consacrée à l'étude de ce nouvel opérateur de Newton et à celle de l'opérateur « limite » :

$$M_f(x) = \lim_{k\to\infty} N_f^k(x)$$

qui, localement, se comporte comme une projection sur V.

Ce chapitre se termine avec deux exemples d'applications. On développe pour les systèmes polynomiaux et pour le problème symétrique des valeurs propres l'étude des erreurs rétrogrades (« backward error analysis » disent les anglophones) : la solution approchée d'un problème est décrite comme la solution exacte d'un problème approché, l'erreur rétrograde est, par définition, la distance entre ces deux problèmes. Nous montrons comment la méthode de Newton permet de calculer de tels problèmes approchés.

4.2 Inverses généralisés

Notons $L : \mathbb{E} \to \mathbb{F}$ un opérateur linéaire et continu entre deux espaces de Hilbert dont l'image est fermée dans $\mathbb{F}$. De ce fait on a deux décompositions en somme directe orthogonale :

$$L : \mathbb{E} = \ker L \oplus (\ker L)^\perp \to \operatorname{im} L \oplus (\operatorname{im} L)^\perp = \mathbb{F}.$$

Notons $i : (\ker L)^\perp \to \mathbb{E}$ l'injection canonique, $\Pi_{(\ker L)^\perp} : \mathbb{E} \to (\ker L)^\perp$ la projection orthogonale de $\mathbb{E}$ sur $(\ker L)^\perp$ et enfin $\Pi_{\operatorname{im} L} : \mathbb{F} \to \operatorname{im} L$ la projection orthogonale de $\mathbb{F}$ sur im L. La restriction de L à $(\ker L)^\perp$ est une bijection entre cet espace et im L, ce qui fait de

$$\mathcal{L} = \Pi_{\operatorname{im} L} \circ L|_{(\ker L)^\perp} : (\ker L)^\perp \to \operatorname{im} L$$

une application linéaire, continue et bijective. Son inverse est aussi continue par le théorème de l'inverse continu : $\mathcal{L}$ est donc un isomorphisme.

Définition 119. *On appelle inverse généralisé de L ou inverse de Moore-Penrose l'application linéaire et continue suivante :*

$$L^\dagger : \mathbb{F} \to \mathbb{E}, \quad L^\dagger = i \circ \left(\Pi_{im\ L} \circ L|_{(\ker L)^\perp}\right)^{-1} \circ \Pi_{im\ L}.$$

Notons que $L^\dagger = L^{-1}$ dès que L est bijectif puisqu'alors $\ker L = \{0\}$, $(\ker L)^\perp = \mathbb{E}$ et $\text{im } L = \mathbb{F}$. Les propriétés essentielles de ce nouvel opérateur sont :

Théorème 120.

1. $\ker L^\dagger = (im\ L)^\perp$ *et* $im\ L^\dagger = (\ker L)^\perp$,
2. $L^\dagger \circ L = \Pi_{(\ker L)^\perp}$ *et* $L \circ L^\dagger = \Pi_{im\ L}$,
3. $L^\dagger \circ L = (L^\dagger \circ L)^*$ *et* $L \circ L^\dagger = (L \circ L^\dagger)^*$, *où* N^* *désigne l'adjoint de* N,
4. $L^\dagger \circ L \circ L^\dagger = L^\dagger$ *et* $L \circ L^\dagger \circ L = L$
5. $L^\dagger$ *est le seul opérateur linéaire et continu* $M : \mathbb{F} \to \mathbb{E}$ *qui vérifie les propriétés suivantes*
   a) $M \circ L$ *et* $L \circ M$ *sont auto-adjoints,*
   b) $M \circ L \circ M = M$ *et* $L \circ M \circ L = L$,
6. *Soit* $M : \mathbb{F} \to \mathbb{E}$ *un opérateur linéaire et continu qui vérifie les deux propriétés suivantes*
   a) $\ker M = (im\ L)^\perp$,
   b) $L \circ M = \Pi_{im\ L}$,
   alors, pour tout $y \in \mathbb{F}$
   $$\|L^\dagger(y)\| \leq \|M(y)\|.$$
7. *Lorsque* L *est surjectif,* $L \circ L^\dagger = id_{\mathbb{F}}$ *et* $L^\dagger = L^* \circ (L \circ L^*)^{-1}$, *où* L^* *est l'adjoint de* L,
8. *Lorsque* L *est injectif,* $L^\dagger \circ L = id_{\mathbb{E}}$ *et* $L^\dagger = (L^* \circ L)^{-1} \circ L^*$.

Preuve La première assertion est une conséquence immédiate de la construction de $L^\dagger$.

Soit $x = \Pi_{\ker L} x + \Pi_{(\ker L)^\perp} x$. On a

$$\begin{aligned} L^\dagger \circ L(x) &= L^\dagger \circ L \circ \Pi_{(\ker L)^\perp}(x) \\ &= i \circ \left(\Pi_{\text{im } L} \circ L|_{(\ker L)^\perp}\right)^{-1} \circ \Pi_{\text{im } L} \circ L \circ \Pi_{(\ker L)^\perp}(x) \\ &= i \circ \Pi_{(\ker L)^\perp}(x) = \Pi_{(\ker L)^\perp}(x) \end{aligned}$$

ce qui prouve que $L^\dagger \circ L(x) = \Pi_{(\ker L)^\perp}$. De façon similaire, soit

$$y = \Pi_{\text{im } L} y + \Pi_{(\text{im } L)^\perp} y = L(x) + \Pi_{(\text{im } L)^\perp} y$$

pour un certain $x \in (\ker L)^\perp$. On a

$$\begin{aligned} L \circ L^\dagger(y) &= L \circ i \circ \left(\Pi_{\text{im } L} \circ L|_{(\ker L)^\perp}\right)^{-1} \circ \Pi_{\text{im } L}(L(x)) \\ &= L \circ i \circ \left(\Pi_{\text{im } L} \circ L|_{(\ker L)^\perp}\right)^{-1} \circ \Pi_{\text{im } L} \circ L|_{(\ker L)^\perp}(x) \\ &= L(x) = \Pi_{\text{im } L} y \end{aligned}$$

et ceci prouve la seconde assertion.

Une projection orthogonale est un endomorphisme auto-adjoint d'où la troisième assertion.

La quatrième est une conséquence facile de la seconde. Pour la cinquième, on procède de la façon suivante :

$$\begin{aligned} M &= MLM = MLL^\dagger LM = (ML)^* L^\dagger (LM)^* = L^* M^* L^\dagger M^* L^* \\ &= L^* M^* (L^\dagger L L^\dagger) M^* L^* = L^* M^* L^\dagger L (L^\dagger L L^\dagger) M^* L^* \\ &= L^* M^* (L^\dagger L)^* L^\dagger (L L^\dagger)^* M^* L^* = (L^* M^* L^*) L^{\dagger^*} L^\dagger L^{\dagger^*} (L^* M^* L^*) \\ &= (L^* L^{\dagger^*}) L^\dagger (L^{\dagger^*} L^*) = (L^\dagger L) L^\dagger (L L^\dagger) = L^\dagger L L^\dagger = L^\dagger. \end{aligned}$$

La sixième assertion se montre ainsi : considérons $y = u + v \in \operatorname{im} L \oplus (\operatorname{im} L)^\perp$. Puisque $L^\dagger$ et M ont même noyau $(\operatorname{im} L)^\perp$, on a $\|L^\dagger y\| = \|L^\dagger u\|$ et $\|My\| = \|Mu\|$, il suffit donc de prouver que $\|L^\dagger y\| \leq \|My\|$ pour tout $y \in \operatorname{im} L$ ou encore que $\|L^\dagger \circ L(x)\| \leq \|M \circ L(x)\|$ pour tout $x \in \mathbb{E}$. Notons que $L(M \circ L(x)) = L(L^\dagger \circ L(x)) = L(x)$ à cause de l'hypothèse faite sur M et de la seconde assertion pour L. Ceci prouve que $M \circ L(x)$ et $L^\dagger \circ L(x)$ sont dans l'image réciproque de $L(x)$ par L. Cette image réciproque est égale à $x + \ker L$. Puisque $L^\dagger \circ L(x) \in (\ker L)^\perp$ c'est le point de cette image réciproque qui est à plus courte distance de 0. D'où l'assertion.

Pour prouver la septième assertion on utilise la cinquième en prenant $M = L^* \circ (L \circ L^*)^{-1}$. La vérification des hypothèses est facile. Le fait que $L \circ L^*$ soit inversible lorsque L est surjectif est classique. Injectivité : si $L \circ L^*(y) = 0$ alors $\langle L \circ L^*(y), z \rangle = 0$ pour tout $z \in \mathbb{F}$ d'où $\langle L^*(y), L^*(z) \rangle = 0$ pour tout $z \in \mathbb{F}$ ce qui pour $z = y$ donne $L^*(y) = 0$. En faisant le produit scalaire avec un $x \in \mathbb{E}$ quelconque on obtient $0 = \langle L^*(y), x \rangle = \langle y, L(x) \rangle$. Puisque L est surjectif on peut choisir x tel que $y = L(x)$ ce qui donne $y = 0$. Surjectivité : nous allons montrer que si $y \in \mathbb{F}$ est orthogonal à l'image de $L \circ L^*$ alors $y = 0$. Par hypothèse $\langle y, L \circ L^*(z) \rangle = 0$ pour tout $z \in \mathbb{F}$, c'est à dire $\langle L^*(y), L^*(z) \rangle = 0$ pour tout z et pour $z = y$ on obtient $L^*(y) = 0$. Multiplions scalairement par un $x \in \mathbb{E}$ quelconque : $0 = \langle L^*(y), x \rangle = \langle y, L(x) \rangle$. Puisque L est surjectif on peut choisir x tel que $y = L(x)$ ce qui donne $y = 0$.

La huitième assertion se prouve par des arguments similaires. □

4.3 Paramétrer une sous-variété

Le concept de sous-variété (voir l'appendice) a déjà été mentionné, dans le contexte d'espaces de Banach, à propos du théorème de la variété stable locale. Nous le retrouvons ici dans des espaces de Hilbert. Cette section donne une première application de l'invariant γ introduit ci-dessous. Le Théorème 124 relie la courbure en x d'une sous-variété donnée par une équation $V = f^{-1}(0)$ à l'invariant $\gamma(f, x)$. Nous montrons ensuite comment calculer une paramétrisation de V.

Notons $f : \mathbb{E} \to \mathbb{F}$ une application analytique entre deux espaces de Hilbert et

$$V = f^{-1}(0) = \{x \in \mathbb{E} \ : \ f(x) = 0\}.$$

Définition 121. *Pour tout $x \in \mathbb{E}$ on pose*

$$\gamma(f,x) = \sup_{k\geq 2} \left\| Df(x)^\dagger \frac{D^k f(x)}{k!} \right\|^{\frac{1}{k-1}}$$

lorsque $Df(x)$ est surjective et $\gamma(f,x) = \infty$ sinon.

Cette définition généralise le Définition 89 : on considère ici des applications dont la dérivée $Df(x)$ est surjective au lieu d'être inversible. De façon tout à fait similaire à la Proposition 90 on a :

Proposition 122. *Le rayon de convergence R de la série de Taylor de f en x vérifie*

$$0 < \frac{1}{\gamma(f,x)} \leq R.$$

Lemme 123. *Soit $x \in \mathbb{E}$ tel que $Df(x)$ soit surjective. Pour tout $y \in \mathbb{E}$ tel que*

$$\nu = \|y-x\|\gamma(f,x) < 1 - \frac{\sqrt{2}}{2}$$

1. *$Df(x)^\dagger Df(y) = \Pi_{(\ker Df(x))^\perp} + B$ où B est un opérateur linéaire et continu tel que*
$$\|B\| \leq \frac{1}{(1-\nu)^2} - 1 < 1,$$
2. *$Df(x)^\dagger Df(y)|_{(\ker Df(x))^\perp} : (\ker Df(x))^\perp \to (\ker Df(x))^\perp$ est un isomorphisme et*
$$\| \left(Df(x)^\dagger Df(y)|_{(\ker Df(x))^\perp} \right)^{-1} \| \leq \frac{(1-\nu)^2}{\psi(\nu)}$$
où $\psi(\nu) = 2\nu^2 - 4\nu + 1$ (cette fonction décroît, sur l'intervalle $[0, 1-\sqrt{2}/2]$, de 1 à 0) ;
3. *$Df(y) : (\ker Df(x))^\perp \to \mathbb{F}$ est un isomorphisme. En conséquence $Df(y) : \mathbb{E} \to \mathbb{F}$ est surjective.*
4. *$\|Df(y)^\dagger Df(x)\| \leq \frac{(1-v)^2}{\psi(v)}$.*

Preuve Nous partons de la formule de Taylor en x pour $Df(y)$:

$$Df(y) = Df(x) + \sum_{k=1}^{\infty} \frac{D^{k+1} f(x)}{k!} (y-x)^k.$$

Composons cette identité à gauche par $Df(x)^\dagger$ et passons aux normes. Puisque $Df(x)^\dagger \circ Df(x) = \Pi_{(\ker Df(x))^\perp}$ (Théorème 120) on a :

$$\|Df(x)^\dagger Df(y) - \Pi_{(\ker Df(x))^\perp}\| \leq$$

$$\sum_{k=1}^{\infty} \left\| Df(x)^\dagger \frac{D^{k+1}f(x)}{k!} \right\| \|y-x\|^k \leq \sum_{k=1}^{\infty} (k+1)\gamma(f,x)^k \|y-x\|^k \leq \frac{1}{(1-\nu)^2} - 1.$$

Ainsi

$$Df(x)^\dagger Df(y) = \Pi_{(\ker Df(x))^\perp} + B$$

où B est un opérateur linéaire tel que

$$\|B\| \leq \frac{1}{(1-\nu)^2} - 1 < 1$$

puisque $\nu < 1 - \sqrt{2}/2$.

Lorsque l'on restreint $Df(x)^\dagger Df(y)$ à $(\ker Df(x))^\perp$ on obtient

$$\begin{aligned} &Df(x)^\dagger Df(y)|_{(\ker Df(x))^\perp} \\ &\quad = \mathrm{id}_{(\ker Df(x))^\perp} + B|_{(\ker Df(x))^\perp} : (\ker Df(x))^\perp \to (\ker Df(x))^\perp \end{aligned}$$

auquel on applique le Lemme 86 : on en déduit que cet opérateur est inversible et on obtient l'inégalité annoncée pour la norme de son inverse.

Pour la troisième assertion, on note que $Df(y)|_{(\ker Df(x))^\perp}$ est injective puisque, par 2, $Df(x)^\dagger Df(y)|_{(\ker Df(x))^\perp}$ est injective. De plus

$$Df(y)|_{(\ker Df(x))^\perp} = Df(x)(Df(x)^\dagger Df(y)|_{(\ker Df(x))^\perp})$$

est la composée de $Df(x)$ qui est surjective et d'un isomorphisme, d'où la surjectivité de $Df(y)|_{(\ker Df(x))^\perp}$ et celle de $Df(y)$.

Pour prouver la quatrième assertion on note que

$$\|Df(y)^\dagger Df(x)\| = \|Df(y)^\dagger Df(x)|_{(\ker Df(x))^\perp}\|$$

et que

$$\begin{aligned} Df(y)^\dagger Df(x)|_{(\ker Df(x))^\perp} &= Df(y)^\dagger Df(y)|_{(\ker Df(x))^\perp} (Df(y)|_{(\ker Df(x))^\perp})^{-1} \\ &\quad \circ Df(x)|_{(\ker Df(x))^\perp} \\ &= \Pi \circ (Df(x)^\dagger Df(y)|_{(\ker Df(x))^\perp})^{-1} \end{aligned}$$

où $\prod$ est la restriction à $(\ker Df(x))^\perp$ de la projection orthogonale sur $(\ker Df(y))^\perp$. On utilise 2 pour conclure. □

Notons $B_1(r)$ et $\bar{B}_1(r)$ (resp. $B_2(r)$ et $\bar{B}_2(r)$) la boule ouverte et la boule fermée de centre 0 et de rayon r dans $E_1 = \ker Df(x)$ (resp. $E_2 = (\ker Df(x))^\perp$), $B(x,r)$ et $\bar{B}(x,r)$ les boules ouvertes et fermées de centre x et de rayon r dans $\mathbb{E}$. On identifiera la somme directe $E_1 \oplus E_2$ au produit cartésien $E_1 \times E_2$. Ainsi $u \in \mathbb{E}$ pourra être vu comme $u = u_1 + u_2 \in E_1 \oplus E_2$ ou bien $u = (u_1, u_2) \in E_1 \times E_2$ suivant le point de vue adopté. Le résultat principal de cette section est le suivant :

Théorème 124. *Soit $x \in V$ tel que $Df(x)$ soit surjective. Alors, $V \cap B\left(x, \frac{2-\sqrt{2}}{2\gamma(f,x)}\right)$ est une sous-variété différentiable de $\mathbb{E}$ et il existe*

une fonction

$$h : \bar{B}_1\left(\frac{3-2\sqrt{2}}{\gamma(f,x)}\right) \to \bar{B}_2\left(\frac{2-\sqrt{2}}{2\gamma(f,x)}\right)$$

telle que

1. *h est analytique,*
2. *$h(0) = 0$ et $Dh(0) = 0$,*
3. *Le graphe de h est tel que*

$$x + graphe(h) = V \cap \bar{B}\left(x, \frac{2-\sqrt{2}}{2\gamma(f,x)}\right) \cap \left(x + \bar{B}_1\left(\frac{3-2\sqrt{2}}{\gamma(f,x)}\right) \times \bar{B}_2\left(\frac{2-\sqrt{2}}{2\gamma(f,x)}\right)\right).$$

   Notons

$$c_1(\lambda) = 1 - \sqrt{\frac{\lambda+1}{2\lambda}}.$$

   Pour tout $1 < \lambda < (23+16\sqrt{2})/17 = 2.68396\ldots$ et pour tout $u \in \bar{B}_1\left(\frac{c_1(\lambda)}{\gamma(f,x)}\right)$ on a
4. *$\|h(u)\| \leq \lambda\|u\|^2\gamma(f,x)$ et*
5. *$\|Dh(u)\| \leq 2\lambda\|u\|\gamma(f,x)$.*

Preuve Supposons pour simplifier que $x = 0$ et donc que $f(0) = 0$. Par le Lemme 123, les Définitions 186 et 188 $Df(y)$ est surjective pour tout $y \in B\left(0, \frac{2-\sqrt{2}}{2\gamma(f,0)}\right)$ et donc

$$V \cap B\left(0, \frac{2-\sqrt{2}}{2\gamma(f,0)}\right)$$

est une sous-variété différentiable de $\mathbb{E}$. Nous allons utiliser le théorème des fonctions implicites du chapitre précédent (Théorème 118 dans le contexte suivant : $f : E_1 \times E_2 \to \mathbb{F}$ et $(x_0, y_0) = (0,0)$. Notons que

$$D_1 f(0,0) = Df(0)|_{E_1} = 0$$

parce que $E_1 = \ker Df(0)$ et que

$$D_2 f(0,0)^{-1} = Df(0)^\dagger$$

puisque $E_2 = (\ker Df(0))^\perp$. Pour cette raison, l'invariant

$$\gamma_2(f,0,0) = \sup_{k\geq 2} \left\| D_2 f(0,0)^{-1} \frac{D^k f(0,0)}{k!} \right\|^{\frac{1}{k-1}}$$

du Théorème 118 est égal à

$$\gamma(f,0) = \sup_{k\geq 2} \left\| Df(0)^\dagger \frac{D^k f(0)}{k!} \right\|^{\frac{1}{k-1}}$$

et

$$\left(1 + \left\| D_2 f(0,0)^{-1} D_1 f(0,0) \right\|^2\right)^{1/2} = 1.$$

Le théorème des fonctions implicites (Théorème 118) prouve l'existence d'une fonction analytique

$$h : \bar{B}_1\left(\frac{3-2\sqrt{2}}{\gamma(f,x)}\right) \to \bar{B}_2\left(\frac{2-\sqrt{2}}{2\gamma(f,x)}\right)$$

qui vérifie $h(0) = 0$, $Dh(0) = 0$ et telle que

$$\text{graphe}(h) = V \cap \bar{B}\left(0, \frac{2-\sqrt{2}}{2\gamma(f,0)}\right) \cap \left(\bar{B}_1\left(\frac{3-2\sqrt{2}}{\gamma(f,0)}\right) \times \bar{B}_2\left(\frac{2-\sqrt{2}}{2\gamma(f,0)}\right)\right).$$

Nous allons estimer $\|Dh(u)\|$ pour tout $u \in B_1\left(\frac{3-2\sqrt{2}}{\gamma(f,0)}\right)$. Notons $v = h(u)$ et $z = (u,v)$. Par le Lemme 123-2, $D_2 f(z) : E_2 \to \mathbb{F}$ est un isomorphisme et, puisque $f(u,h(u)) = 0$ pour tout u, on a

$$Dh(u) = -D_2 f(z)^{-1} D_1 f(z).$$

Par le Lemme 123-1, dont nous utilisons les notations,

$$Df(z) = Df(0) \circ (\Pi_{E_2} + B)$$

et par restriction à E_1 et E_2 :

$$D_1 f(z) = D_2 f(0) \circ B|_{E_1} \text{ et } D_2 f(z) = D_2 f(0) \circ (\mathrm{id}_{E_2} + B|_{E_2}).$$

Notons que $Df(0)|_{E_2}$ est inversible de sorte que

$$Dh(u) = -D_2 f(z)^{-1} D_1 f(z) = -\left(\mathrm{id}_{E_2} + B|_{E_2}\right)^{-1} \circ B|_{E_1}.$$

En utilisant les Lemmes 86 et 123 pour majorer $\| \left(\mathrm{id}_{E_2} + B|_{E_2}\right)^{-1} \|$ on a

$$\begin{aligned}\|Dh(u)\| &\leq \| \left(\mathrm{id}_{E_2} + B|_{E_2}\right)^{-1} \| \|B|_{E_1}\| \\ &\leq \frac{1}{1 - \left(\frac{1}{(1-\nu)^2} - 1\right)} \times \left(\frac{1}{(1-\nu)^2} - 1\right) = \frac{2\nu - \nu^2}{\psi(\nu)}\end{aligned}$$

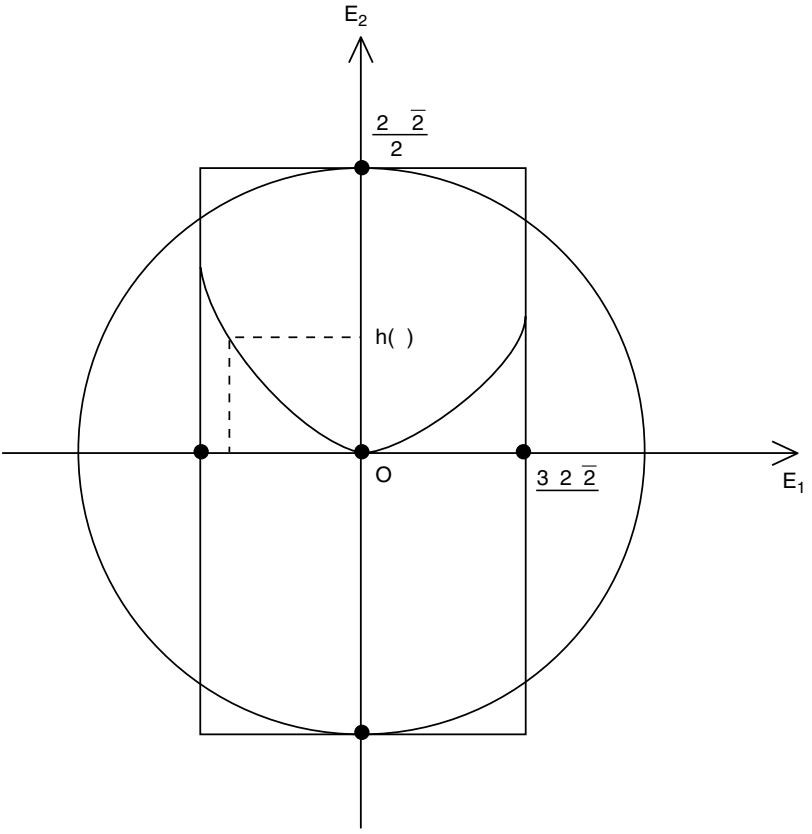


Fig. 4.1. La fonction h.

avec $\nu = \|u\|\gamma(f,0)$ et $\psi(\nu) = 1 - 4\nu + 2\nu^2$. Nous venons de prouver que

$$\|Dh(u)\| \leq \frac{2 - \|u\|}{\psi(\nu)}\|u\|\gamma(f,0) \leq \frac{2}{\psi(\nu)}\|u\|\gamma(f,0)$$

pour tout $u \in B_1$. Pour rendre plus lisible cette inégalité, notons que, pour tout $\lambda > 1$ donné, $2/\psi(\nu) \leq 2\lambda$ dès que

$$0 \leq \nu \leq c_1(\lambda) = 1 - \sqrt{\frac{\lambda + 1}{2\lambda}}.$$

Comme il faut aussi que $\nu = \|u\|\gamma(f,0) < 3 - 2\sqrt{2}$ ceci impose une condition sur λ à savoir

$$\lambda < \frac{23 + 16\sqrt{2}}{17} = 2.68396\ldots$$

En résumé, pour tout $\lambda \in]1, 2.68396[$,

$$\|Dh(u)\| \leq 2\lambda\|u\|\gamma(f,0) \quad \text{dès que} \quad \|u\|\gamma(f,0) \leq c_1(\lambda).$$

Ceci prouve la cinquième assertion. Puisque $h(0) = 0$ on a

$$h(u) = \int_0^1 Dh(tu)(u)dt$$

de sorte que

$$\begin{aligned}\|h(u)\| &\leq \left\| \int_0^1 Dh(tu)(u)dt \right\| \leq \int_0^1 \|Dh(tu)\|\|u\|dt \\ &\leq \int_0^1 2\lambda\|u\|^2 t\gamma(f,0)dt = \lambda\|u\|^2\gamma(f,0)\end{aligned}$$

d'où la quatrième assertion, ce qui achève la preuve de ce théorème. □

Remarque 4. Comment paramétrer V ? Le théorème précédent permet d'envisager le calcul d'une paramétrisation locale de cette sous-variété au voisinage de $x \in V$ de la façon suivante.

- Calculer l'espace tangent en x : $\ker Df(x)$ ainsi que son orthogonal,
- Calculer un majorant de $\gamma(f, x)$,
- Pour tout $u \in B_1\left(\frac{3-2\sqrt{2}}{\gamma(f,x)}\right)$, en vertu du théorème précédent, l'équation $f(x+(u,v)) = 0$ possède une unique solution dans $B\left(x, \frac{2-\sqrt{2}}{2\gamma(f,x)}\right)$ et l'on a $v = h(u)$. Cette solution peut être calculée par la méthode de Newton appliquée au système $f(x+(u,v)) = 0$, où l'inconnue est v, initialisée en $v_0 = 0$. La suite de Newton ainsi construite converge quadratiquement vers $x + (u, h(u))$ en vertu du Théorème 102.

4.4 La méthode de Newton dans le cas surjectif

Soit $f : \mathbb{E} \to \mathbb{F}$ une application analytique entre deux espaces de Hilbert et soit $V = f^{-1}(0)$ l'ensemble des zéros de f. On suppose que, pour tout $\zeta \in V$, $Df(\zeta)$ est surjectif. Ainsi V est une sous-variété de $\mathbb{E}$ et l'espace tangent à V en $\zeta \in V$ est $T_\zeta V = \ker Df(\zeta)$. Cette situation est typiquement celle d'un système d'équations sous-déterminé, $f_i(x_1, \ldots, x_n) = 0$, $1 \leq i \leq m$, où le nombre d'inconnues n est plus grand que le nombre d'équations m.

L'opérateur de Newton est défini par

$$N_f(x) = x - Df(x)^\dagger f(x).$$

Comme dans le cas inversible ses points fixes correspondent aux zéros de f :

Proposition 125. *Soit $\zeta \in \mathbb{E}$ avec $Df(\zeta)$ surjectif. Alors $N_f(\zeta) = \zeta$ si et seulement si $f(\zeta) = 0$.*

Preuve Puisque $Df(\zeta)$ est surjectif, par le Théorème 120, $Df(\zeta)Df(\zeta)^\dagger = \mathrm{id}_{\mathbb{F}}$ de sorte que $Df(\zeta)^\dagger f(\zeta) = 0$ équivaut à $f(\zeta) = 0$. □

Le premier résultat principal de cette section est une version dans le cas surjectif du Théorème 96 énoncé au chapitre précédent. Etant donné un point $x \in \mathbb{E}$, nous énonçons une condition nécessaire portant sur x pour que la suite de Newton correspondante converge quadratiquement vers un zéro de f. Par là même nous prouvons l'existence d'un tel zéro. La démonstration que nous donnons de ce théorème est différente de celle donnée dans le cas inversible.

Définition 126. *Nous noterons $\psi(u) = 1 - 4u + 2u^2$, $\alpha_0 = 0.13071\ldots$ la constante positive pour laquelle $2u \leq \psi(u)^2$ si et seulement si $0 \leq u \leq \alpha_0$ et enfin*

$$1.63281\ldots = \sum_{k=0}^{\infty} \left(\frac{1}{2}\right)^{2^k-1}.$$

Définition 127. *Notons*

$$\beta(f,x) = \|Df(x)^\dagger f(x)\|$$

la longueur de la correction de Newton et

$$\alpha(f,x) = \beta(f,x)\gamma(f,x) = \|Df(x)^\dagger f(x)\| \sup_{k\geq 2} \left\| Df(x)^\dagger \frac{D^k f(x)}{k!} \right\|^{\frac{1}{k-1}}.$$

Théorème 128. *Soit $x \in \mathbb{E}$ tel que $\alpha(f,x) \leq \alpha_0$. La suite de Newton $x_k = N_f^k(x)$ vérifie les propriétés suivantes :*

1. $\|x_{k+1} - x_k\| \leq \left(\frac{1}{2}\right)^{2^k-1} \beta(f,x)$,
2. *Elle converge vers un zéro ζ de f tel que $\|\zeta - x\| \leq 1.63281\ldots\beta(f,x)$,*
3. $\|x_k - \zeta\| \leq 1.63281\ldots\left(\frac{1}{2}\right)^{2^k-1} \beta(f,x)$ *pour tout $k \geq 0$.*

Le second résultat principal de cette section étend au cas surjectif le Théorème 91. Il décrit un « tube » autour de V tel que, pour tout x pris dans ce tube, la suite de Newton initialisée en x converge quadratiquement vers un point de V.

Définition 129. *Nous noterons $u_0 = 0.05992\ldots$ la constante positive pour laquelle $u \leq \alpha_0\psi(u)^2$ et $1.63281\ldots(1-u) \leq 2\psi(u)$ si et seulement si $0 \leq u \leq u_0$.*

Théorème 130. *Considérons l'ensemble suivant*

$$\mathcal{V} = \{x \in \mathbb{E} \quad : \quad \exists \zeta \in V \quad \|x - \zeta\|\gamma(f, \zeta) \leq u_0\}.$$

Soient $x \in \mathcal{V}$ et $\zeta \in V$ tels que $\|x - \zeta\|\gamma(f, \zeta) \leq u_0$. La suite de Newton $x_k = N_f(x)$ vérifie les propriétés suivantes :

1. *Elle converge vers vers un zéro de f que l'on note $M_f(x)$,*
2. *Pour tout $k \geq 0$, $\|x_k - M_f(x)\| \leq 2\left(\frac{1}{2}\right)^{2^k - 1} \|x - \zeta\|$,*
3. *En particulier $\|x - M_f(x)\| \leq 2\|x - \zeta\|$.*

La démonstration de ce résultat repose sur les lemmes suivants.

Lemme 131. *Soient $x,\ x_1 \in \mathbb{E}$ avec $u = \|x - x_1\|\gamma(f, x) < 1 - (\sqrt{2}/2)$ et $Df(x)$ surjectif. Alors,*

- $\beta(f, x_1) \leq \dfrac{1-u}{\psi(u)}((1-u)\beta(f, x) + \|x_1 - x\|)$,
- $\gamma(f, x_1) \leq \dfrac{\gamma(f, x)}{(1-u)\psi(u)}$,
- $\alpha(f, x_1) \leq \dfrac{(1-u)\alpha(f, x) + u}{\psi(u)^2}$.

La preuve du Lemme 131 repose sur l'énoncé suivant :

Lemme 132. *Avec les hypothèses du Lemme 131, pour tout $k \geq 2$, on a :*

- $\left\| Df(x_1)^\dagger \dfrac{D^k f(x_1)}{k!} \right\| \leq \dfrac{1}{\psi(u)} \left(\dfrac{\gamma(f, x)}{1-u} \right)^{k-1}$,
- $\|Df(x)^\dagger f(x_1)\| \leq \beta(f, x) + \dfrac{\|x_1 - x\|}{1-u}$.

Preuve du Lemme 132. Notons que, puisque $Df(x)$ est surjectif et que $u < 1 - (\sqrt{2}/2)$, $Df(x_1)$ est surjectif par le Lemme 123 et $Df(x_1)Df(x_1)^\dagger = \mathrm{id}$ par le Théorème 120.

Pour prouver la première assertion nous utilisons un développement de Taylor en x pour $D^k f(x_1)$ et nous le composons à gauche par $Df(x_1)^\dagger = Df(x_1)^\dagger Df(x)Df(x)^\dagger$. Cela donne

$$Df(x_1)^\dagger \frac{D^k f(x_1)}{k!} = Df(x_1)^\dagger Df(x) \sum_{l=0}^{\infty} Df(x)^\dagger \frac{D^{k+l} f(x)}{k!l!} (x_1 - x)^l.$$

En passant aux normes, on a

$$\left\| Df(x_1)^\dagger \frac{D^k f(x_1)}{k!} \right\| \leq \|Df(x_1)^\dagger Df(x)\| \sum_{l=0}^{\infty} \frac{(k+l)!}{k!l!} \left\| Df(x)^\dagger \frac{D^{k+l} f(x)}{(k+l)!} \right\| \|x_1 - x\|^l.$$

De plus,

$$\|Df(x_1)^\dagger Df(x)\| \leq \frac{(1-u)^2}{\psi(u)}$$

par le Lemme 123. On obtient donc

$$\begin{aligned}\left\|Df(x_1)^\dagger \frac{D^k f(x_1)}{k!}\right\| &\leq \frac{(1-u)^2}{\psi(u)} \sum_{l=0}^{\infty} \frac{(k+l)!}{k!l!} \gamma(f,x)^{k+l-1} \|x_1 - x\|^l \\ &= \frac{(1-u)^2}{\psi(u)} \gamma(f,x)^{k-1} \frac{1}{(1-u)^{k+1}}\end{aligned}$$

ce qui prouve la première assertion. Pour la seconde, le développement de Taylor en x de $f(x_1)$, composé à gauche par $Df(x)^\dagger$ donne

$$\begin{aligned}Df(x)^\dagger f(x_1) = Df(x)^\dagger f(x) + Df(x)^\dagger Df(x)(x_1 - x) \\ + \sum_{k=2}^{\infty} Df(x)^\dagger \frac{D^k f(x)}{k!}(x_1 - x)^k,\end{aligned}$$

Puisque $Df(x)^\dagger Df(x)$ est une projection orthogonale (voir le Théorème 120) on a :

$$\begin{aligned}\|Df(x)^\dagger f(x_1)\| &\leq \|Df(x)^\dagger f(x)\| + \|x_1 - x\| \\ &\quad + \sum_{k=2}^{\infty} \left\|Df(x)^\dagger \frac{D^k f(x)}{k!}\right\| \ \|x_1 - x\|^k \\ &\leq \beta(f,x) + \|x_1 - x\| + \sum_{k=2}^{\infty} \gamma(f,z)^{k-1} \ \|x_1 - x\|^k \\ &= \beta(f,x) + \|x_1 - x\| \left(1 + \left(\frac{1}{1-u} - 1\right)\right) \\ &= \beta(f,x) + \frac{\|x_1 - x\|}{1-u}. \ \square\end{aligned}$$

Preuve du Lemme 131. Pour β on utilise les arguments déjà développés dans la preuve du Lemme 132 pour obtenir :

$$\begin{aligned}\beta(f,x_1) = \|Df(x_1)^\dagger f(x_1)\| &\leq \|Df(x_1)^\dagger Df(x)\| \ \|Df(x)^\dagger f(x_1)\| \\ &\leq \frac{(1-u)^2}{\psi(u)} \left(\beta(f,x) + \frac{\|x_1 - x\|}{1-u}\right).\end{aligned}$$

L'estimation sur γ est une conséquence du Lemme 132 :

$$\begin{aligned}\gamma(f,x_1) = \sup_{k\geq 2} \left\|Df(x_1)^\dagger \frac{D^k f(x_1)}{k!}\right\|^{\frac{1}{k-1}} &\leq \sup_{k\geq 2} \left(\frac{1}{\psi(u)}\right)^{\frac{1}{k-1}} \frac{\gamma(f,x)}{1-u} \\ = \frac{\gamma(f,x)}{(1-u)\psi(u)}.\end{aligned}$$

En effet, pour $u < 1 - \sqrt{2}/2$ on a $\psi(u) < 1$ et ce sup est atteint pour $k = 2$. La troisième inégalité est obtenue en multipliant les deux premières entre-elles. □

Lemme 133. *Soit $x \in \mathbb{E}$ tel que $Df(x)$ soit surjectif et soit $x_1 = N_f(x)$. Supposons que $u = \alpha(f,x) = \|x - x_1\|\gamma(f,x) < 1 - (\sqrt{2}/2)$. Alors,*

- $\beta(f,x_1) \leq \dfrac{1-u}{\psi(u)}\alpha(f,x)\beta(f,x)$,
- $\alpha(f,x_1) \leq \dfrac{\alpha(f,x)^2}{\psi(u)^2}$.

Preuve La première inégalité est donnée par

$$\begin{aligned}\beta(f,x_1) &= \|Df(x_1)^\dagger f(x_1)\| = \|Df(x_1)^\dagger Df(x)Df(x)^\dagger f(x_1)\| \\ &\leq \|Df(x_1)^\dagger Df(x)\|\|Df(x)^\dagger f(x_1)\|.\end{aligned}$$

On majore le premier terme à l'aide du Lemme 123 :

$$\|Df(x_1)^\dagger Df(x)\| \leq \frac{(1-u)^2}{\psi(u)}$$

et le second terme par l'argument suivant

$$f(x_1) = f(x) + Df(x)(x_1 - x) + \sum_{k=2}^{\infty} \frac{D^k f(x)}{k!}(x_1 - x)^k = \sum_{k=2}^{\infty} \frac{D^k f(x)}{k!}(x_1 - x)^k$$

puisque $x_1 = N_f(x)$. On compose à gauche par $Df(x)^\dagger$ puis on majore la norme de cette expression par

$$\begin{aligned}\|Df(x)^\dagger f(x_1)\| &\leq \sum_{k=2}^{\infty} \|Df(x)^\dagger \frac{D^k f(x)}{k!}\|\|x_1 - x\|^k \leq \sum_{k=2}^{\infty} \gamma(f,x)^{k-1}\|x_1 - x\|^k \\ &= \frac{\gamma(f,x)\|x_1 - x\|^2}{1-u} = \frac{\alpha(f,x)\beta(f,x)}{1-u}.\end{aligned}$$

Ceci prouve la première inégalité. La seconde est une conséquence de la première et du Lemme 131. □

Preuve du Théorème 128. Notons $\beta_k = \beta(f,x_k)$, $\gamma_k = \gamma(f,x_k)$ et $u_k = \alpha_k = \alpha(f,x_k)$. Nous allons prouver par récurrence que

$$\alpha_k \leq \left(\frac{1}{2}\right)^{2^k - 1} \alpha(f,x).$$

Pour $k = 0$ il n'y a rien à démontrer puis, par le Lemme 133

$$\begin{aligned}\alpha_{k+1} &\leq \frac{\alpha_k^2}{\psi(u_k)^2} \leq \frac{1}{\psi(u_k)^2}\left(\frac{1}{2}\right)^{2^{k+1}-2} \alpha(f,x)^2 \\ &\leq \frac{2\alpha(f,x)}{\psi(\alpha(f,x))^2}\left(\frac{1}{2}\right)^{2^{k+1}-1} \alpha(f,x).\end{aligned}$$

Par définition de α_0 l'expression $2\alpha(f,x)/\psi(\alpha(f,x))^2$ est inférieure ou égale à 1 d'où

$$\alpha_{k+1} \leq \left(\frac{1}{2}\right)^{2^{k+1}-1} \alpha(f,x).$$

A l'aide de cette inégalité nous allons prouver par récurrence que

$$\|x_{k+1} - x_k\| = \beta_k \leq \left(\frac{1}{2}\right)^{2^k-1} \beta(f,x).$$

Le cas $k = 0$ est trivial puis, par le Lemme 133

$$\begin{aligned}\beta_{k+1} &\leq \frac{1-u_k}{\psi(u_k)}\alpha_k\beta_k \leq \frac{1-u_k}{\psi(u_k)}\left(\frac{1}{2}\right)^{2^k-1}\alpha(f,x)\left(\frac{1}{2}\right)^{2^k-1}\beta(f,x)\\ &\leq \frac{2(1-u_k)\alpha(f,x)}{\psi(u_k)}\left(\frac{1}{2}\right)^{2^{k+1}-1}\beta(f,x)\\ &\leq \frac{2(1-\alpha(f,x))\alpha(f,x)}{\psi(\alpha(f,x))}\left(\frac{1}{2}\right)^{2^{k+1}-1}\beta(f,x).\end{aligned}$$

De par la définition de α_0 on a

$$\frac{2(1-\alpha(f,x))\alpha(f,x)}{\psi(\alpha(f,x))} = \frac{2\alpha(f,x)}{\psi(\alpha(f,x))^2}(1-\alpha(f,x))\psi(\alpha(f,x)) \leq 1$$

de sorte que

$$\beta_{k+1} \leq \left(\frac{1}{2}\right)^{2^{k+1}-1}\beta(f,x)$$

et ceci prouve la première assertion. On en déduit que la suite (x_k) est de Cauchy, donc converge. Notons ζ sa limite. L'inégalité

$$\|x_k - \zeta\| \leq \sum_{p=0}^{\infty} \|x_{k+p+1} - x_{k+p}\|$$

donne la troisième assertion et aussi la surjectivité de $Df(\zeta)$ par le Lemme 123. Enfin, puisque

$$\beta_k = \|Df(x_k)^\dagger f(x_k)\| \to 0$$

lorsque $k \to \infty$ on a $Df(\zeta)^\dagger f(\zeta) = 0$ donc $f(\zeta) = 0$ car $Df(\zeta)$ est surjectif. □

Preuve du Théorème 130. Soit $u = \|x - \zeta\|\gamma(f,\zeta)$. Du Lemme 131 nous déduisons

$$\alpha(f,x) \leq \frac{(1-u)\alpha(f,\zeta)+u}{\psi(u)^2} = \frac{u}{\psi(u)^2} \leq \alpha_0$$

puisque, par hypothèse, $u \leq u_0$. Donc, par le Théorème 128, la suite x_k converge vers un zéro de f que l'on note $M_f(x)$ et l'on a

$$\|x_k - M_f(x)\| \leq 1.63281\ldots\left(\frac{1}{2}\right)^{2^k-1}\beta(f,x).$$

Toujours par le Lemme 131 nous obtenons

$$\beta(f,x) \leq \frac{1-u}{\psi(u)}((1-u)\beta(f,\zeta)+\|x-\zeta\|) = \frac{1-u}{\psi(u)}\|x-\zeta\| \leq \frac{2}{1.63281\ldots}\|x-\zeta\|$$

puisque $u \leq u_0$ et le théorème est démontré. □

4.5 Le cas des espaces euclidiens

Nous supposons, tout au long de cette section, que $\mathbb{E}$ et $\mathbb{F}$ sont des espaces euclidiens de dimension n et m respectivement. $f : \mathbb{E} \to \mathbb{F}$ est une application analytique et $V = f^{-1}(0)$ est l'ensemble de ses zéros. On suppose que, pour tout $\zeta \in V$, $Df(\zeta)$ est surjectif. Ainsi V est une sous-variété de $\mathbb{E}$ de dimension $n-m$. Notre objectif est ici d'étudier les propriétés de l'opérateur M_f défini au paragraphe précédent.

Définition 134. *Nous définissons*

$$\mathcal{T} = \{x \in \mathbb{E} \quad : \quad \exists \zeta \in V \quad d(x,V) = \|x-\zeta\| \quad et \quad \|x-\zeta\|\gamma(f,\zeta) \leq u_0\}.$$

Proposition 135. *$\mathcal{T}$ est un voisinage fermé de V contenu dans $\mathcal{V}$.*

Pour démontrer cette proposition nous utiliserons le résultat suivant :

Proposition 136. *L'application $x \in \mathbb{E} \to \gamma(f,x)$ est continue en tout point où $Df(x)$ est surjective.*

Preuve $\gamma(f,x)$ est semi-continue inférieurement puisqu'elle est l'enveloppe supérieure de la famille de fonctions continues $\gamma_k(f,x) = \|Df(x)^\dagger D^k f(x)/k!\|^{1/(k-1)}$, $k \geq 2$. La semi-continuité supérieure de $\gamma(f,x)$ est une conséquence du Lemme 131. □

Preuve de la Proposition 135. Le fait que $\mathcal{T}$ soit contenu dans $\mathcal{V}$ est immédiat. Pour tout $x \in \mathcal{T}$, $Df(x)$ est surjectif. Cela résulte du Lemme 123 et du fait que $u_0 < 1 - \sqrt{2}/2$. En conséquence $\gamma(f,x)$ est continue sur $\mathcal{T}$. $\mathcal{T}$ est fermé par continuité des fonctions $d(.,V)$ et $\gamma(f,.)$. Montrons enfin que pour tout $\zeta \in V$ il existe $r > 0$ tel que $B(\zeta,r) \subset \mathcal{T}$. Ceci prouvera que $\mathcal{T}$ est un voisinage de V. Prenons $r > 0$ qui vérifie les deux conditions suivantes :

1. $\sup_{\zeta' \in V, \|\zeta'-\zeta\| \leq 2r} \gamma(f,\zeta') \leq \gamma(f,\zeta) + 1,$
2. $r(\gamma(f,\zeta)+1) \leq u_0.$

Un tel r existe puisque $\gamma(f,.)$ est continue. Soit $x \in B(\zeta,r)$ et soit $\zeta_x \in V$ tel que $d(x,V) = \|x-\zeta_x\|$. On a

$$\|x-\zeta_x\| = d(x,V) \leq \|x-\zeta\| < r.$$

D'autre part

$$\|\zeta_x - \zeta\| \leq \|\zeta_x - x\| + \|x - \zeta\| \leq 2r$$

de sorte que $\gamma(f, \zeta_x) \leq \gamma(f, \zeta) + 1$ et donc

$$\|\zeta_x - x\|\gamma(f, \zeta_x) \leq r(\gamma(f, \zeta) + 1) \leq u_0.$$

Ceci prouve que $x \in \mathcal{T}$. □

Théorème 137. *L'application $M_f : \mathcal{T} \to V$ donnée par $M_f(x) = \lim_{k\to\infty} N_f^k(x)$ est définie et continue sur $\mathcal{T}$ et de classe C^1 sur l'intérieur de $\mathcal{T}$. Pour tout $x \in \mathcal{T}$ on a*

$$d(x, V) \leq \|x - M_f(x)\| \leq 2d(x, V),$$

de plus, pour tout $\zeta \in V$, $DM_f(\zeta) = \Pi_{\ker Df(\zeta)}$ la projection orthogonale sur l'espace tangent à V en ζ.

Preuve Le fait que M_f soit définie sur $\mathcal{T}$ et l'encadrement par les distances est une conséquence immédiate du Théorème 130.

On sait, par ce même théorème que, pour tout $k \geq 0$,

$$\|N_f^k(x) - M_f(x)\| \leq 2\left(\frac{1}{2}\right)^{2^k-1} \|x - \zeta\|.$$

Considérons un compact $K \subset \mathcal{T}$. Puisque les applications $d(., V)$ et $\gamma(f, .)$ sont continues, on peut majorer tous les nombres $\|x - \zeta\|$ par une constante $b > 0$ uniformément pour $x \in K$. De ce fait

$$\|N_f^k(x) - M_f(x)\| \leq 2b\left(\frac{1}{2}\right)^{2^k-1}$$

pour tout $x \in K$ et ceci prouve que la suite de fonctions analytiques $(N_f^k(x))$ converge uniformément vers $M_f(x)$ sur K. D'où la continuité de M_f sur $\mathcal{T}$.

Prouver que M_f est de classe C^1 est nettement plus difficile. Il suffit d'établir que, pour tout $x \in \text{int} - \mathcal{T}$, la suite des dérivées $(DN_f^k(y))_{k\geq 1}$ converge uniformément sur une boule fermée $\bar{B}(x, r) \subset \text{int} - \mathcal{T}$ et de prouver que ces dérivées sont continues. La dérivée de $N_f^k(x)$ est donnée par

$$\begin{aligned} DN_f^k(x) &= DN_f(N_f^{k-1}(x)) \circ DN_f(N_f^{k-2}(x)) \circ \ldots \circ DN_f(x) \\ &= \prod_{i=0}^{k-1} DN_f(N_f^i(x)) = \prod_{i=0}^{k-1} DN_f(x_i). \end{aligned}$$

Notons que pour tout $x \in \mathcal{T}$ il existe $\zeta \in V$ tel que $\|x - \zeta\|\gamma(f, \zeta) < 1 - \sqrt{2}/2$ de sorte que, par le Lemme 123 et puisque $Df(\zeta)$ est surjectif, $Df(x)$ est surjectif. Ceci prouve que

$$x \to Df(x)^\dagger = Df(x)^*(Df(x)Df(x)^*)^{-1}$$

est analytique sur $\operatorname{int} -\mathcal{T}$. Notons que

$$DN_f(x) = \mathrm{id} - Df(x)^\dagger Df(x) - D(Df(x)^\dagger)f(x) = \Pi_{\ker Df(x)} - D(Df(x)^\dagger)f(x)$$

où $\Pi_{\ker Df(x)}$ est la projection orthogonale sur $\ker Df(x)$. On en déduit que

$$DN_f(x_i) = \Pi_{\ker Df(M_f(x))} + \Pi_{\ker Df(x_i)} - \Pi_{\ker Df(M_f(x))} - D(Df(x_i)^\dagger)f(x_i).$$

Notons pour simplifier $P(x) = \Pi_{\ker Df(M_f(x))}$. On a $DN_f(x_i) = P(x) + Df(M_f(x))^\dagger Df(M_f(x)) - Df(x_i)^\dagger Df(x_i) - D(Df(x_i)^\dagger)(f(x_i) - f(M_f(x))) = P(x) + R_i(x)$.

Soit $r > 0$ tel que $\bar{B}(x,r) \subset \operatorname{int} -\mathcal{T}$. En utilisant l'analyticité de f, celle de $Df^\dagger$, la continuité de M_f et le Théorème 130 on montre que

$$\|R_i(y)\| \le D\|y_i - M_f(y)\| \le C\left(\frac{1}{2}\right)^{2^i} = \eta_i$$

pour tout $y \in \bar{B}(x,r)$, pour tout $i \ge 0$ et pour des constantes C et D convenables, indépendantes de y et i. En d'autres termes $DN_f(x_i)$ est une petite perturbation de la projection orthogonale $P(x)$. On a $DN_f^k(y) = \prod_{i=0}^{k-1}(P + R_i)(y)$. Nous allons montrer que la suite $(DN_f^k(y))_k$ est de Cauchy. Cela établira sa convergence. On a, pour $k > l$,

$$\begin{aligned}\left\|DN_f^k(y) - DN_f^l(y)\right\| &= \left\|\left(\prod_{i=l}^{k-1}(P+R_i) - P + P - \mathrm{id}\right)\prod_{i=0}^{l-1}(P+R_i)\right\| \\ &\le \left\|\prod_{i=l}^{k-1}(P+R_i) - P\right\|\left\|\prod_{i=0}^{l-1}(P+R_i)\right\| \\ &\quad + \left\|(P-\mathrm{id})R_{l-1}\prod_{i=0}^{l-2}(P+R_i)\right\|.\end{aligned}$$

Nous allons maintenant établir par récurrence que

$$\left\|\prod_{i=l}^{k-1}(P+R_i) - P\right\| \le \sum_{j=l}^{k-1}\eta_j \prod_{i=j+1}^{k-1}(1+\eta_i).$$

Pour $k = l+1$ il n'y a rien à démontrer. Le passage de k à $k+1$ se fait ainsi :

$$\begin{aligned}\left\|\prod_{i=l}^{k}(P+R_i) - P\right\| &= \left\|(P+R_k)\prod_{i=l}^{k-1}(P+R_i) - P + R_kP\right\| \\ &\le (1+\eta_k)\sum_{j=l}^{k-1}\eta_j\prod_{i=j+1}^{k-1}(1+\eta_i) + \eta_k = \sum_{j=l}^{k}\eta_j\prod_{i=j+1}^{k}(1+\eta_i).\end{aligned}$$

Notons aussi que la série $\eta = \sum_{j=0}^{\infty} \eta_j$ est convergente ainsi que le produit infini $\xi = \Pi_{j=0}^{\infty}(1+\eta_j)$. En mettant tout cela ensemble on a

$$\|DN_f^k(y) - DN_f^l(y)\| \leq \xi^2 \sum_{j=l}^{k-1} \eta_j + \xi\eta_{l-1} \leq \xi^2 \sum_{j=l-1}^{k-1} \eta_j$$

ce qui prouve bien que notre suite est de Cauchy.

Le calcul de la dérivée de M_f se déduit de celui de N_f^k puisque $DM_f(x) = \lim_{k\to\infty} DN_f^k(x)$. Pour $k=1$ on a

$$DN_f(x) = \mathrm{id}_{\mathbb{E}} - D(Df(x)^\dagger)f(x) - Df(x)^\dagger Df(x)$$

et pour $x = \zeta \in V$

$$DN_f(\zeta) = \mathrm{id}_E - Df(\zeta)^\dagger Df(\zeta) = id_{\mathbb{E}} - \Pi_{(\ker Df(\zeta))^\perp} = \Pi_{\ker Df(\zeta)}.$$

Par induction et puisque ζ est un point fixe de N_f on a

$$DN_f^{k+1}(\zeta) = D(N_f^k \circ N_f)(\zeta) = DN_f^k(N_f(\zeta)) \circ DN_f(\zeta)$$

$$= DN_f^k(\zeta) \circ DN_f(\zeta) = \Pi_{\ker Df(\zeta)} \circ \Pi_{\ker Df(\zeta)} = \Pi_{\ker Df(\zeta)}$$

de sorte que $DM_f(\zeta) = \lim_{k\to\infty} DN_f^k(\zeta) = \Pi_{\ker Df(\zeta)}$. □

Remarque 5. Il existe une version « classe C^k » de ce théorème : lorsque f est de classe C^{k+1} et que $Df(\zeta)$ est surjective pour tout $\zeta \in V$ alors M_f est définie et de classe C^k (on perd un degré de régularité) dans un voisinage de V. Ces résultats sont dus à Beyn [2]. Dans le cadre analytique que nous avons choisi, M_f est de classe C^∞. On ignore si elle est analytique.

Corollaire 138. *Pour tout $\zeta \in V$ notons $\mathcal{T}_\zeta$ l'ensemble des $x \in int - \mathcal{T}$ tels que $M_f(x) = \zeta$. Il existe un voisinage ouvert $\mathcal{V}$ de V tel que $\mathcal{V} \cap \mathcal{T}_\zeta$ soit une sous-variété analytique de $\mathbb{E}$ de dimension m. Cette sous-variété est invariante par N_f et contient ζ. L'espace tangent en ce point est $T_\zeta\mathcal{T}_\zeta = \ker Df(\zeta)^\perp$. De plus*

$$\bigcup_{\zeta\in V} \mathcal{V} \cap \mathcal{T}_\zeta = \mathcal{V}.$$

Preuve Pour prouver ce corollaire nous utilisons l'exemple 7 de l'appendice : « *Soient $\mathbb{E}$ et $\mathbb{F}$ deux espaces euclidiens, U un ouvert de $\mathbb{E}$ et $F : U \to \mathbb{F}$ une application de classe C^r, $r \geq 1$. Si le rang de $DF(x)$ est constant pour tout $x \in U$ alors, $V = F^{-1}(0)$ est une sous-variété de classe C^r.* »

On va prendre ici $F = M_f - \zeta$ et $V = f^{-1}(0)$. Nous devons vérifier que $DM_f(x)$ est de rang constant. D'une part rang $DM_f(x) \leq n - m$ puisque cette application est à valeurs dans $T_\zeta V = \ker Df(\zeta)$, d'autre part rang $DM_f(\zeta) = n - m$ pour tout $\zeta \in V$ puisque $DM_f(\zeta) = \Pi_{\ker Df(\zeta)}$ dans

ce cas. Comme le rang est semi-continu inférieurement, ceci prouve qu'il est constant et égal $n-m$ dans un voisinage $\mathcal{V}$ de V, donc $DM(x)$ est de rang $n-m$ dans ce voisinage. On obtient ainsi la description de $T_\zeta\mathcal{T}_\zeta$ et le fait que $\mathcal{V}\cap\mathcal{T}_\zeta$ est une sous-variété de $\mathbb{E}$ de dimension m. Le fait que cette sous-variété soit invariante par N_f et qu'elle contienne ζ est évident. Que l'union de ces sous-variétés remplisse $\mathcal{V}$ est une conséquence du fait que M_f est bien défini sur $\mathcal{V}$. □

4.6 Exemple : la fonction d'évaluation

Dans cette section, nous appliquons les résultats établis au cours de ce chapitre à la sous-variété « problèmes-solutions » qui est l'ensemble des zéros de la « fonction d'évaluation » :

$$\text{Eval }(F,x)=F(x)$$

où F est un système et x l'inconnue. Cette fonction a une dérivée partout surjective ce qui fait de

$$V=\text{Eval }^{-1}(0)=\{(F,x)\ :\ F(x)=0\}$$

une sous-variété différentiable. Les résultats que nous avons en vue sont du type suivant : si $F(x)$ est petit, il existe un système G proche de F et un vecteur y proche de x tels que $G(y)=0$.

Qu'entendons nous par « système » ? C'est ici un système polynomial et l'ensemble de ces systèmes est muni d'une structure euclidienne très particulière que nous étudions soigneusement. Nous calculons ensuite l'invariant $\gamma(\text{Eval },F,x)$.

Définition 139. *Nous notons $\mathcal{P}_d$ l'espace des polynômes à coefficients réels et de degré $\leq d$*

$$f(z)=\sum_{|\alpha|\leq d}a_\alpha z^\alpha$$

où $z=(z_1,\ldots,z_n)$, $\alpha=(\alpha_1,\ldots,\alpha_n)$, $|\alpha|=\alpha_1+\ldots+\alpha_n$ et $z^\alpha=z^{\alpha_1}\ldots z^{\alpha_n}$. L'espace $\mathcal{P}_d$ est muni de la structure euclidienne suivante

$$\langle f,g\rangle=\sum_{|\alpha|\leq d}\binom{d}{\alpha}^{-1}a_\alpha b_\alpha$$

avec $g(z)=\sum_{|\alpha|\leq d}b_\alpha z^\alpha$ et où

$$\binom{d}{\alpha}=\frac{d!}{(d-|\alpha|)!\alpha_1!\ldots\alpha_n!}$$

est appelé « coefficient multinomial ».

Remarque 6. Soit $\beta \in \mathbb{N}^{n+1}$, $\beta = (\beta_0, \beta_1, \ldots, \beta_n)$ tel que $|\beta| = \beta_0 + \beta_1 + \ldots + \beta_n = d$. Les coefficients multinomiaux sont égaux à

$$\binom{d}{\beta} = \frac{d!}{\beta_0! \beta_1! \ldots \beta_n!}.$$

Leur propriété essentielle est donnée par le développement suivant

$$(x_0 + x_1 + \ldots + x_n)^d = \sum_{\substack{\beta \in \mathbb{N}^{n+1} \\ |\beta| = d}} \binom{d}{\beta} x^\beta.$$

Dans le contexte de la définition précédente nous prenons $\beta = (d - |\alpha|, \alpha_1, \ldots, \alpha_n)$ de sorte que $\binom{d}{\alpha} = \binom{d}{\beta}$. On obtient alors le développement

$$(1 + z_1 + \ldots + z_n)^d = \sum_{\substack{\alpha \in \mathbb{N}^n \\ |\alpha| \leq d}} \binom{d}{\alpha} z^\alpha.$$

Proposition 140. *Donnons nous $x \in \mathbb{R}^n$ et notons $H(., x)$ le polynôme*

$$H(z, x) = (1 + \langle z, x \rangle)^d \in \mathcal{P}_d.$$

Pour tout $f \in \mathcal{P}_d$ on a

1. $f(x) = \langle f, H(., x) \rangle$,
2. $H(x, x) = \|H(., x)\|^2 = (1 + \|x\|^2)^d$,
3. $|f(x)| \leq \|f\| (1 + \|x\|^2)^{d/2}$.

Preuve La première assertion résulte de la définition du produit scalaire et de la formule du développement multinomial. La seconde s'obtient à partir de la première en prenant $f = H(., x)$, la troisième est donnée par l'inégalité de Cauchy-Schwarz. □

Proposition 141. *Donnons nous un entier $k \geq 1$, $x, u_1, \ldots, u_k \in \mathbb{R}^n$ et notons $H_k(., x, u_1, \ldots, u_k)$ le polynôme*

$$H_k(z, x, u_1, \ldots, u_k) = d \ldots (d - k + 1) \langle z, u_1 \rangle \ldots \langle z, u_k \rangle (1 + \langle z, x \rangle)^{d-k} \in \mathcal{P}_d.$$

Pour tout $f \in \mathcal{P}_d$ on a

1. $D^k f(x)(u_1, \ldots, u_k) = \langle f, H_k(., x, u_1, \ldots, u_k) \rangle$,
2. $\left| D^k f(x)(u_1, \ldots, u_k) \right| \leq d \ldots (d-k+1) \|f\| (1+\|x\|^2)^{(d-k)/2} \|u_1\| \ldots \|u_k\|$,
3. $\left\| D^k f(x) \right\| \leq d \ldots (d - k + 1) \|f\| (1 + \|x\|^2)^{(d-k)/2}$.

Preuve La preuve de la première assertion se fait par récurrence sur k. Le cas $k = 0$ est traité dans la Proposition 140. On a ensuite

$$\begin{aligned} D^{k+1}f(x)(u_1,\dots,u_k,u_{k+1}) &= \frac{d}{dt}D^k f(x+tu_{k+1})(u_1,\dots,u_k)\,|_{t=0} \\ &= \frac{d}{dt}\left\langle f, d\dots(d-k+1)\left\langle .,u_1\right\rangle\dots\left\langle .,u_k\right\rangle(1+\left\langle .,x+tu_{k+1}\right\rangle)^{d-k}\right\rangle|_{t=0} \\ &= \left\langle f, d\dots(d-k+1)(d-k)\left\langle .,u_1\right\rangle\dots\left\langle .,u_k\right\rangle\left\langle .,u_{k+1}\right\rangle(1+\left\langle .,x\right\rangle)^{d-k-1}\right\rangle. \end{aligned}$$

La seconde assertion est beaucoup plus difficile. Cette inégalité résulte de l'inégalité de Cauchy-Schwarz appliquée au produit scalaire

$$D^k f(x)(u_1,\dots,u_k) = \langle f, H_k(.,x,u_1,\dots,u_k)\rangle$$

qui donne

$$\left|D^k f(x)(u_1,\dots,u_k)\right| \le \|f\|\,\|H_k(.,x,u_1,\dots,u_k)\|.$$

Il faut donc estimer la norme de H_k. Nous allons y arriver à l'aide d'une formule intégrale pour le produit scalaire $\langle f,g\rangle$. Commençons par associer à $f(z) = \sum_{|\alpha|\le d} a_\alpha z^\alpha$ le polynôme homogène

$$f_h(z_0,z) = \sum_{|\alpha|\le d} a_\alpha z_0^{d-|\alpha|} z^\alpha.$$

On procède de même avec $g(z) = \sum_{|\beta|\le d} b_\beta z^\beta$. Notons enfin $\mathbb{S}^{2n+1}$ la sphère unité dans $\mathbb{C}^{n+1}$ c'est à dire

$$\mathbb{S}^{2n+1} = \Big\{(z_0,z_1,\dots,z_n)\in\mathbb{C}^{n+1}\ :\ |z_0|^2+|z_1|^2+\dots+|z_n|^2 = 1\Big\}.$$

Elle est équipée de l'unique mesure unitairement invariante, notée $d\mathbb{S}^{2n+1}$, pour laquelle la mesure totale de la sphère est égale à

$$\int_{\mathbb{S}^{2n+1}} d\mathbb{S}^{2n+1} = \frac{2\pi^{n+1}}{n!}.$$

Pour cette mesure, par un calcul que nous ne détaillerons pas ici, on a

$$\int_{\mathbb{S}^{2n+1}} z_0^{d-|\alpha|} z^\alpha \overline{z_0^{d-|\beta|} z^\beta} d\mathbb{S}^{2n+1} = \begin{cases} 0 & \text{si } \alpha\neq\beta, \\ 2\pi^{n+1}\frac{d!}{(d+n)!}\binom{d}{\alpha}^{-1} & \text{sinon.} \end{cases}$$

A partir de ces intégrales on obtient la description suivante de la structure euclidienne de $\mathcal{P}_d$:

$$\begin{aligned}\int_{\mathbb{S}^{2n+1}} f_h(z_0,z)\overline{g_h(z_0,z)}d\mathbb{S}^{2n+1} &= \sum_{|\alpha|\le d,\ |\beta|\le d} a_\alpha b_\beta \int_{\mathbb{S}^{2n+1}} z_0^{d-|\alpha|} z^\alpha \overline{z_0^{d-|\beta|} z^\beta} d\mathbb{S}^{2n+1} \\ &= \sum_{|\alpha|\le d} a_\alpha b_\alpha 2\pi^{n+1}\frac{d!}{(d+n)!}\binom{d}{\alpha}^{-1} = 2\pi^{n+1}\frac{d!}{(d+n)!}\langle f,g\rangle .\end{aligned}$$

Ce calcul va nous permettre d'arriver à nos fins :

$$\|H_k\|^2 = \langle H_k, H_k\rangle = \frac{(d+n)!}{2\pi^{n+1}d!}\int_{\mathbb{S}^{2n+1}} |H_{k,h}(z_0,z)|^2\, d\mathbb{S}^{2n+1}.$$

Comme

$$H_{k,h}(z_0,z) = d\dots(d-k+1)\,\langle z,u_1\rangle\dots\langle z,u_k\rangle\,(z_0+\langle z,x\rangle)^{d-k}$$

et que $(z_0,z)\in\mathbb{S}^{2n+1}$ on obtient la majoration

$$|H_{k,h}(z_0,z)| \le d\dots(d-k+1)\,\|u_1\|\dots\|u_k\|\,(1+\|x\|^2)^{(d-k)/2}.$$

Réunissons nos victuailles :

$$\|H_k\|^2 \le \frac{(d+n)!}{2\pi^{n+1}d!}\int_{\mathbb{S}^{2n+1}} d^2\dots(d-k+1)^2\,\|u_1\|^2\dots\|u_k\|^2\,(1+\|x\|^2)^{d-k}d\mathbb{S}^{2n+1}.$$

La fonction à intégrer ne dépendant plus de (z_0,z), l'intégrale est donnée par la mesure de la sphère

$$\begin{aligned}\|H_k\|^2 &\le \frac{(d+n)!}{2\pi^{n+1}d!}d^2\dots(d-k+1)^2\,\|u_1\|^2\dots\|u_k\|^2\,(1+\|x\|^2)^{d-k}\frac{2\pi^{n+1}}{n!} \\ &= \frac{(d+n)!}{n!d!}d^2\dots(d-k+1)^2\,\|u_1\|^2\dots\|u_k\|^2\,(1+\|x\|^2)^{d-k}\end{aligned}$$

de sorte que

$$\|H_k\| \le d\dots(d-k+1)\,\|u_1\|\dots\|u_k\|\,(1+\|x\|^2)^{(d-k)/2}.$$

La troisième assertion est une conséquence de la seconde. □

Définition 142. *Soit $(d) = (d_1,\dots,d_n)$ un vecteur d'entiers ≥ 1. Nous notons*

$$\mathcal{P}_{(d)} = \mathcal{P}_{d_1}\times\dots\times\mathcal{P}_{d_n}$$

l'espace des systèmes polynomiaux $F=(f_1,\dots,f_n)$ avec $f_i\in\mathcal{P}_{d_i}$. Cet espace est muni de la structure euclidienne produit :

$$\langle F,G\rangle = \sum_{i=1}^n \langle f_i,g_i\rangle_{\mathcal{P}_{d_i}}$$

avec $G=(g_1,\dots,g_n)$.

En réunissant les Propositions 140 et 141 on obtient facilement le résultat suivant :

Proposition 143. *Pour tout entier $k \geq 1$, $x \in \mathbb{R}^n$ et le système $F \in \mathcal{P}_{(d)}$ on a :*

1. $\|F(x)\| \leq \|F\| \left(1 + \|x\|^2\right)^{D/2}$,
2. $\left\|D^k F(x)\right\| \leq D \dots (D-k+1) \|F\| \left(1 + \|x\|^2\right)^{(D-k)/2}$

où $D = \max\{d_i \ : \ 1 \leq i \leq n\}$ est le degré du système.

Considérons maintenant la fonction d'évaluation

$$\text{Eval } : \mathcal{P}_{(d)} \times \mathbb{R}^n \to \mathbb{R}^n, \ \text{Eval } (F, x) = F(x).$$

Nous lui associons l'ensemble suivant :

$$\Sigma = \left\{(F, x) \in \mathcal{P}_{(d)} \times \mathbb{R}^n \ : \ F(x) = 0\right\}.$$

Proposition 144. *Σ est une sous-variété différentiable dans l'espace produit $\mathcal{P}_{(d)} \times \mathbb{R}^n$.*

Preuve Il suffit de montrer que $D\text{Eval } (F, x)$ est surjective pour tout $(F, x) \in \Sigma$ et d'utiliser l'exemple 3 de l'appendice. On a

$$D\text{Eval } (F, x)(\dot{F}, \dot{x}) = \dot{F}(x) + DF(x)\dot{x}.$$

Cette dérivée est surjective, $\dot{F}$ suffit. □

Nous allons prouver le résultat suivant où les constantes α_0 et $1.63281\dots$ qui y figurent sont celles du Théorème 128 :

Théorème 145. *Soient $F \in \mathcal{P}_{(d)}$ et $x \in \mathbb{R}^n$ qui vérifient*

$$\|F(x)\| \leq \frac{\alpha_0}{\gamma(\mathit{Eval}\ , F, x)}.$$

Il existe $G \in \mathcal{P}_{(d)}$ et $y \in \mathbb{R}^n$ tels que $G(y) = 0$ et

$$\left(\|F - G\|^2 + \|x - y\|^2\right)^{1/2} \leq 1.63281\dots \|F(x)\|.$$

De plus, la quantité $\gamma(\mathit{Eval}\ , F, x)$ vérifie

$$\gamma(\mathit{Eval}\ , F, x) \leq \binom{D+1}{2} \left(1 + \|F\|^2\right)^{1/2} \left(1 + \|x\|^2\right)^{\frac{D-1}{2}}.$$

Ce théorème est une conséquence des trois lemmes suivants :

Lemme 146. *Soient A, B et E des matrices $n \times n$, réelles et symétriques, telles que $B = A+E$ et que E soit définie positive (resp. semi-définie positive). Notons $\lambda_1 \geq \ldots \geq \lambda_n$ les valeurs propres de A et $\mu_1 \geq \ldots \geq \mu_n$ celles de B. Alors $\lambda_i < \mu_i$ pour tout i (resp. $\lambda_i \leq \mu_i$).*

Preuve Rappelons que, puisque A est réelle est symétrique, ses valeurs propres sont réelles et qu'il existe une base orthonormée de $\mathbb{R}^n$ faite de vecteurs propres de A. Pour prouver le lemme, nous utilisons la description suivante des valeurs propres de A :

$$\lambda_i = \max_{\dim \mathcal{X}=i} \min_{x\in\mathcal{X},\ \|x\|=1} x^T A x$$

où $\mathcal{X}$ est un sous-espace de dimension i de $\mathbb{R}^n$. Cette formule s'obtient elle-même en deux étapes. Notons $x_1, \ldots, x_n$ une base orthonormée de $\mathbb{R}^n$ où x_k est un vecteur propre de A associé à λ_k. Soit $\mathcal{X}_i$ le sous-espace engendré par $x_1, \ldots, x_i$. On a

$$\lambda_i = \min_{x\in\mathcal{X}_i,\ \|x\|=1} x^T A x$$

ce qui prouve que

$$\lambda_i \leq \max_{\dim \mathcal{X}=i} \min_{x} x^T A x.$$

Pour obtenir l'autre inégalité, il faut montrer que, pour tout sous-espace $\mathcal{X}$ de dimension i, il existe $x \in \mathcal{X}$ de norme 1 tel que

$$\lambda_i \geq x^T A x.$$

On obtient un tel x dans l'intersection de $\mathcal{X}$ et du sous-espace engendré par $x_i, \ldots, x_n$. Ces deux sous-espaces n'ont pas une intersection réduite à $\{0\}$ parce que leurs dimensions respectives sont i et $n - i + 1$.

Revenons à l'inégalité $\lambda_i < \mu_i$. Lorsque E est définie positive, pour tout $x \neq 0$, on a :

$$x^T A x < x^T A x + x^T E x = x^T B x$$

de sorte que

$$\lambda_i = \max_{\dim \mathcal{X}=i} \min_{x\in\mathcal{X},\ \|x\|=1} x^T A x < \max_{\dim \mathcal{X}=i} \min_{x\in\mathcal{X},\ \|x\|=1} x^T B x = \mu_i.$$

Lorsque E est semi-définie positive, on obtient par le même argument, une inégalité large. □

Lemme 147. *Considérons les matrices $n \times n$ suivantes :*

$$\mathcal{D} = \mathrm{Diag}\left((1 + \|x\|^2)^{d_i}\right), \quad \mathcal{E} = \mathcal{D} + DF(x)DF(x)^*.$$

L'inverse généralisé de $DEval\,(F,x)$ *est donné par*

$$DEval\,(F,x)^{\dagger}\mu = (\lambda_1 H_1(.,x),\ldots,\lambda_n H_n(.,x), DF(x)^*\lambda)$$

où $H_i(z,x) = (1+\langle z,x\rangle)^{d_i} \in \mathcal{P}_{d_i}$ *a été introduit Proposition 140 et où* $\lambda = \begin{pmatrix}\lambda_1\\ \vdots\\ \lambda_n\end{pmatrix}$, *vérifie* $\mathcal{E}\lambda = \mu$. *De plus*

$$\left\|DEval\,(F,x)^{\dagger}\right\| \leq 1.$$

Preuve Commençons par calculer l'inverse généralisé de $D\text{Eval}\,(F,x)$. Pour tout $\mu \in \mathbb{R}^n$, $D\text{Eval}\,(F,x)^{\dagger}\mu$ est l'unique inverse de μ dans $(\ker D\text{Eval}\,(F,x))^{\perp}$. Calculons cet orthogonal. Tout d'abord, $(\dot{F},\dot{x}) \in \ker D\text{Eval}\,(F,x)$ si et seulement si $\dot{F}(x) + DF(x)\dot{x} = 0$ c'est à dire, pour tout i, en vertu de la Proposition 140,

$$\left\langle \dot{f}_i, H_i(.,x)\right\rangle + \langle \dot{x}, Df_i(x)^*\rangle = 0$$

où l'on note $F = (f_1,\ldots,f_n)$, $Df_i(x) = \left(\frac{\partial f_i}{\partial x_1}(x),\ldots,\frac{\partial f_i}{\partial x_n}(x)\right)$, $x = \begin{pmatrix}x_1\\ \vdots\\ x_n\end{pmatrix}$ et $\dot{F} = (\dot{f}_1,\ldots,\dot{f}_n)$. En termes du produit scalaire produit de $\mathcal{P}_{(d)} \times \mathbb{R}^n$ on obtient

$$\left\langle \dot{F}, (0,\ldots,H_i(.,x),\ldots,0)\right\rangle + \langle \dot{x}, Df_i(x)^*\rangle = 0$$

ce qui prouve que $(\ker D\text{Eval}\,(F,x))^{\perp}$ est le sous-espace engendré par les

$$(0,\ldots,H_i(.,x),\ldots,0,Df_i(x)^*)\,,\ 1 \leq i \leq n,$$

c'est à dire l'ensemble des

$$(\lambda_1 H_1(.,x),\ldots,\lambda_n H_n(.,x), Df(x)^*\lambda),$$

$\lambda \in \mathbb{R}^n$. La condition

$$D\text{Eval}\,(F,x)\,(\lambda_1 H_1(.,x),\ldots,\lambda_n H_n(.,x), DF(x)^*\lambda) = \mu$$

devient

$$\begin{pmatrix}\lambda_1 H_1(x,x)\\ \vdots\\ \lambda_1 H_n(x,x)\end{pmatrix} + DF(x)DF(x)^*\lambda = \mathcal{E}\lambda = \mu.$$

La matrice $\mathcal{E} = \mathcal{D} + DF(x)DF(x)^*$ est la somme de la matrice diagonale $\mathcal{D}$ et de la matrice semi-définie positive $DF(x)DF(x)^*$. Le Lemme 146 prouve que les valeurs propres de $\mathcal{E}$ sont plus grandes que celles de $\mathcal{D}$, elle mêmes égales

à $H_i(x,x) = (1+\|x\|)^{d_i} \geq 1$. L'inverse de $\mathcal{E}$ est lui aussi symétrique et ses valeurs propres, les inverses des valeurs propres de $\mathcal{E}$, sont positives et ≤ 1. La norme de $\mathcal{E}^{-1}$ qui, pour une matrice symétrique réelle, est le plus grand des modules des valeurs propres, satisfait donc

$$\left\|\mathcal{E}^{-1}\right\| \leq 1.$$

Calculons maintenant la norme de $D\text{Eval}\,(F,x)^\dagger$. Soient λ et $\mu \in \mathbb{R}^n$ avec $\mathcal{E}\lambda = \mu$. On a

$$\left\|D\text{Eval}\,(F,x)^\dagger \mu\right\|^2 = \lambda_1^2 \left\|H_1(.,x)\right\|^2 + \ldots + \lambda_n^2 \left\|H_n(.,x)\right\|^2 + \left\|DF(x)^*\lambda\right\|^2$$

ce qui, par la Proposition 140, est égal à

$$\lambda^T \mathcal{D}\lambda + \lambda^T DF(x)DF(x)^*\lambda = \lambda^T \mathcal{E}\lambda = \mu^T \mathcal{E}^{-1}\mu \leq \left\|\mathcal{E}^{-1}\right\| \|\mu\|^2 \leq \|\mu\|^2$$

de sorte que

$$\left\|D\text{Eval}\,(F,x)^\dagger\right\| \leq 1. \;\square$$

Lemme 148. *Avec les notations du Théorème 145 on a :*

$$\gamma(\mathit{Eval}\,, F, x) \leq \binom{D+1}{2} \left(1+\|F\|^2\right)^{1/2} \left(1+\|x\|^2\right)^{\frac{D-1}{2}}.$$

Preuve La première étape consiste à calculer $D^k\text{Eval}$. Pour tous $(\dot{F}_i, \dot{x}_i) \in \mathcal{P}_{(d)} \times \mathbb{R}^n$, $1 \leq i \leq k$, on a

$$D^k\text{Eval}\,(F,x)\left(\dot{F}_1, \dot{x}_1, \ldots, \dot{F}_k, \dot{x}_k\right)$$

$$= D^k F(x)(\dot{x}_1, \ldots, \dot{x}_k) + \sum_{j=1}^{k} D^{k-1}\dot{F}_j(x)\left(\dot{x}_1, \ldots, \overline{\dot{x}_j}, \ldots, \dot{x}_k\right)$$

où l'expression $\overline{\dot{x}_j}$ exprime que le terme $\dot{x}_j$ est manquant. Passons aux normes :

$$\left\|D^k\text{Eval}\,(F,x)\left(\dot{F}_1, \dot{x}_1, \ldots, \dot{F}_k, \dot{x}_k\right)\right\|$$

$$\leq \left\|D^k F(x)\left(\dot{x}_1, \ldots, \dot{x}_k\right)\right\| + \sum_{j=1}^{k} \left\|D^{k-1}\dot{F}_j(x)\left(\dot{x}_1, \ldots, \overline{\dot{x}_j}, \ldots, \dot{x}_k\right)\right\|$$

que l'on majore à l'aide de la Proposition 143, ce qui donne :

$$\leq D\ldots(D-k+1)\,\|F\|\left(1+\|x\|^2\right)^{\frac{D-k}{2}} \|\dot{x}_1\|\ldots\|\dot{x}_k\|$$

$$+\sum_{j=1}^{k} D\ldots(D-k+2)\left\|\dot{F}_j\right\|\left(1+\|x\|^2\right)^{\frac{D-k+1}{2}} \|\dot{x}_1\|\ldots\overline{\|\dot{x}_j\|}\ldots\|\dot{x}_k\|.$$

On majore dans ces expressions $\|\dot{x}_j\|$ et $\left\|\dot{F}_j\right\|$ par $\left\|(\dot{F}_j, \dot{x}_j)\right\|$ pour obtenir

$$\left\|D^k\text{Eval }(F,x)\right\|$$

$$\leq D\dots(D-k+1)\,\|F\|\left(1+\|x\|^2\right)^{\frac{D-k}{2}} + D\dots(D-k+2)k\left(1+\|x\|^2\right)^{\frac{D-k+1}{2}}.$$

Passons maintenant à

$$\gamma(\text{Eval }, F, x) = \sup_{k\geq 2}\left\|D\text{Eval }(F,x)^\dagger \frac{D^k\text{Eval }(F,x)}{k!}\right\|^{\frac{1}{k-1}}.$$

A l'aide des deux lemmes précédents, on majore $\left\|D\text{Eval }(F,x)^\dagger\right\|$ par 1 et la dérivée $k-$ième comme ci-dessus pour obtenir

$$\begin{aligned}
&\left\|D\text{Eval }(F,x)^\dagger \frac{D^k\text{Eval }(F,x)}{k!}\right\|^{\frac{1}{k-1}} \\
&\quad\leq \left(\binom{D}{k}\|F\|\left(1+\|x\|^2\right)^{\frac{D-k}{2}} + \binom{D}{k-1}\left(1+\|x\|^2\right)^{\frac{D-k+1}{2}}\right)^{\frac{1}{k-1}} \\
&\quad\leq \left(\binom{D+1}{k}(1+\|F\|^2)^{1/2}\left(1+\|x\|^2\right)^{\frac{D-k+1}{2}}\right)^{\frac{1}{k-1}}.
\end{aligned}$$

Il nous reste à montrer que le supremum, pour $k \geq 2$, de cette expression est obtenu pour $k = 2$. Il est facile de voir que

$$(1+\|F\|^2)^{\frac{1}{2(k-1)}} \leq (1+\|F\|^2)^{\frac{1}{2}}$$

et que

$$\left(1+\|x\|^2\right)^{\frac{D-k+1}{2(k-1)}} \leq \left(1+\|x\|^2\right)^{\frac{D-1}{2}}.$$

Pour montrer que

$$\binom{D+1}{k}^{\frac{1}{k-1}} \leq \binom{D+1}{2}$$

on prouve, par un calcul élémentaire et embêtant, que cette suite est décroissante, d'où le lemme. □

Preuve du Théorème 145. Ce théorème est une conséquence du Théorème 128. L'hypothèse $\alpha(\text{Eval }, F, x) \leq \alpha_0$ du Théorème 128 est satisfaite dès que

$$\|F(x)\| \leq \frac{\alpha_0}{\gamma(\text{Eval }, F, x)}$$

qui est l'hypothèse du Théorème 145. En effet,

$$\begin{aligned}\alpha(\text{Eval }, F, x) &= \beta(\text{Eval }, F, x)\gamma(\text{Eval }, F, x)\\ &= \left\|D\text{Eval }(F,x)^{\dagger}\text{Eval }(F,x)\right\|\gamma(\text{Eval }, F, x)\\ &\leq \|F(x)\|\,\gamma(\text{Eval }, F, x)\end{aligned}$$

en vertu du Lemme 147. □

Remarque 7. Le Théorème 128, que nous avons utilisé ici, montre aussi que, sous les hypothèses du Théorème 145, la méthode de Newton appliquée à la fonction Eval et initialisée en (F, x) converge quadratiquement vers le couple (G, y) du Théorème 145. On a ici

$$N_{\text{Eval}}\,(F,x) = (F,x) - D\text{Eval }(F,x)^{\dagger}\text{Eval }(F,x) = (F,x) - D\text{Eval }(F,x)^{\dagger}F(x)$$

et, comme nous l'avons vu au Lemme 147,

$$D\text{Eval }(F,x)^{\dagger}F(x) = (\lambda_1 H_1(.,x), \ldots, \lambda_n H_n(.,x), DF(x)^{*}\lambda)$$

et

$$\lambda = \left(\text{Diag}\left((1+\|x\|^2)^{d_i}\right) + DF(x)DF(x)^{*}\right)^{-1} F(x).$$

4.7 Exemple : le problème symétrique des valeurs propres

Notons $\mathcal{S}_n$ l'espace des matrices $n \times n$ réelles et symétriques. Si A est une telle matrice, ses valeurs propres sont réelles et elle possède une base orthonormée faite de vecteurs propres autrement dit, il existe une matrice diagonale réelle D et une matrice orthogonale U telles que

$$U^T A U = D.$$

C'est le fameux théorème spectral. Ceci nous conduit à considérer l'ensemble

$$\mathcal{V} = \{(A, x, \lambda) \in \mathcal{S}_n \times \mathbb{R}^n \times \mathbb{R} \ : \ F(A, x, \lambda) = 0\}$$

où F est définie par

$$F(A, x, \lambda) = \begin{pmatrix} (\lambda I - A)x \\ \frac{1}{2}\left(\|x\|^2 - 1\right) \end{pmatrix} \in \mathbb{R}^n \times \mathbb{R}.$$

L'ensemble $\mathcal{V}$ est donc constitué de triplets (A, x, λ) où λ est une valeur propre de A et x un vecteur propre normalisé associé à λ. Le coefficient $1/2$ qui apparaît dans la définition de F est présent pour des raisons cosmétiques : il

va disparaître au cours d'une dérivation. La structure de $\mathcal{V}$ est décrite par la proposition suivante :

Proposition 149. *$\mathcal{V}$ est une sous-variété analytique de $\mathcal{S}_n \times \mathbb{R}^n \times \mathbb{R}$ de dimension $n(n+1)/2$.*

Preuve F est une fonction polynomiale (donc analytique) et sa dérivée est

$$DF(A,x,\lambda)(\dot{A},\dot{x},\dot{\lambda}) = \begin{pmatrix} (\lambda I - A)\dot{x} + (\dot{\lambda} I - \dot{A})x \\ x^T \dot{x} \end{pmatrix}.$$

Pour tout triplet $(A,x,\lambda) \in \mathcal{V}$ cette dérivée est surjective : étant donnés $y \in \mathbb{R}^n$ et $\mu \in \mathbb{R}$, on a

$$DF(A,x,\lambda)(\dot{A},\dot{x},\dot{\lambda}) = \begin{pmatrix} y \\ \mu \end{pmatrix}$$

dès que $\dot{A} = -yx^T$, $\dot{x} = \mu x$ et $\dot{\lambda} = 0$. On conclut à l'aide de l'Exemple 3. □

Le théorème que nous avons en vue relie la quantité $\|F(A,x,\lambda)\|$ à la distance de (A,x,λ) à la sous-variété $\mathcal{V}$, quelque chose comme «si $\|F(A,x,\lambda)\|$ est petit, on est proche de $\mathcal{V}$ ». La distance que nous considérons ici est la distance euclidienne sur $\mathcal{S}_n \times \mathbb{R}^n \times \mathbb{R}$ associée au produit scalaire

$$\langle (A,x,\lambda),(A',x',\lambda') \rangle = \langle A, A' \rangle_F + \langle x, x' \rangle + \lambda\lambda'$$

avec

$$\langle A, A' \rangle_F = \text{trace } (A^T A') = \sum_{i,j=1}^{n} a_{ij} {a'}_{ij}.$$

Théorème 150. *Soit $(A,x,\lambda) \in \mathcal{S}_n \times \mathbb{R}^n \times \mathbb{R}$ qui vérifie*

$$\|(\lambda I - A)x\| \leq \frac{\alpha_0}{2\sqrt{2}},$$
$$\|x\| = 1.$$

Il existe un triplet $(B,y,\mu) \in \mathcal{V}$ qui vérifie

$$\|B - A\|_F^2 + \|y - x\|^2 + |\mu - \lambda|^2 \leq 5.4\, \|(\lambda I - A)x\|^2.$$

De plus, la suite de Newton $(A_{p+1}, x_{p+1}, \lambda_{p+1}) = N_F(A_p, x_p, \lambda_p)$, où $(A_0, x_0, \lambda_0) = (A,x,\lambda)$, converge quadratiquement vers (B,y,μ).

Remarque 8. La constante α_0 qui figure dans l'énoncé ci-dessus ainsi que l'expression « converge quadratiquement » sont à prendre au sens du Théorème 128.

Au cours des démonstrations des lemmes qui suivent et qui vont nous conduire au Théorème 150, l'invariance orthogonale va jouer un rôle très important. Précisons cela :

Lemme 151. *Pour toute matrice $n \times n$ orthogonale U notons*

$$\mathcal{U}(A, x, \lambda) = (U^T A U, U^T x, \lambda)$$

ainsi que

$$\mathcal{U} \otimes \mathcal{U}((A, x, \lambda), (A', x', \lambda')) = (\mathcal{U}(A, x, \lambda), \mathcal{U}(A', x', \lambda')).$$

Pour tout $(A, x, \lambda) \in \mathcal{S}_n \times \mathbb{R}^n \times \mathbb{R}$ on a

1. $\mathcal{U}(A, x, \lambda) \in \mathcal{V}$ *si et seulement si* $(A, x, \lambda) \in \mathcal{V}$,
2. $F(A, x, \lambda) = \begin{pmatrix} U & 0 \\ 0 & 1 \end{pmatrix} F \circ \mathcal{U}(A, x, \lambda)$,
3. $DF(A, x, \lambda) = \begin{pmatrix} U & 0 \\ 0 & 1 \end{pmatrix} DF\left(\mathcal{U}(A, x, \lambda)\right) \circ \mathcal{U}$,
4. $DF(A, x, \lambda)^\dagger = \mathcal{U}^{-1} \circ DF\left(\mathcal{U}(A, x, \lambda)\right)^\dagger \begin{pmatrix} U^T & 0 \\ 0 & 1 \end{pmatrix}$,
5. $D^2F(A, x, \lambda) = \begin{pmatrix} U & 0 \\ 0 & 1 \end{pmatrix} D^2F\left(\mathcal{U}(A, x, \lambda)\right) \circ (\mathcal{U} \otimes \mathcal{U})$,
6. $\gamma(F, A, x, \lambda) = \gamma(F, U^T A U, U^T x, \lambda)$.

Preuve Les deux premières assertions sont évidentes et la troisième résulte du théorème de dérivation des fonctions composées. Pour prouver la quatrième on utilise l'assertion 3, le Théorème 120-5 et le fait que $\mathcal{U}$ est une transformation orthogonale. La cinquième assertion est encore due au théorème de dérivation des fonctions composées. La sixième provient des quatrièmes et cinquièmes : puisque F est polynomiale de degré 2 on a

$$\begin{aligned}
\gamma(F, A, x, \lambda) &= \frac{1}{2} \left\| DF(A, x, \lambda)^\dagger D^2F\left(A, x, \lambda\right) \right\| \\
&= \frac{1}{2} \left\| \mathcal{U}^{-1} \circ DF\left(\mathcal{U}(A, x, \lambda)\right)^\dagger \begin{pmatrix} U^T & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} U & 0 \\ 0 & 1 \end{pmatrix} \right. \\
&\quad \left. \times\, D^2F\left(\mathcal{U}(A, x, \lambda)\right) \circ (\mathcal{U} \otimes \mathcal{U}) \right\| \\
&= \frac{1}{2} \left\| \mathcal{U}^{-1} \circ DF\left(\mathcal{U}(A, x, \lambda)\right)^\dagger D^2F\left(\mathcal{U}(A, x, \lambda)\right) \circ (\mathcal{U} \otimes \mathcal{U}) \right\|.
\end{aligned}$$

Comme $\mathcal{U}^{-1}$ et $\mathcal{U} \otimes \mathcal{U}$ sont des transformations orthogonales l'expression ci-dessus est égale à

$$\frac{1}{2} \left\| DF\left(\mathcal{U}(A, x, \lambda)\right)^\dagger D^2F\left(\mathcal{U}(A, x, \lambda)\right) \right\| = \gamma(F, U^T A U, U^T x, \lambda). \quad \square$$

Remarque 9. L'essentiel de la démonstration du Théorème 150 consiste à calculer $\gamma(F, A, x, \lambda)$. Puisque $A \in \mathcal{S}_n$ est orthogonalement semblable à une matrice diagonale ($U^T AU = D$ avec D diagonale et U orthogonale), en vertu du lemme précédent, pour calculer $\gamma(F, A, x, \lambda)$ il suffit de considérer le cas $A = \mathrm{Diag}(\lambda_1, \ldots, \lambda_n)$, $x = e_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$ et $\lambda = \lambda_1$ ce que nous ferons désormais.

Lemme 152. $\ker DF(A, e_1, \lambda_1)$ *est constitué des triplets* $(\dot{A}, \dot{x}, \dot{\lambda}) \in \mathcal{S}_n \times \mathbb{R}^n \times \mathbb{R}$ *tels que :*

1. $\dot{a}_{11} = \dot{\lambda}$,
2. $\dot{a}_{i1} = \dot{a}_{1i} = (\lambda_1 - \lambda_i)\dot{x}_i$, $2 \leq i \leq n$,
3. $\dot{x}_1 = 0$.

Preuve Il suffit d'écrire que $DF(A, e_1, \lambda_1)(\dot{A}, \dot{x}, \dot{\lambda}) = 0$. □

Lemme 153. $(\ker DF(A, e_1, \lambda_1))^{\perp}$ *est constitué des triplets* $(B, y, \mu) \in \mathcal{S}_n \times \mathbb{R}^n \times \mathbb{R}$ *tels que :*

1. $b_{11} + \mu = 0$,
2. $2(\lambda_1 - \lambda_i)b_{i1} + y_i = 0$, $2 \leq i \leq n$,
3. $b_{ij} = 0$, $2 \leq i, j \leq n$.

Preuve En vertu du lemme précédent, la relation d'orthogonalité

$$\left\langle (B, y, \mu), (\dot{A}, \dot{x}, \dot{\lambda}) \right\rangle = 0$$

pour tout $(\dot{A}, \dot{x}, \dot{\lambda}) \in \ker DF(A, e_1, \lambda_1)$ devient

$$b_{11}\dot{\lambda} + 2\sum_{i=2}^{n} b_{i1}(\lambda_1 - \lambda_i)\dot{x}_i + \sum_{i,j=2}^{n} b_{ij}\dot{a}_{ij} + \sum_{i=2}^{n} y_i\dot{x}_i + \mu\dot{\lambda} = 0$$

pour tout $\dot{a}_{ij}$, $2 \leq i, j \leq n$, $\dot{x}_i$, $2 \leq i \leq n$, et $\dot{\lambda}$. On obtient ainsi les égalités annoncées dans le lemme. □

Lemme 154. *Pour tout* $(z, \nu) \in \mathbb{R}^n \times \mathbb{R}$ *on a* $DF(A, e_1, \lambda_1)^{\dagger}(z, \nu) = (B, y, \mu)$ *avec*

1. $b_{11} = -\frac{z_1}{2}$,
2. $b_{i1} = b_{1i} = -\frac{z_i}{1+2(\lambda_1-\lambda_i)^2}$, $2 \leq i \leq n$,
3. $b_{ij} = 0$, $2 \leq i, j \leq n$,
4. $y_1 = \nu$,
5. $y_i = \frac{2(\lambda_1-\lambda_i)z_i}{1+2(\lambda_1-\lambda_i)^2}$, $2 \leq i \leq n$,
6. $\mu = \frac{z_1}{2}$.

De plus $\left\|DF(A, e_1, \lambda_1)^{\dagger}\right\| \leq \sqrt{2}$.

Preuve (B, y, μ) est caractérisé par

$$(B, y, \mu) \in (\ker DF(A, e_1, \lambda_1))^{\perp}$$

et

$$DF(A, e_1, \lambda_1)(B, y, \mu) = (z, \nu).$$

Cette appartenance et cette équation s'expriment en le système suivant :

$$\begin{array}{l}
\mu + b_{11} = 0 \\
y_i + 2(\lambda_1 - \lambda_i) b_{i1} = 0,\ 2 \leq i \leq n, \\
b_{ij} = 0,\ 2 \leq i, j \leq n, \\
\mu - b_{11} = z_1, \\
(\lambda_1 - \lambda_i) y_i - b_{i1} = z_i,\ 2 \leq i \leq n, \\
y_1 = \nu,
\end{array}$$

qui donne les identités décrites dans le lemme. De plus,

$$\begin{aligned}
\left\| DF(A, e_1, \lambda_1)^{\dagger}(z, \nu) \right\|^2 &= \sum b_{ij}^2 + \sum y_i^2 + \mu^2 \\
&= \frac{z_1^2}{4} + 2 \sum_{i=2}^{n} \frac{z_i^2}{(1 + 2(\lambda_1 - \lambda_i)^2)^2} + \nu^2 + \sum_{i=2}^{n} \frac{4(\lambda_1 - \lambda_i)^2 z_i^2}{(1 + 2(\lambda_1 - \lambda_i)^2)^2} + \frac{z_1^2}{4} \\
&= \frac{z_1^2}{2} + 2 \sum_{i=2}^{n} \frac{z_i^2}{1 + 2(\lambda_1 - \lambda_i)^2} + \nu^2 \\
&\leq \|(z, \nu)\|^2 \max \left\{ 1, \frac{2}{1 + 2(\lambda_1 - \lambda_i)^2} \right\} \leq 2 \|(z, \nu)\|^2
\end{aligned}$$

et ainsi $\left\| DF(A, e_1, \lambda_1)^{\dagger} \right\| \leq \sqrt{2}$. □

Lemme 155. $\gamma(F, A, e_1, \lambda_1) \leq 2$.

Preuve Notons que $DF(A, x, \lambda)$ est surjectif dès que $x \neq 0$ ce qui est le cas ici (pour tout $y \in \mathbb{R}^n$ et $\mu \in \mathbb{R}$ on a $DF(A, x, \lambda)(-yx^T / \|x\|^2, \mu x / \|x\|^2, 0) = (y, \mu)$). Ainsi

$$\begin{aligned}
\gamma(F, A, e_1, \lambda_1) &= \frac{1}{2} \left\| DF(A, e_1, \lambda_1)^{\dagger} D^2 F(A, e_1, \lambda_1) \right\| \\
&\leq \frac{1}{2} \left\| DF(A, e_1, \lambda_1)^{\dagger} \right\| \left\| D^2 F(A, e_1, \lambda_1) \right\| \\
&\leq \frac{\sqrt{2}}{2} \left\| D^2 F(A, e_1, \lambda_1) \right\|.
\end{aligned}$$

Cette dérivée seconde vérifie

$$D^2 F(A, e_1, \lambda_1)(\dot{A}, \dot{x}, \dot{\lambda})^2 = \begin{pmatrix} 2(\dot{\lambda} - \dot{A})\dot{x} \\ \dot{x}^T \dot{x} \end{pmatrix}$$

de sorte que

$$\begin{aligned}\left\|D^2F(A,e_1,\lambda_1)(\dot{A},\dot{x},\dot{\lambda})^2\right\|^2 &= 4\left\|(\dot{\lambda}-\dot{A})\dot{x}\right\|^2+\|\dot{x}\|^4\\ &\leq 8\left\|\dot{\lambda}\dot{x}\right\|^2+8\left\|\dot{A}\dot{x}\right\|^2+\|\dot{x}\|^4\leq 8\left(\left\|\dot{\lambda}\right\|^2+\left\|\dot{A}\right\|_F^2+\|\dot{x}\|^2\right)^2.\end{aligned}$$

Ceci donne $\left\|D^2F(A,e_1,\lambda_1)\right\|\leq 2\sqrt{2}$ et $\gamma(F,A,e_1,\lambda_1)\leq 2\sqrt{2}\frac{\sqrt{2}}{2}=2$. □

Preuve du Théorème 150. Ce théorème est obtenu à partir du Théorème 128. Les Lemmes 154, 155 et 151 prouvent que

$$\beta(F,A,x,\lambda)\leq\sqrt{2}\,\|(\lambda I-A)x\|$$

et que

$$\alpha(F,A,x,\lambda)\leq 2\sqrt{2}\,\|(\lambda I-A)x\|\,.$$

L'hypothèse faite ici prouve que $\alpha(F,A,x,\lambda)\leq\alpha_0$ ce qui est l'hypothèse du Théorème 128. La conclusion s'en suit. □

5

La méthode de Newton-Gauss pour des systèmes sur-déterminés

5.1 Introduction

Nous considérons ici le cas de systèmes d'équations $f(x) = 0$, $f = (f_1, \ldots, f_m)$, $x \in \mathbb{R}^n$, où le nombre d'équations est plus grand que celui des inconnues. Un exemple académique est donné par la recherche d'une droite dans le plan qui doit passer par $m > 2$ points. Si ces points ne sont pas alignés une telle droite n'existe pas, autrement dit, le système correspondant n'a pas de solution. Pour de tels systèmes, on introduit un autre concept de solution : la solution au sens des moindres carrés ; on recherche ζ qui réalise le minimum de la fonction

$$F(x) = \frac{1}{2}\|f(x)\|^2 = \frac{1}{2}\sum_{k=1}^{m} |f_k(x)|^2$$

appelée « fonction résidu ». Notons que $F(\zeta) = 0$ si et seulement si $f(\zeta) = 0$: le concept de solution au sens des moindres carrés est plus général que celui de solution. Minimiser la fonction résidu est un problème d'optimisation globale donc, à priori, difficile à résoudre. Il est affaibli soit en un problème d'optimisation locale (recherche des minimum locaux de $F(x)$), soit en la recherche des points stationnaires de la fonction résidu ($DF(x) = 0$). Lorsque $F(x)$ est convexe, c'est le cas si les équations $f_k(x) = 0$ sont affines, ces trois concepts de solution au sens des moindres carrés coïncident. Ce n'est pas le cas en général.

Pour des problèmes linéaires, c'est à dire lorsque l'on recherche la solution d'un système $L(x) = b$, $L : \mathbb{R}^n \to \mathbb{R}^m$, si L est injective ou, ce qui revient au même, de rang n, ou bien $b \in \text{im } L$ et une unique solution existe, ou bien $b \notin \text{im } L$ et l'on recherche une solution au sens des moindres carrés :

$$\min_{x \in \mathbb{R}^n} \frac{1}{2}\|L(x) - b\|^2.$$

Elle est unique et donnée par

$$x = L^{\dagger}(b) = (L^*L)^{-1}L^*(b)$$

puisque l'inverse généralisé $L^\dagger$ de L est l'inverse à gauche de norme minimale de L (Théorème 120).

Dans le cas non linéaire on peut ramener l'étude des solutions au sens des moindres carrés à celle du système $DF(x) = 0$ qui a autant d'équations que d'inconnues et lui appliquer les méthodes de résolution habituelles. La méthode de Newton est donnée dans ce cas par l'opérateur

$$N_{DF}(x) = x - (Df(x)^* Df(x) + D^2 f(x)^* f(x))^{-1} Df(x)^* f(x)$$

en convenant que $D^2 f(x)^* f(x) v = (D^2 f(x)(., v))^* f(x)$.

Une autre statégie, introduite par Gauss en 1809, consiste à linéariser le système $f(x) = 0$ au voisinage d'un point x puis à résoudre ce nouveau système au sens des moindres carrés. Le problème linéarisé s'écrit

$$f(x) + Df(x)(y - x) = 0$$

et, lorsque $Df(x)$ est injectif, sa solution est égale à

$$y = x - Df(x)^\dagger f(x) = x - (Df(x)^* Df(x))^{-1} Df(x)^* f(x).$$

On obtient ainsi un opérateur de Newton associé à des problèmes surdéterminés. On le note encore $N_f(x)$. La méthode itérative correspondante, qui consiste à construire la suite $x_{k+1} = N_f(x_k)$, est connue sous le nom de méthode de Newton-Gauss. Ses propriétés sont extrêmement différentes du cas sous-déterminé : les points fixes de N_f correspondent aux points stationnaires de F et pas seulement aux zéros de f, ces points fixes ne sont pas nécessairement attractifs, si un point fixe ζ est attractif il correspond à un minimum local de la fonction résidu F et la convergence de la suite de Newton $x_{k+1} = N_f(x_k)$ vers ζ est quadratique si $F(\zeta) = 0$ et linéaire sinon.

5.2 Premières propriétés de la méthode de Newton-Gauss

5.2.1 L'inverse de Moore-Penrose pour des opérateurs injectifs

Soient $\mathbb{E}$ et $\mathbb{F}$ deux espaces de Hilbert. Notons $\mathcal{L}(\mathbb{E}, \mathbb{F})$ l'espace des applications linéaires continues $L : \mathbb{E} \to \mathbb{F}$ et $\mathcal{GL}(\mathbb{E}, \mathbb{F})$ l'ensemble des applications linéaires, continues, injectives $L : \mathbb{E} \to \mathbb{F}$ dont l'image est fermée dans $\mathbb{F}$. Rappelons que l'inverse de Moore-Penrose ou inverse généralisé de $L \in \mathcal{GL}(\mathbb{E}, \mathbb{F})$ est donné par

$$L^\dagger = (L^* L)^{-1} L^*$$

de sorte que $L^\dagger L = \mathrm{id}_{\mathbb{E}}$ et que $LL^\dagger = \Pi_{\mathrm{im}\ L}$ la projection orthogonale sur im L.

Lemme 156. *L'ensemble $\mathcal{GL}(\mathbb{E}, \mathbb{F})$ est ouvert dans l'espace $\mathcal{L}(\mathbb{E}, \mathbb{F})$, l'application $L \in \mathcal{GL}(\mathbb{E}, \mathbb{F}) \to L^\dagger \in \mathcal{L}(\mathbb{F}, \mathbb{E})$ est de classe C^∞ et sa dérivée a pour expression*

$$DL^\dagger(F) = -L^\dagger F L^\dagger + (L^*L)^{-1}F^*\Pi_{(im\ L)^\perp}$$

où $\Pi_{(im\ L)^\perp}$ *désigne la projection orthogonale sur* $(im\ L)^\perp$.

Preuve Soit $A \in \mathbb{L}(\mathbb{E},\mathbb{F})$ avec $\|A\| < \|L^\dagger\|^{-1}$. Nous allons voir que $L+A$ est injectif. En effet $L^\dagger(L+A) = \mathrm{id} + L^\dagger A$ puisque L est injectif et comme

$$\|L^\dagger A\| \leq \|L^\dagger\|\|A\| < \|L^\dagger\|\|L^\dagger\|^{-1} = 1,$$

par le Lemme 86 cet opérateur est inversible. On en déduit que

$$\Pi_{\mathrm{im}\ L}(L+A) = (LL^\dagger)(L+A) = L(L^\dagger(L+A))$$

est la composée d'un opérateur injectif et d'un opérateur inversible. Donc $\Pi_{\mathrm{im}\ L}(L+A)$ est injectif et, a fortiori, $L+A$. Nous allons maintenant prouver que si $\|A\| < \varepsilon$ pour un ε convenable alors $L+A$ est injectif et d'image fermée. Pour ce faire nous utilisons le résultat suivant (voir Brezis [10] sect. II.7)
« Soit $L \in \mathcal{L}(\mathbb{F},\mathbb{E})$. *Alors im* L *est fermé si et seulement s'il existe une constante* $C > 0$ *telle que*

$$d(x, \ker L) \leq C\|Lx\|$$

pour tout $x \in \mathbb{E}$. *»*
Soit $A \in \mathcal{L}(\mathbb{F},\mathbb{E})$ avec $\|A\| < \|L^\dagger\|^{-1}$ de sorte que $L+A$ est injectif. Puisque L est injectif et d'image fermée il existe $C > 0$ pour lequel

$$C^{-1}\|x\| \leq \|Lx\|.$$

On a alors

$$(C^{-1} - \|A\|)\|x\| \leq \|Lx\| - \|Ax\| \leq \|(L+A)x\|.$$

Si l'on prend $\|A\| < C^{-1}$ on obtient

$$\|x\| \leq (C^{-1} - \|A\|)^{-1}\|(L+A)x\|$$

ce qui prouve que $L+A$ est d'image fermée et que $\mathcal{GL}(\mathbb{E},\mathbb{F})$ est ouvert dans $\mathcal{L}(\mathbb{E},\mathbb{F})$.

Puisque L est injectif, $L^\dagger = (L^*L)^{-1}L^*$ qui est une application C^∞. Elle se dérive en

$$DL^\dagger(F) = -(L^*L)^{-1}(F^*L + L^*F)(L^*L)^{-1}L^* + (L^*L)^{-1}F^*$$

en utilisant le fait que $DA^{-1}(E) = -A^{-1}EA^{-1}$. On obtient

$$DL^\dagger(F) = -L^\dagger F L^\dagger + (L^*L)^{-1}F^*(\mathrm{id}_\mathbb{F} - LL^\dagger)$$

puis on note que $\mathrm{id}_\mathbb{F} - LL^\dagger = \Pi_{(\mathrm{im}\ L)^\perp}$ par le Théorème 120. □

Lemme 157. *Soit $M \in \mathcal{GL}(\mathbb{E},\mathbb{F})$. On a*

$$\|M^\dagger\|^2 = \|(M^*M)^{-1}\|.$$

Preuve On utilise le fait que $\|A\|^2 = \|AA^*\|$ et l'identité $M^\dagger = (M^*M)^{-1}M^*$.□

Lemme 158. *Soient L et $M \in \mathcal{GL}(\mathbb{E},\mathbb{F})$. Notons*

$$\mu_L = \inf_{\|x\|=1} \|Lx\|.$$

On a :

$$\|L^\dagger\| = \mu_L^{-1} \quad et \quad \|\mu_L - \mu_M\| \leq \|L - M\|.$$

Preuve La première inégalité provient de

$$\begin{aligned}\|L^\dagger\| &= \sup_{\|y\|=1} \|L^\dagger y\| = \sup_{\|y\|=1} \|L^\dagger \left(\Pi_{\text{im } L} y\right)\| \\ &= \sup_x \|L^\dagger \left(\frac{Lx}{\|Lx\|}\right)\| = \sup_x \frac{\|x\|}{\|Lx\|} = \frac{1}{\inf\limits_x \dfrac{\|Lx\|}{\|x\|}} = \mu_L^{-1}.\end{aligned}$$

La seconde inégalité se prouve ainsi :

$$\begin{aligned}\mu_L - \mu_M &= \inf_{\|x\|=1} \|Lx\| - \inf_{\|y\|=1} \|My\| = \inf_{\|x\|=1} \sup_{\|y\|=1} \|Lx\| - \|My\| \\ &\leq \sup_{\|y\|=1} \|Ly\| - \|My\| \leq \|L - M\|. \ \square\end{aligned}$$

Lemme 159. *Soient L et $M \in \mathcal{GL}(\mathbb{E},\mathbb{F})$. On a*

$$\|M^\dagger - L^\dagger\| \leq \sqrt{2}\|L^\dagger\|\|M^\dagger\|\|L - M\|.$$

Preuve On a

$$M^\dagger - L^\dagger = -M^\dagger(M - L)L^\dagger + (M^*M)^{-1}(M - L)^*\Pi_{(\text{im } L)^\perp}.$$

En effet ce second membre vaut

$$\begin{aligned}&-(M^\dagger M)L^\dagger + M^\dagger(LL^\dagger) + ((M^*M)^{-1}M^*)\Pi_{(\text{im } L)^\perp} - (M^*M)^{-1}(L^*\Pi_{(\text{im } L)^\perp}) \\ &\quad = -L^\dagger + M^\dagger\Pi_{(\text{im } L)} + M^\dagger\Pi_{(\text{im } L)^\perp} - M^\dagger 0 = M^\dagger - L^\dagger.\end{aligned}$$

Supposons que $\|M^\dagger\| \leq \|L^\dagger\|$. Dans le cas contraire il faudrait permuter les rôles de L et M. Soit $v \in \mathbb{F}$, $v = v_1 + v_2 \in \text{im } L \oplus (\text{im } L)^\perp$. On a :

$$(M^\dagger - L^\dagger)v = -M^\dagger(M - L)L^\dagger v_1 + (M^*M)^{-1}(M - L)^*\Pi_{(\text{im } L)^\perp} v_2$$

et comme $\|(M^*M)^{-1}\| = \|M^\dagger\|^2 \leq \|L^\dagger\|\|M^\dagger\|$ par l'hypothèse et le Lemme 157 on obtient

$$\begin{aligned}\|(M^\dagger - L^\dagger)v\| &\leq \|L^\dagger\|\|M^\dagger\|\|L - M\|(\|v_1\| + \|v_2\|) \\ &\leq \sqrt{2}\|L^\dagger\|\|M^\dagger\|\|L - M\|\|v\|. \ \square\end{aligned}$$

5.2.2 L'opérateur de Newton-Gauss et ses points fixes

Soient $\mathbb{E}$ et $\mathbb{F}$ deux espaces de Hilbert et soit $f : \mathbb{E} \to \mathbb{F}$ une application de classe C^1 définie sur $\mathbb{E}$ ou sur un ouvert U de $\mathbb{E}$. Supposons que l'image de $Df(x)$ soit fermée dans $\mathbb{F}$ de sorte que l'inverse généralisé de $Df(x)$ soit défini. On définit l'opérateur de Newton-Gauss par

$$N_f(x) = x - Df(x)^\dagger f(x)$$

et la fonction résidu par

$$F(x) = \frac{1}{2}\|f(x)\|^2.$$

Lemme 160. *Pour tout $u \in \mathbb{E}$ on a*

$$DF(x)u = \langle Df(x)u, f(x)\rangle = \langle u, Df(x)^* f(x)\rangle .$$

Preuve On applique le théorème de dérivation des fonctions composées à $F = \frac{1}{2}\|.\|^2 \circ f$ en notant que

$$D\left(\frac{1}{2}\|.\|^2\right)(y) = \langle ., y\rangle .$$

Ainsi

$$\begin{aligned} DF(x)u &= D\left(\frac{1}{2}\|.\|^2\right)(f(x)) \circ Df(x)(u) = \langle Df(x)u, f(x)\rangle \\ &= \langle u, Df(x)^* f(x)\rangle . \square \end{aligned}$$

Proposition 161. *Les énoncés suivant sont équivalents :*

1. $N_f(x) = x$,
2. $f(x) \in \ker Df(x)^\dagger$,
3. $f(x) \in (im\ Df(x))^\perp$,
4. $f(x) \in \ker Df(x)^*$,
5. $DF(x) = 0$.

   Lorsque $Df(x)$ est surjectif ces énoncés sont équivalents à
6. $f(x) = 0$.

Preuve 1 signifie que $Df(x)^\dagger f(x) = 0$ c'est à dire 2 ou 3 ou bien 4 puisque $\ker Df(x)^\dagger = \text{im}\ Df(x)^\perp = \ker Df(x)^*$ par le Théorème 120. Ceci est équivalent à 5 par le lemme 160. Lorsque $Df(x)$ est surjectif on a $\text{im}\ Df(x)^\perp = 0$ et on conclut par 3. $\square$

Lemme 162. *Supposons que f soit de classe C^2. Alors*

$$D^2F(x) = Df(x)^* Df(x) + D^2 f(x)^* f(x)$$

c'est à dire que, pour tout $u, v \in \mathbb{E}$,

$$D^2F(x)(u,v) = \left\langle u, (D^2 f(x)(., v))^* f(x)\right\rangle + \langle u, Df(x)^* Df(x)v\rangle .$$

Preuve On dérive la formule donnée en 160 :

$$D^2F(x)(u,v) = \langle D^2f(x)(u,v), f(x)\rangle + \langle Df(x)u, Df(x)v\rangle$$
$$= \langle u, (D^2f(x)(.,v))^* f(x)\rangle + \langle u, Df(x)^* Df(x)v\rangle. \square$$

Lemme 163. *Lorsque f est de classe C^2, $Df(x)$ injectif et im $Df(x)$ fermée on a :*

1. $N_f(x) = x - Df(x)^\dagger f(x) = x - (Df(x)^* Df(x))^{-1} Df(x)^* f(x)$,
2. $DN_f(x) = Df(x)^\dagger D^2f(x)Df(x)^\dagger f(x)$
$$-(Df(x)^* Df(x))^{-1} D^2f(x)^* \Pi_{(im\ Df(x))^\perp} f(x),$$
en convenant que
$$D^2f(x)^* \Pi_{(im\ Df(x))^\perp} f(x)u = (D^2f(x)(.,u))^* \Pi_{(im\ Df(x))^\perp} f(x),$$
3. *Si de plus $F(x) = 0$ alors $DN_f(x) = -(Df(x)^* Df(x))^{-1} D^2f(x)^* f(x)$.*

Preuve La première formule provient de la description de $Df(x)^\dagger$ dans le cas injectif. La seconde est une conséquence du Lemme 156 et du théorème de dérivation des fonctions composées. Pour prouver le troisième énoncé on note que $Df(x)^\dagger f(x) = 0$ et que $f(x) \in (\text{im } Df(x))^\perp$ lorsque $F(x) = 0$ par la Proposition 161. $\square$

A la différence du cas surjectif les points fixes de N_f ne sont pas nécessairement attractifs. C'est ce que nous prouvons dans le résultat suivant :

Théorème 164. *Supposons que f soit de classe C^2, soit ζ un zéro de f au sens des moindres carrés, c'est à dire tel que $DF(\zeta) = 0$, et supposons enfin que $Df(\zeta)$ soit injectif. Alors*

1. *Les valeurs spectrales de $DN_f(\zeta)$ sont réelles et ce sont des valeurs propres,*
2. *Si ζ est un point fixe attractif de N_f alors c'est un minimum local strict de F,*
3. *Si ζ est un maximum local strict de F alors c'est un point fixe répulsif de N_f.*

Preuve Nous avons vu aux Lemmes 163 et 162 que

$$DN_f(\zeta) = -(Df(\zeta)^* Df(\zeta))^{-1} D^2f(\zeta)^* f(\zeta)$$

et que

$$D^2F(\zeta) = Df(\zeta)^* Df(\zeta) + D^2f(\zeta)^* f(\zeta).$$

Ecrivons cela $D^2F(\zeta) = b + a$ et $DN_f(\zeta) = -b^{-1}a$ avec des notations évidentes. Notons que a c'est un opérateur symétrique parce que b et $D^2F(\zeta)$ le sont. Quant à b, c'est un opérateur défini positif puisque

$$\langle bx, x\rangle = \langle Df(\zeta)^* Df(\zeta)x, x\rangle = \langle Df(\zeta)x, Df(\zeta)x\rangle > 0$$

dès que $x \neq 0$ puisque $Df(\zeta)$ est injectif. Pour un tel opérateur il existe une unique application linéaire, continue et positive dont le carré soit b : on la note $b^{1/2}$, elle s'appelle le racine carrée de b (voir [39] Th. 12.33). Notons Spec (b) le spectre de b (paragraphe 2.4.1). Pour deux opérateurs linéaires et continus l et m, si l possède un inverse continu alors Spec (lm) = Spec (ml). En effet $\lambda \,\mathrm{id} - ml = l^{-1}(\lambda \,\mathrm{id} - lm)l$ de sorte que $\lambda \,\mathrm{id} - ml$ est inversible si et seulement si $\lambda \,\mathrm{id} - lm$ est inversible. On a :

$$\begin{aligned}\text{Spec } DN_f(\zeta) &= \text{Spec } (-b^{-1}a) = \text{Spec } (-b^{-1/2}b^{-1/2}a)\\ &= \text{Spec } (-b^{-1/2}ab^{-1/2}) = \text{Spec } (\mathrm{id} - b^{-1/2}(a+b)b^{-1/2}).\end{aligned}$$

Remarquons que $b^{-1/2}(a+b)b^{-1/2}$ est un opérateur réel et symétrique. Ses valeurs spectrales sont donc des valeurs propres et elles sont réelles (voir [56] Chap. XI-8, Théorème 1). Il en est donc de même pour celles de $DN_f(\zeta)$.

Si ζ est un point fixe attractif pour N_f, par le Théorème 20, Spec $(DN_f(\zeta))$ est contenu dans l'intervalle $]-1, 1[$ et donc

$$\sigma(b^{-1/2}(a+b)b^{-1/2}) \subset]0, 2[.$$

Cela signifie que $b^{-1/2}(a+b)b^{-1/2}$ est positif (voir [39] Th. 12.32) donc aussi $a+b$ puisque $b^{-1/2}$ est symétrique et inversible. Ainsi $D^2F(\zeta)$ est positif. C'est un résultat classique en optimisation qu'un minimum local strict ζ d'une fonction F de classe C^2 soit caractérisé par les conditions $DF(\zeta) = 0$ et $D^2F(\zeta)$ positif. Ceci établit la seconde assertion.

La dernière assertion se prouve par des arguments similaires : lorsque ζ est un maximum local strict pour F on a $-D^2F(\zeta)$ est positif, Spec $(a+b)$ $\subset]-\infty, 0[$, de même

$$\text{Spec } (b^{-1/2}(a+b)b^{-1/2}) \subset]-\infty, 0[$$

et donc

$$\text{Spec } (DN_f(\zeta)) \subset]0, \infty[.$$

En vertu du Théorème 20 cela fait de ζ un point fixe répulsif. □

Remarquons le bon comportement de la méthode de Newton-Gauss vis à vis des solutions du problème au sens des moindres carrés. Si la méthode converge, on est assuré de calculer un minimum local de la fonction résidu et pas seulement un de ses points stationnaires.

Considérons l'exemple suivant :

$$f(x) = \begin{pmatrix} x \\ x^2 + a \end{pmatrix}, \quad f : \mathbb{R} \to \mathbb{R}^2.$$

Lorsque $a = 0$, $x = 0$ est un zéro de f et lorsque $a \neq 0$, $f(0) \neq 0$. L'itération de Newton-Gauss est donnée par

$$N_f(x) = x - \frac{2x^3 + (2a+1)x}{4x^2 + 1}$$

et la fonction résidu par

$$F(x) = \frac{1}{2}(x^4 + (2a+1)x^2 + a^2).$$

De plus $DN_f(0) = -2a$, $DF(0) = 0$ et $D^2F(0) = 2a + 1$. 0 est un point fixe de N_f, super-attractif si $a = 0$, attractif si $|a| < 1/2$ et dans ce cas 0 est le minimum de F. Lorsque $|a| > 1/2$, 0 est un point fixe répulsif : si $a < -1/2$ c'est un maximum local de F et si $a > 1/2$ c'est le minimum de F. Cet exemple montre bien que la solution au sens des moindres carrés n'est pas nécessairement accessible par la méthode de Newton-Gauss.

5.3 Théorèmes de convergence pour la méthode de Newton-Gauss

Le théorème qui suit est du type « Kantorovitch ». On y décrit une condition suffisante pour qu'un point fixe de N_f soit attractif ainsi que la vitesse de convergence de la suite des itérés.

Théorème 165. *Supposons que f soit de classe C^2. Soit $\zeta \in \mathbb{E}$ tel que $Df(\zeta)$ soit injective et d'image fermée.*

1. *Si $f(\zeta) = 0$, il existe $r > 0$ tel que, pour tout $x \in \mathbb{E}$, $\|x - \zeta\| \leq r$, la suite de Newton-Gauss $x_k = N_f^k(x)$ soit définie, converge vers ζ et vérifie*

$$\|x_k - \zeta\| \leq \left(\frac{1}{2}\right)^{2^k - 1} \|x - \zeta\|$$

   pour tout $k \geq 0$.
2. *Si $DF(\zeta) = 0$ et si*

$$\|Df(\zeta)^\dagger\|^2 \|D^2 f(\zeta)\| \|f(\zeta)\| < 1$$

   il existe $r > 0$ et $0 \leq \lambda < 1$ tels que, pour tout $x \in \mathbb{E}$, $\|x - \zeta\| \leq r$, la suite de Newton-Gauss $x_k = N_f^k(x)$ soit définie, converge vers ζ et vérifie

$$\|x_k - \zeta\| \leq \lambda^k \|x - \zeta\|$$

   pour tout $k \geq 0$.

Preuve Lorsque $f(\zeta) = 0$, en vertu du Lemme 163, on a

$$DN_f(\zeta) = -(Df(\zeta)^* Df(\zeta))^{-1} D^2 f(\zeta)^* f(\zeta) = 0$$

ce qui fait de ζ un point fixe super-attractif par le Théorème 7. Lorsque $DF(\zeta) = 0$, on a

$$\|DN_f(\zeta)\| = \|(Df(\zeta)^* Df(\zeta))^{-1} D^2 f(\zeta)^* f(\zeta)\| \leq \|(Df(\zeta)^* Df(\zeta))^{-1}\|^2 \times \|D^2 f(\zeta)\| \|f(\zeta)\|$$

ce qui, par le Lemme 157 et l'hypothèse, donne

$$\|DN_f(\zeta)\| \leq \|Df(\zeta))^\dagger\|^2 \|D^2 f(\zeta)\| \|f(\zeta)\| < 1.$$

En vertu du Lemme 156, N_f est de classe C^1 au voisinage de ζ de sorte qu'on peut supposer que

$$\|DN_f(x)\| \leq \lambda < 1$$

pour tout $x \in \mathbb{E}$, $\|x - \zeta\| \leq r$, pour des constantes $r > 0$ et λ, $0 \leq \lambda < 1$, convenables. Ceci fait de N_f une contraction sur la boule fermée ainsi définie, de constante de contraction λ et de point fixe ζ. Il suffit alors d'appliquer le Théorème 5. □

Nous allons préciser les résultats du théorème précédent lorsque f est analytique. Le point de vue que nous adoptons est celui de la théorie alpha de Smale, les résultats présentés proviennent de Dedieu-Shub [14] et Dedieu-Kim [13]. Le contexte de cette section est le suivant : $f : \mathbb{E} \to \mathbb{F}$ est une application analytique entre deux espaces de Hilbert ou bien définie sur un ouvert de $\mathbb{E}$. On suppose que $Df(x)$ est d'image fermée dans $\mathbb{F}$ pour tout x dans le domaine de définition de f.

Nous utiliserons les invariants $\alpha(f, x)$, $\beta(f, x)$ et $\gamma(f, x)$ introduits à la Définition 127 dans le contexte des systèmes sous-déterminés. Nous devons les redéfinir dans notre nouveau contexte. En effet, $Df(x)^\dagger$ a pour noyau $(\text{im } Df(x))^\perp$ de sorte que l'action de cet opérateur sur un vecteur ne prend en compte que la composante de ce vecteur contenue dans im $Df(x)$.

Définition 166. *Pour tout $x \in \mathbb{E}$ posons*

- $\alpha_1(f, x) = \beta_1(f, x)\gamma_1(f, x)$,
- $\beta_1(f, x) = \|Df(x)^\dagger\| \|(f(x))\|$,
- $\gamma_1(f, x) = \sup_{k \geq 2} \left(\|Df(x)^\dagger\| \left\| \frac{D^k f(x)}{k!} \right\| \right)^{\frac{1}{k-1}}$.

5.3.1 Enoncé des résultats principaux

Rappelons que $\psi(v) = 1 - 4v + 2v^2$. Cette fonction décroit de 1 à 0 sur l'intervalle $[0, 1 - \frac{\sqrt{2}}{2}]$. Le bassin d'attraction quadratique d'un zéro de f est donné par :

Théorème 167. *Soient x et $\zeta \in \mathbb{E}$ tels que $f(\zeta) = 0$, que $Df(\zeta)$ soit injectif et que*

$$v = \|x - \zeta\|\gamma_1(f, \zeta) \leq \frac{3 - \sqrt{7}}{2}.$$

Alors la suite de Newton $x_k = N_f^k(x)$ vérifie

$$\|x_k - \zeta\| \leq \left(\frac{1}{2}\right)^{2^k - 1} \|x - \zeta\|.$$

Pour un zéro au sens des moindres carrés on a :

Théorème 168. *Soient x et $\zeta \in \mathbb{E}$ tels que $Df(\zeta)^\dagger f(\zeta) = 0$, que $Df(\zeta)$ soit injectif et que*

$$v = \|x - \zeta\|\gamma_1(f, \zeta) < 1 - \frac{\sqrt{2}}{2}.$$

Supposons que

$$\lambda = \frac{1}{\psi(v)}(v + \sqrt{2}(2 - v)\alpha_1(f, \zeta)) < 1.$$

Alors la suite de Newton $x_k = N_f^k(x)$ vérifie

$$\|x_k - \zeta\| \leq \lambda^k \|x - \zeta\|.$$

Remarque 10. Puisque $v \to 0$ lorsque $x \to \zeta$ la condition $\lambda < 1$ is satisfaite pour tout x dans une boule de centre ζ et de rayon convenable dès que

$$\alpha_1(f, \zeta) < \frac{1}{2\sqrt{2}}.$$

Dans le théorème qui suit on donne une condition suffisante pour qu'une suite de Newton converge vers un zéro du système au sens des moindres carrés. A la différence du théorème précédent l'existence de ce zéro est prouvée en cours de route et non pas donnée en hypothèse.

Théorème 169. *Soit $x \in \mathbb{E}$ tel que $Df(x)$ soit injectif. Notons*

$$\kappa = \|Df(x)\|\|Df(x)^\dagger\|,$$

$$\lambda = \frac{1}{8\kappa + 16},$$

$$\Lambda = 4\frac{1 - \lambda}{\psi(\lambda)^2}\left(\frac{1}{16\kappa + 32} + \frac{\lambda^2}{1 - \lambda} + \kappa\lambda\right).$$

On a $0 \leq \Lambda < 1$. Supposons que

$$\alpha_1(f, x) \leq \frac{1}{16\kappa + 32}.$$

Alors

1. N_f *envoie* $\bar{B}\left(x, \dfrac{\lambda}{\gamma_1(f,x)}\right)$ *dans elle-même,*
2. N_f *est une contraction sur cette boule, de constante de contraction* Λ*,*
3. *Il existe un unique* $\zeta \in \mathbb{E}$ *tel que* $Df(\zeta)^\dagger f(\zeta) = 0$ *et*

$$\|\zeta - x\| < \frac{\lambda}{\gamma_1(f,x)},$$

4. *La suite de Newton* $x_k = N_f^k(x)$ *converge vers* ζ *et*

$$\|x_k - \zeta\| \leq \Lambda^k \|x - \zeta\|.$$

5.3.2 Démonstration des résultats principaux : lemmes préliminaires

Lemme 170. *Lorsque* $Df(x)$ *est injectif et que*

$$u = \|x - y\|\gamma_1(f,x) < 1 - \frac{\sqrt{2}}{2}$$

alors

1. $Df(y)$ *et* $\Pi_{im\ Df(x)} Df(y)$ *sont injectifs,*
2. $Df(x)^\dagger Df(y)$ *est inversible et son inverse est égal à*

$$\left(\Pi_{im\ Df(x)} Df(y)\right)^\dagger\ Df(x)$$

3. $\|(Df(x)^\dagger Df(y))^{-1}\| \leq \dfrac{(1-u)^2}{\psi(u)}$.

Preuve Par la formule de Taylor

$$Df(x)^\dagger(Df(x) - Df(y)) = -Df(x)^\dagger \sum_{k\geq 2} k \frac{D^k f(x)}{k!}(y-x)^{k-1}$$

de sorte que

$$\begin{aligned}\|Df(x)^\dagger(Df(x) - Df(y))\| &\leq \sum_{k\geq 2} k\|Df(x)^\dagger\| \frac{\|D^k f(x)\|}{k!}\|y-x\|^{k-1} \\ &\leq \sum_{k\geq 2} k\gamma_1(f,x)^{k-1}\|y-x\|^{k-1} = \frac{1}{(1-u)^2} - 1 < 1\end{aligned}$$

parce que $u < 1 - \frac{\sqrt{2}}{2}$. Par le Lemme 86 id $- Df(x)^\dagger(Df(x) - Df(y)$ $= Df(x)^\dagger Df(y)$ est inversible et la norme de son inverse est bornée par

$$\|(Df(x)^\dagger Df(y))^{-1}\| \leq \frac{1}{1-\left(\frac{1}{(1-u)^2}-1\right)} = \frac{(1-u)^2}{\psi(u)}.$$

De plus

$$\begin{aligned}&(\Pi_{\text{im } Df(x)} Df(y))^\dagger Df(x)(Df(x)^\dagger Df(y))\\ &\quad = (\Pi_{\text{im } Df(x)} Df(y))^\dagger \circ \left(\prod_{\text{im } Df(x)} Df(y)\right) = \text{id}\end{aligned}$$

ce qui prouve 2. $Df(y)$ est injectif parce que $\Pi_{\text{im } Df(x)} Df(y)$ est aussi injectif. □

Lemme 171. *Lorsque $Df(x)$ est injectif et si*

$$u = \|x-y\|\gamma_1(f,x) < 1 - \frac{\sqrt{2}}{2}$$

alors

$$\|Df(y)^\dagger\| \leq \|Df(x)^\dagger\| \frac{(1-u)^2}{\psi(u)}.$$

Preuve Par la formule de Taylor on a

$$Df(y) = Df(x) + \sum_{k\geq 2} k \frac{D^k f(x)}{k!}(y-x)^{k-1}$$

de sorte que, par une majoration désormais familière,

$$\|Df(y) - Df(x)\| \leq \|Df(x)^\dagger\|^{-1}\left(\frac{1}{(1-u)^2}-1\right).$$

Par le Lemme 170, $Df(x)$ et $Df(y)$ sont tous deux injectifs aussi, par le Lemme 158,

$$|\mu_{Df(x)} - \mu_{Df(y)}| \leq \|Df(y) - Df(x)\|.$$

De plus $\mu^{-1}_{Df(y)} = \|Df(y)^\dagger\|$, $\mu^{-1}_{Df(x)} = \|Df(x)^\dagger\|$ ce qui donne

$$\begin{aligned}\mu_{Df(y)} &\geq \mu_{Df(x)} - \|Df(y) - Df(x)\| \geq \mu_{Df(x)}\left(2 - \frac{1}{(1-u)^2}\right)\\ &= \mu_{Df(x)} \frac{\psi(u)}{(1-u)^2}\end{aligned}$$

puisque $u < 1 - \frac{\sqrt{2}}{2}$. On en déduit que

$$\begin{aligned}\|Df(y)^\dagger\| = \mu^{-1}_{Df(y)} &= \mu^{-1}_{Df(x)}(\mu_{Df(x)}\mu^{-1}_{Df(y)}) \leq \mu^{-1}_{Df(x)} \frac{(1-u)^2}{\psi(u)}\\ &= \|Df(x)^\dagger\| \frac{(1-u)^2}{\psi(u)}.\ \square\end{aligned}$$

Lemme 172. *Soit $\zeta \in \mathbb{E}$ avec $Df(\zeta)^\dagger f(\zeta) = 0$ et $Df(\zeta)$ injectif. Pour tout $x \in \mathbb{E}$ tel que*

$$v = \|\zeta - x\|\gamma_1(f,\zeta) < 1 - \frac{\sqrt{2}}{2}$$

on a

$$\|Df(x)^\dagger f(\zeta)\| \leq \sqrt{2}\frac{2v - v^2}{\psi(v)}\beta_1(f,\zeta).$$

Preuve Par le Lemme 159

$$\begin{aligned}\|Df(x)^\dagger f(\zeta)\| &= \|(Df(x)^\dagger - Df(\zeta)^\dagger)f(\zeta)\| \\ &\leq \sqrt{2}\|Df(x)^\dagger\|\|Df(\zeta)^\dagger\|\|Df(x) - Df(\zeta)\|\|f(\zeta)\|.\end{aligned}$$

On utilise le Lemme 171 pour majorer $\|Df(x)^\dagger\|$ ainsi que l'inégalité

$$\|Df(x) - Df(\zeta)\| \leq \|Df(\zeta)^\dagger\|^{-1}\left(\frac{1}{(1-v)^2} - 1\right)$$

que l'on obtient comme dans la preuve du Lemme 171. Ainsi

$$\begin{aligned}\|Df(x)^\dagger f(\zeta)\| \leq \sqrt{2}\|Df(\zeta)^\dagger\|&\frac{(1-v)^2}{\psi(v)}\|Df(\zeta)^\dagger\|\|Df(\zeta)^\dagger\|^{-1} \\ &\times\left(\frac{1}{(1-v)^2} - 1\right)\|f(\zeta)\| = \sqrt{2}\frac{2v - v^2}{\psi(v)}\beta_1(f,\zeta)\end{aligned}$$

d'où le résultat. □

Lemme 173. *Soit $\zeta \in \mathbb{E}$ avec $Df(\zeta)^\dagger f(\zeta) = 0$ et $Df(\zeta)$ injectif. Pour tout $x \in \mathbb{E}$ tel que*

$$v = \|\zeta - x\|\gamma_1(f,\zeta) < 1 - \frac{\sqrt{2}}{2}$$

on a

$$\|N_f(x) - \zeta\| \leq \|x - \zeta\|\frac{v}{\psi(v)} + \sqrt{2}\frac{2v - v^2}{\psi(v)}\beta_1(f,\zeta).$$

Preuve On a

$$\begin{aligned}N_f(x) - \zeta &= x - \zeta - Df(x)^\dagger f(x) \\ &= Df(x)^\dagger((Df(x)(x-\zeta) - f(x) + f(\zeta)) - Df(x)^\dagger f(\zeta))\end{aligned}$$

de sorte que

$$\|N_f(x) - \zeta\| \leq \|Df(x)^\dagger\|\|Df(x)(x-\zeta) + f(\zeta) - f(x)\| + \|Df(x)^\dagger f(\zeta)\|.$$

De plus, par la formule de Taylor,

$$Df(x)(x-\zeta) - f(x) + f(\zeta) = \sum_{k\geq 1}(k-1)\frac{D^k f(\zeta)}{k!}(z-\zeta)^k$$

d'où l'estimation

$$\|Df(x)(x-\zeta)-f(x)+f(\zeta)\| \le \|Df(\zeta)^\dagger\|^{-1}\|x-\zeta\|\frac{v}{(1-v)^2}.$$

On applique alors les Lemmes 171 et 172 ce qui donne

$$\|N_f(x)-\zeta\| \le \frac{(1-v)^2}{\psi(v)}\|x-\zeta\|\frac{v}{(1-v)^2}+\sqrt{2}\frac{2v-v^2}{\psi(v)}\beta_1(f,v). \ \square$$

Lemme 174. *Lorsque $Df(x)$ est injectif on a*

$$\|DN_f(x)\| \le 4\alpha_1(f,x).$$

Preuve C'est une conséquence des Lemmes 163 et 157. On obtient $\|DN_f(x)\|$

$$= \left\|Df(x)^\dagger D^2f(x)Df(x)^\dagger f(x)-(Df(x)^*Df(x))^{-1}D^2f(x)^*\Pi_{\mathrm{im}\ Df(x)^\perp}f(x)\right\|$$

$$\le 2\|Df(x)^\dagger\|\|D^2f(x)\|\|Df(x)^\dagger\|\|f(x)\| \le 4\gamma_1(f,x)\beta_1(f,x) = 4\alpha_1(f,x). \ \square$$

Lemme 175. *Lorsque $Df(x)$ est injectif et si*

$$u = \|x-y\|\gamma_1(f,x) < 1-\frac{\sqrt{2}}{2}$$

on a

1. $\beta_1(f,y) \le \dfrac{(1-u)^2}{\psi(u)}\left(\beta_1(f,x)+\dfrac{u}{1-u}\|y-x\|+\|Df(x)\|\|Df(x)^\dagger\|\|y-x\|\right),$
2. $\gamma_1(f,y) \le \dfrac{\gamma_1(f,x)}{(1-u)\psi(u)},$
3. $\alpha_1(f,y) \le \dfrac{1-u}{\psi(u)^2}\left(\alpha_1(f,x)+\dfrac{u^2}{1-u}+\|Df(x)\|\|Df(x)^\dagger\|u\right).$

Preuve 3 est une conséquence de 1 et 2. Pour prouver 1 on utilise

$$f(y) = f(x)+Df(x)(y-x)+\sum_{k\ge 2}\frac{D^kf(x)}{k!}(y-x)^k$$

qui donne

$$\|f(y)\| \le \|f(x)\|+\|Df(x)\|\|y-x\|+\|Df(x)^\dagger\|^{-1}\|y-x\|\frac{u}{1-u}$$

d'où, par le Lemme 171,

$$\begin{aligned}\beta_1(f,y) &= \|Df(y)^\dagger\|\|f(y)\| \le \|Df(x)^\dagger\|\frac{(1-u)^2}{\psi(u)}\|f(y)\| \\ &\le \|Df(x)^\dagger\|\frac{(1-u)^2}{\psi(u)}\Big(\|f(x)\|+\|Df(x)\|\|y-x\| \\ &\quad +\|Df(x)^\dagger\|^{-1}\|y-x\|\frac{u}{1-u}\Big) \\ &= \frac{(1-u)^2}{\psi(u)}\left(\beta_1(f,x)+\frac{u}{1-u}\|y-x\|+\|Df(x)\|\|Df(x)^\dagger\|\|y-x\|\right).\end{aligned}$$

Pour prouver 2 on part de

$$\frac{D^k f(y)}{k!} = \sum_{l=0}^{\infty} \frac{D^{k+l} f(x)}{k!l!}(y-x)^l$$

de sorte que

$$\begin{aligned}\frac{\left\|D^k f(y)\right\|}{k!} &\leq \sum_l \binom{k+l}{l} \frac{\left\|D^{k+l} f(x)\right\|}{(k+l)!}\|y-x\|^l \\ &\leq \|Df(x)^\dagger\|^{-1} \sum_l \binom{k+l}{l} \frac{\gamma_1(f,x)^{k+l-1}}{(k+l)!}\|y-x\|^l \\ &= \|Df(x)^\dagger\|^{-1} \frac{\gamma_1(f,x)^{k-1}}{(1-u)^{k+1}}.\end{aligned}$$

Par le Lemme 171 on a

$$\|Df(y)^\dagger\| \frac{\|D^k f(y)\|}{k!} \leq \frac{(1-u)^2}{\psi(u)} \frac{\gamma_1(f,x)^{k-1}}{(1-u)^{k+1}}$$

et donc

$$\gamma_1(f,y) \leq \frac{\gamma_1(f,x)}{(1-u)\psi(u)}$$

puisque dans l'intervalle $0 \leq u < 1 - \sqrt{2}/2$ on a $0 < \psi(u) \leq 1$. □

5.3.3 Démonstration du Théorème 167

Lorsque $f(\zeta) = 0$ et $v = \|\zeta - x\|\gamma_1(f,\zeta) \leq \frac{3-\sqrt{7}}{2} < 1 - \frac{\sqrt{2}}{2}$ on a, par le Lemme 173,

$$\|N_f(x) - \zeta\| \leq \frac{v}{\psi(v)}\|x - \zeta\|.$$

Montrons par récurrence que

$$\|N_f^k(x) - \zeta\| \leq \left(\frac{v}{\psi(v)}\right)^{2^k-1} \|x - \zeta\|.$$

Posons $x_k = N_f^k(x)$ et $v_k = \|\zeta - x_k\|\gamma_1(f,\zeta)$. Puisque $v \leq \frac{3-\sqrt{7}}{2}$ on a $\frac{v}{\psi(v)} \leq 1$ et donc, par l'hypothèse de récurrence, $v_k \leq v < 1 - \frac{\sqrt{2}}{2}$. On peut donc appliquer le Lemme 173 à x_k ce qui donne

$$\|x_{k+1} - \zeta\| \leq \frac{v_k}{\psi(v_k)}\|x_k - \zeta\| = \frac{\gamma_1(f,\zeta)}{\psi(v_k)}\|x_k - \zeta\|^2$$

et par l'hypothèse de récurrence

$$\|x_{k+1} - \zeta\| \leq \frac{\gamma_1(f,\zeta)}{\psi(v_k)} \left(\left(\frac{v}{\psi(v)}\right)^{2^k-1} \|x - \zeta\|\right)^2.$$

Il suffit alors de remarquer que $\psi(v_k) \geq \psi(v)$ pour obtenir

$$\|x_{k+1} - \zeta\| \leq \left(\frac{v}{\psi(v)}\right)^{2^{k+1}-1} \|x - \zeta\|.$$

Le théorème s'obtient en notant que $\frac{v}{\psi(v)} \leq 1/2$ dès que $v \leq \frac{3-\sqrt{7}}{2}$. □

5.3.4 Démonstration du Théorème 168

Puisque $v = \|\zeta - x\|\gamma_1(f,\zeta) \leq \frac{3-\sqrt{7}}{2} < 1 - \frac{\sqrt{2}}{2}$ on a, par le Lemme 173,

$$\|N_f(x) - \zeta\| \leq \frac{v}{\psi(v)}\|x - \zeta\| + \sqrt{2}\frac{2v - v^2}{\psi(v)}\beta_1(f,\zeta),$$

ou, de façon équivalente,

$$\|N_f(x) - \zeta\| \leq \left(\frac{v}{\psi(v)} + \sqrt{2}\frac{2-v}{\psi(v)}\alpha_1(f,\zeta)\right)\|x - \zeta\| = \lambda\|x - \zeta\|.$$

Nous savons, par hypothèse, que

$$\lambda = \frac{v}{\psi(v)} + \sqrt{2}\frac{2-v}{\psi(v)}\alpha_1(f,\zeta) < 1.$$

On en déduit par récurrence que

$$\|N_f^k(x) - \zeta\| \leq \lambda^k\|x - \zeta\|. \ \square$$

5.3.5 Démonstration du Théorème 169

Par les Lemmes 174 et 175 on a

$$\|DN_f(y)\| \leq 4\frac{1-u}{\psi(u)^2}\left(\alpha_1(f,x) + \frac{u^2}{1-u} + \|Df(x)\|\|Df(x)^\dagger\|u\right)$$

pour tout y tel que $u = \|x-y\|\gamma_1(f,x) < 1 - \frac{\sqrt{2}}{2}$. Soit $\lambda < 1 - \frac{\sqrt{2}}{2}$. En utilisant le théorème des valeurs intermédiaires, pour tout y tel que $u = \|x-y\|\gamma_1(f,x) \leq \lambda$, on a

$$\begin{aligned}\|N_f(y) - x\| &\leq \|N_f(y) - N_f(x)\| + \|N_f(x) - x\| \\ &\leq \sup_{\|x-z\|\gamma_1(f,x)\leq\lambda} \|DN_f(z)\|\|x - y\| + \beta_1(f,x)\end{aligned}$$

puis, par l'inégalité ci-dessus et le fait que son second membre est une fonction croissante de u,

$$\begin{aligned}\|N_f(y) - x\| \leq 4\frac{1-\lambda}{\psi(\lambda)^2}&\left(\alpha_1(f,x) + \frac{\lambda^2}{1-\lambda} + \|Df(x)\|\|Df(x)^\dagger\|\lambda\right) \\ &\times\|x - y\| + \beta_1(f,x).\end{aligned}$$

Posons $\kappa = \|Df(x)\|\|Df(x)^\dagger\|$. Notons que $\kappa \geq 1$ puisque

$$1 = \|\mathrm{id}\| = \|Df(x)^\dagger Df(x)\| \leq \|Df(x)^\dagger\|\|Df(x)\| = \kappa.$$

On a

$$\|N_f(y) - x\| \leq 4\frac{1-\lambda}{\psi(\lambda)^2}(\alpha_1(f,x) + \frac{\lambda^2}{1-\lambda} + \kappa\lambda)\frac{\lambda}{\gamma_1(f,x)} + \frac{\alpha_1(f,x)}{\gamma_1(f,x)}.$$

Nous voulons prouver que N_f est une contraction qui envoie la boule fermée $\bar{B}\left(x, \frac{\lambda}{\gamma_1(f,x)}\right)$ dans elle même. Ce programme sera réalisé si la condition suivante est satisfaite

$$4\frac{1-\lambda}{\psi(\lambda)^2}(\alpha_1(f,x) + \frac{\lambda^2}{1-\lambda} + \kappa\lambda)\lambda + \alpha_1(f,x) \leq \lambda$$

c'est à dire si

$$\alpha_1(f,x) \leq \frac{\lambda\psi(\lambda)^2 - 4\lambda^3 - 4\kappa\lambda^2(1-\lambda)}{\psi(\lambda)^2 + 4\lambda(1-\lambda)}.$$

Comme $\psi(\lambda)^2 + 4\lambda(1-\lambda) \leq 1$ l'inégalité précédente est vérifiée si

$$\alpha_1(f,x) \leq \lambda\psi(\lambda)^2 + 4\lambda^3(\kappa - 1) - 4\kappa\lambda^2.$$

On affaiblit cette condition en remarquant que $1-8\lambda \leq \psi(\lambda)^2$ dans le domaine considéré et que $0 \leq \kappa - 1$ ce qui donne

$$\alpha_1(f,x) \leq \lambda(1-8\lambda) - 4\kappa\lambda^2.$$

Afin que cette condition ne soit pas vide il faut que son membre de droite soit positif pour un $\lambda < 1 - \frac{\sqrt{2}}{2}$. Ceci est réalisé avec

$$\lambda = \frac{1}{8(\kappa+2)} \leq \frac{1}{24} < 1 - \frac{\sqrt{2}}{2}$$

qui est la valeur pour laquelle $\lambda(1-8\lambda) - 4\kappa\lambda^2$ est maximum. Notons que, pour un tel λ, on obtient la condition suivante

$$\alpha_1(f,x) \leq \frac{1}{16(\kappa+2)}.$$

Dans ce cas N_f envoie la boule de centre x et de rayon $\frac{\lambda}{\gamma_1(f,x)}$ dans elle même. De plus N_f est lipschitzienne de constante de Lipschitz

$$\begin{aligned}\sup_{\|x-z\|\gamma_1(f,x)\leq\lambda} \|DN_f(z)\| &\leq 4\frac{1-\lambda}{\psi(\lambda)^2}\left(\alpha_1(f,x) + \frac{\lambda^2}{1-\lambda} + \kappa\lambda\right) \\ &\leq 4\frac{1-\lambda}{\psi(\lambda)^2}\left(\frac{1}{16(\kappa+2)} + \frac{\lambda^2}{1-\lambda} + \kappa\lambda\right) = \Lambda.\end{aligned}$$

Un calcul élémentaire montre que

$$\Lambda \leq \frac{1}{2\psi(\lambda)^2} = \frac{1}{2\psi\left(\frac{1}{8(\kappa+2)}\right)^2} \leq \frac{1}{2\psi\left(\frac{1}{24}\right)^2} = 0.71403\cdots < 1$$

ce qui prouve que N_f est une contraction. En conséquence N_f possède dans cette boule un unique point fixe ζ. Il vérifie $Df(\zeta)^\dagger f(\zeta) = 0$ et la suite des approximation successives $x_k = N_f^k(x)$ converge vers ζ à la vitesse

$$\|x_k - \zeta\| \leq \Lambda^k \|x_k - \zeta\|. \ \square$$

5.4 Exemples

Les exemples que nous présentons conduisent tous naturellement à des systèmes surdéterminés. Certains d'entre eux sont (ou peuvent être) résolus par la méthode de Newton-Gauss mais ce n'est pas la seule approche possible.

5.4.1 Le calcul de racines multiples de polynômes

Nous considérons ici des polynômes ayant une structure particulière de racines multiples. Elle est donnée par des entiers d, $l_1, \ldots, l_m$ qui vérifient $l_k \geq 1$ et $l_1 + \ldots + l_m = d$. Notons $l = (l_1, \ldots, l_m)$ et $\mathcal{P}_l$ l'ensemble des polynômes qui s'écrivent

$$p(z) = \prod_{k=1}^{m} (z - r_k)^{l_k}$$

pour des nombres complexes $r_1, \ldots, r_m$ convenables. On note aussi $r = (r_1, \ldots, r_m)$. Si l'on développe ce produit on obtient les coefficients de $p(z)$

$$p(z) = z^d + g_1(r) z^{d-1} + \ldots + g_d(r).$$

Supposons maintenant que $p(z)$ soit donné via ses coefficients

$$p(z) = z^d + a_1 z^{d-1} + \ldots + a_d$$

et sa structure de multiplicités : $p(z) \in \mathcal{P}_l$. On trouvera ses racines $r_1, \ldots, r_m$ en résolvant le système

$$g_k(r_1, \ldots, r_m) = a_k, \quad 1 \leq k \leq d.$$

C'est un système de d équations à m inconnues donc surdéterminé. On envisage de résoudre ce système par la méthode de Newton-Gauss. Pour ce faire il faut s'assurer que $Dg(r)$ est injective. On a :

Proposition 176. *Si $r_1, \ldots, r_m$ sont deux à deux distincts alors $Dg(r)$ est injective.*

Preuve La matrice jacobienne J de $g(r)$ est donnée par $Dg(r)_{ij} = \frac{\partial g_i(r)}{\partial r_j}$. Sa $j-$ème colonne a pour entrées les coefficients du polynôme

$$\frac{\partial p(r)}{\partial r_j} = \sum_{i=1}^{d} \frac{\partial g_i(r)}{\partial r_j} z^{d-i} = -l_j (z - r_j)^{l_j - 1} \prod_{k \neq j} (z - r_k)^{l_k}.$$

Soit $c \in \mathbb{C}^m$ et supposons que $Jc = 0$. Nous voulons prouver que $c = 0$. On a

$$\sum_{j=1}^{m} c_j J_j = 0$$

d'où

$$\sum_{j=1}^{m} c_j \frac{\partial p(r)}{\partial r_j} = 0$$

de sorte que

$$\sum_{j=1}^{m} c_j l_j (z - r_j)^{l_j - 1} \prod_{k \neq j} (z - r_k)^{l_k} = \prod_{k=1}^{m} (z - r_k)^{l_k - 1} \sum_{j=1}^{m} c_j l_j \prod_{k \neq j} (z - r_k) = 0.$$

Cette dernière expression est le produit de deux polynômes. Comme $\prod_{k=1}^{m} (z - r_k)^{l_k - 1}$ n'est pas nul on a

$$\sum_{j=1}^{m} c_j l_j \prod_{k \neq j} (z - r_k) = 0.$$

Prenons la valeur de cette expression en $z = r_j$ on obtient

$$c_j l_j \prod_{k \neq j} (r_j - r_k) = 0$$

ce qui prouve que $c_j = 0$ puisque les r_k sont distincts. □

5.4.2 Les triangulations géodésiques

Il s'agit de déterminer les coordonnées de l'ensemble des sommets d'une triangulation à partir de mesures d'angles et de longueurs d'arêtes. Ce problème avait été posé à Gauss par le gouvernement allemand : il fallait recalculer les coordonnées de repères géodésiques à partir d'un ensemble de nouvelles mesures. C'est pour répondre à cette question que Gauss inventa la méthode des moindres carrés. Il linéarisa les équations du problème au voisinage du vecteur des anciennes coordonnées puis il résolut le problème linéaire ainsi obtenu par la méthode des moindres carrés. C'est exactement un pas de la méthode de Newton-Gauss.

Les problèmes de ce type sont, en général, surdéterminés : il contiennent plus d'équations que d'inconnues. Si les mesures de longueur et d'angles

étaient faites avec une précision infinie, un tel problème possèderait une solution exacte mais comme ces mesures sont approchées, une telle solution n'a pas de raison d'exister et l'on considère l'approche au sens des moindres carrés.

Pour simplifier, nous supposons que le pays considéré est plan. Les points $M_i(x_i, y_i)$, $1 \leq i \leq N$, sont les sommets d'une triangulation. Deux types de mesures sont faites : des distances entre certains points

$$(x_i - x_j)^2 + (y_i - y_j)^2 = \delta_{ij}^2$$

et des angles. Si l'on note θ_{ijk} l'angle au sommet M_i du triangle $M_iM_jM_k$, son cosinus est donné par

$$\frac{(x_j - x_i)(x_k - x_i) + (y_j - y_i)(y_k - y_i)}{\sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}\sqrt{(x_i - x_k)^2 + (y_i - y_k)^2}} = \cos\theta_{ijk}.$$

La configuration exacte d'un tel système dépend de la triangulation et des mesures faites. Notons que les équations ci-dessus sont invariantes par les transformations affines qui conservent les distances : translations, rotations, symétries. Pour déterminer l'ensemble de ces points à une symétrie près, il faut fixer l'un des sommets et la droite supportant l'une des arêtes issues de ce sommet.

Un cas particulier important est celui où les seules mesures faites sont la longueur d'un côté de la triangulation, qui sert d'unité de longueur, et les angles des différents triangles. C'est ainsi que Delambre et Méchain ont mesuré la longueur du méridien terrestre entre Dunkerke et Barcelonne (1792-1798). Sur la bande de terrain qui contenait l'arc de méridien à mesurer, ils choisirent, à l'est et à l'ouest de l'arc, un certain nombre de points (des clochers, des sommets de collines) visibles les uns des autres : ces points formaient les sommets d'une triangulation couvrant l'arc. Depuis chaque sommet de ces triangles, ils mesurèrent les angles par des visées vers les autres sommets, ce qui définissait tous les triangles par leurs angles. A partir de la mesure d'un des côtés (d'une longueur de 6000 toises prise du côté de Melun) ils calculèrent de proche en proche les longueurs des autres côtés d'où s'en déduisait celle du méridien.

Cette façon de procéder a l'inconvénient majeur d'inclure à chaque étape de calcul les erreurs commises précédemment. La méthode des moindres carrés, qui aurait permis d'éviter de telles accumulations par un calcul global n'existait pas encore, ni les moyens de calcul nécessaires à sa mise en oeuvre.

5.4.3 Reconstruction de molécules

Le problème de la reconstruction de molécules consiste à trouver la configuration spatiale d'une molécule constituée de N atomes A_i, $1 \leq i \leq N$, à partir de distances entre ces atomes :

$$(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2 = d_{ij}^2$$

où $(i,j) \in I \subset \{1,\ldots,N\} \times \{1,\ldots,N\}$. Les distances d_{ij} sont des données expérimentales et, pour cette raison, le système précédent n'a pas nécessairement de solution. On recherche donc une solution au sens des moindres carrés.

Ces problème sont du type « triangulation géodésique » sans mesures d'angles. Les ordres de grandeur de N varient entre 10 et 10^4. C'est un domaine de recherche actif.

5.4.4 Des octaèdres dont les longueurs des arètes sont données

Considérons le robot parallèle suivant (figure ci-dessous) constitué de deux triangles indéformables ABC (appelé plateforme inférieure) et DEF (la plateforme supérieure) reliés entre-eux par six vérins de longueurs variables : AD, AE, BE, BF, CF, CD. Les articulations entre les vérins et les plateformes sont de type cardan et permettent des rotations suivant les trois axes de coordonnées. Les roboticiens dénomment ce type de robot « plateforme de Stewart » en hommage à son inventeur qui conçut sur ce principe un simulateur de vol pour avions. Les qualités de ces robots sont la précision du positionnement de la plateforme et la rigidité d'un tel assemblage.

Supposons connues les positions des sommets A, B et C de la plateforme inférieure ainsi que les longueurs des six vérins. Le « problème géométrique direct »consiste à déterminer, à partir de ces données, les coordonnées D, E, F des sommets de la plateforme supérieure. Ce problème découle de la difficulté à piloter un tel robot. Il s'agit d'amener la plateforme supérieure d'une position de départ I à une position d'arrivée II. La position d'arrivée II étant connue, il est facile d'en déduire les longueurs des vérins. La résolution du problème géométrique direct permet, au cours du déplacement de la plateforme supérieure, d'en contrôler sa position.

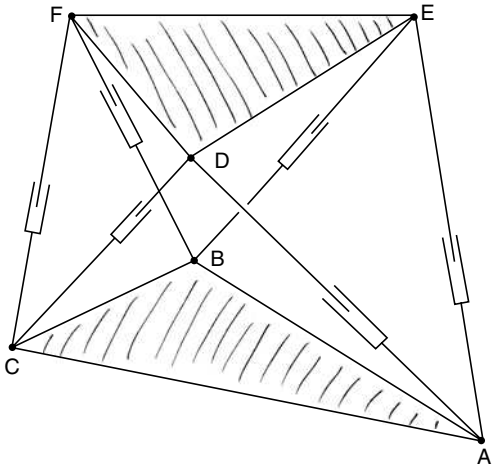


Fig. 5.1. La plateforme de Stewart

Ce problème de robotique peut être formulé différemment. Les deux plateformes et les six vérins constituent un octaèdre ABCDEF dont on connaît les longueurs des douze arêtes ainsi que les coordonnées de trois sommets situés sur une même face (A, B et C). Il s'agit alors de calculer les coordonnées des trois autres sommets (D, E et F).

Suivant un résultat d'unicité dû à Cauchy (1812), ce problème ne possède qu'au plus une solution convexe, à une symétrie près par rapport au plan ABC. D'autre part Bricard (1887) a étudié et caractérisé les octaèdres qui peuvent être déformés sans déformer leurs faces. Ces octaèdres articulés ne sont bien sûr pas convexes. Leur étude a été reprise par Lebesgue qui en a donné une description géométrique. Dans le cas des octaèdres articulés de Bricard, notre problème possède une infinité de solutions.

Le problème proposé conduit à un système de 9 équations polynomiales dont les inconnues sont les 9 coordonnées des sommets D $= (x_1, y_1, z_1)$, E $= (x_2, y_2, z_2)$ et F $= (x_3, y_3, z_3)$. Ces équations sont données par les carrés des distances suivantes :

$$\begin{aligned}
(x_1 - x_A)^2 + (y_1 - y_A)^2 + (z_1 - z_A)^2 &= d(A, D)^2, \\
(x_2 - x_A)^2 + (y_2 - y_A)^2 + (z_2 - z_A)^2 &= d(A, E)^2, \\
(x_2 - x_B)^2 + (y_2 - y_B)^2 + (z_2 - z_B)^2 &= d(B, E)^2, \\
(x_3 - x_B)^2 + (y_3 - y_B)^2 + (z_3 - z_B)^2 &= d(B, F)^2, \\
(x_3 - x_C)^2 + (y_3 - y_C)^2 + (z_3 - z_C)^2 &= d(C, F)^2, \\
(x_1 - x_C)^2 + (y_1 - y_C)^2 + (z_1 - z_C)^2 &= d(C, D)^2, \\
(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2 &= d(D, E)^2, \\
(x_2 - x_3)^2 + (y_2 - y_3)^2 + (z_2 - z_3)^2 &= d(E, F)^2, \\
(x_3 - x_1)^2 + (y_3 - y_1)^2 + (z_3 - z_1)^2 &= d(F, E)^2.
\end{aligned}$$

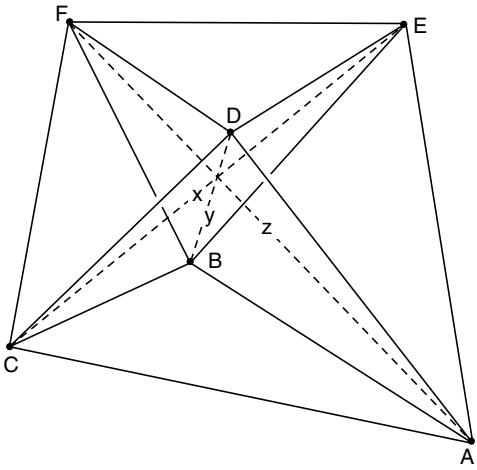


Fig. 5.2. L'octaèdre ABCDEF.

Une autre approche à ce problème consiste à introduire comme inconnues les carrés des longueurs des diagonales de l'octaèdre (CE, BD et AF) :

$$x = d(C,E)^2, \quad y = d(B,D)^2, \quad z = d(A,F)^2.$$

Une fois connues ces quantités, il est facile de calculer les coordonnées de D, E et F. Par exemple, D est l'un des deux points d'intersection des sphères centrées en A, B et C et de rayons respectifs $d(A,D)$, $\sqrt{y}$ et $d(C,D)$.

Nous allons voir que les quantités x, y et z sont données par un système de 6 équations à 3 inconnues. Pour ce faire, nous utilisons une identité, due à Lagrange, entre les distances mutuelles de cinq points de l'espace :

Proposition 177. *Soient M_i, $1 \leq i \leq p$, des points de $\mathbb{R}^n$ avec $p \geq n+2$. Notons $L_{ij} = d(M_i, M_j)^2$ le carré de la distance euclidienne de M_i et M_j. On a :*

$$\det \begin{pmatrix} 0 & 1 & 1 & \dots & 1 \\ 1 & 0 & L_{12} & \dots & L_{1p} \\ 1 & L_{21} & 0 & \dots & L_{2p} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & L_{p1} & L_{p2} & \dots & 0 \end{pmatrix} = 0.$$

Preuve Il suffit de faire le produit des matrices S et T suivantes pour obtenir la matrice ci-dessus. Ces deux matrices sont de dimension $(p+1) \times (p+1)$:

$$S = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 & 0 \dots 0 \\ 1 & \|M_1\|^2 & x_{11} & \dots & x_{1n} & 0 \dots 0 \\ 1 & \|M_2\|^2 & x_{21} & \dots & x_{2n} & 0 \dots 0 \\ \vdots & \vdots & \vdots & \dots & \vdots & \vdots \dots \vdots \\ 1 & \|M_p\|^2 & x_{p1} & \dots & x_{pn} & 0 \dots 0 \end{pmatrix},$$

$$T = \begin{pmatrix} 0 & \|M_1\|^2 & \|M_2\|^2 & \dots & \|M_p\|^2 \\ 0 & 1 & 1 & \dots & 1 \\ 0 & -2x_{11} & -2x_{21} & \dots & -2x_{p1} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 0 & -2x_{1n} & -2x_{2n} & \dots & -2x_{pn} \\ 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 0 & 0 & 0 & \dots & 0 \end{pmatrix}$$

où $M_i = (x_{i1}, \dots, x_{in})$ et $\|M_i\|^2 = x_{i1}^2 + \dots + x_{in}^2$. □

Nous sommes à même de décrire le système vérifié par x, y et z :

Proposition 178. *Les carrés x, y, z des longueurs des diagonales CE, BD et AF satisfont les équations suivantes*

$$
\begin{aligned}
E_1(x,y) &= \det(A,B,C,D,E) = 0,\\
E_2(x,y) &= \det(F,B,C,D,E) = 0,\\
E_3(x,z) &= \det(D,A,C,F,E) = 0,\\
E_4(x,z) &= \det(B,A,C,F,E) = 0,\\
E_5(y,z) &= \det(C,A,D,F,B) = 0,\\
E_6(y,z) &= \det(E,A,D,F,B) = 0,
\end{aligned}
$$

où $\det(A,B,C,D,E)$ *est le déterminant de la proposition précédente avec* $n=3$, $p=5$, $M_1=A$, $M_2=B$, $M_3=C$, $M_4=D$ *et* $M_5=E$ *et avec des notations similaires pour les cinq autres équations. Chacune de ces équations est du type*

$$E(u,v) = u^2v^2 + c_2u^2v + c_3uv^2 + c_4u^2 + c_5v^2 + c_6uv + c_7u + c_8v + c_9.$$

Preuve Il suffit d'appliquer la proposition précédente en prenant à chaque fois cinq des six somments de l'octaèdre. La forme des équations $E(u,v)$ se trouve en développant le déterminant correspondant. □

L'approche du problème fondée sur les identités de Lagrange permet de se ramener à des systèmes qui se résolvent aisément (systèmes de dimension 2×2 pour les longueurs des diagonales, intersections de sphères pour les coordonnées des sommets) contrairement à l'approche spontanée qui conduit à un système 9×9 difficile à résoudre.

Il faut aussi noter le « petit miracle » ci-dessus : le système surdéterminé de dimension 6×3 se scinde en trois systèmes 2×2.

5.4.5 Moindres carrés totaux

Notons $\mathcal{M}_{m,n}$ l'espace des matrices réelles, $m\times n$; cet espace est équipé du produit scalaire

$$\langle U,V\rangle = \text{trace } (V^TU) = \sum_{i,j} U_{i,j}V_{i,j}$$

et de la norme associée. Elle est notée

$$\|U\|_F^2 = \text{trace } (U^TU),$$

c'est la norme de Frobenius.

Le problème que nous considérons ici est le suivant : étant donné deux matrices réelles $A\in\mathcal{M}_{m,n}$ et $B\in\mathcal{M}_{m,p}$ trouver une matrice $X\in\mathcal{M}_{n,p}$ telle que

$$AX = B.$$

Un premier résultat précise sous quelles conditions une telle matrice existe :

Proposition 179. *Une condition nécessaire et suffisante pour qu'il existe* X *avec* $AX=B$ *est que im* $B\subset$ *im* A. *Dans ce cas,* $X_0 = A^\dagger B$ *est une solution de ce problème.*

Preuve Si $AX = B$ et si $v = Bu$ alors $v = A(Xu)$ ce qui prouve que im $B \subset$ im A. Réciproquement, si im $B \subset$ im A, par le Théorème 120-2, on a

$$AX_0 = AA^\dagger B = \Pi_{\text{im } A} B = B. \ \square$$

Qu'une solution de $AX = B$ existe ou non, on peut toujours considérer le problème des moindres carrés suivant :

$$\min_{X \in \mathcal{M}_{n,p}} \|AX - B\|_F^2 .$$

Par définition même de l'inverse généralisé, la solution de norme minimale de ce problème est

$$X = \mathcal{L}_A^\dagger(B)$$

où $\mathcal{L}_A^\dagger$ est l'inverse généralisé de l'opérateur linéaire

$$\mathcal{L}_A : \mathcal{M}_{n,p} \to \mathcal{M}_{m,p},\ \mathcal{L}_A(X) = AX.$$

Cet inverse généralisé est donné par

$$\mathcal{L}_A^\dagger : \mathcal{M}_{m,p} \to \mathcal{M}_{n,p},\ \mathcal{L}_A^\dagger(Y) = A^\dagger Y,$$

c'est à dire $\mathcal{L}_A^\dagger = \mathcal{L}_{A^\dagger}$. Pour établir cette égalité on peut, par exemple, utiliser le Théorème 120-5. On vient donc de prouver que

Proposition 180. *La solution de norme (de Frobenius) minimale du problème*

$$\min_{X \in \mathcal{M}_{n,p}} \|AX - B\|_F^2$$

est $X_0 = A^\dagger B$.

Une instance de ce problème est la régression linéaire. Etant donné une famille de mesures (x_i, y_i), $1 \leq i \leq N$, dont on pense qu'elles suivent une relation linéaire

$$y = ax + b,$$

on cherche à préciser cette relation. La régression linéaire consiste à minimiser la quantité

$$\sum_{i=1}^{N} (ax_i + b - y_i)^2.$$

On retrouve le problème précédent en prenant

$$A = \begin{pmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_N & 1 \end{pmatrix}, \quad X = \begin{pmatrix} a \\ b \end{pmatrix}, \quad B = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}.$$

Si au moins deux des x_i sont distincts, on obtient pour unique solution

$$a = \frac{\Gamma(x,y)}{\sigma^2(x)},\ b = \bar{y} - a\bar{x},$$

avec

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i,\ \bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i,\ \sigma^2(x) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2,\ \Gamma(x,y)$$
$$= \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}).$$

Revenons à notre problème initial. La méthode des moindres carrés totaux propose une autre approche. Elle consiste à minimiser la quantité

$$\|P - A\|_F^2 + \|Q - B\|_F^2$$

sous la contrainte $Q = PX$. Soyons plus précis : notons

$$\mathcal{V} = \{(P,Q,X) \in \mathcal{M}_{m,n} \times \mathcal{M}_{m,p} \times \mathcal{M}_{n,p}\ :\ Q = PX\}.$$

Le problème des moindres carrés totaux est

$$\min_{(P,Q,X)\in\mathcal{V}} \|P - A\|_F^2 + \|Q - B\|_F^2 .$$

L'exemple de la régression linéaire devient, sous cette nouvelle approche,

$$\min \sum_{i=1}^{N} (p_{i1} - x_i)^2 + (p_{i2} - 1)^2 + (ap_{i1} + bp_{i2} - y_i)^2$$

où le minimum est pris pour

$$P = \begin{pmatrix} p_{11} & p_{12} \\ \vdots & \vdots \\ p_{N1} & p_{N2} \end{pmatrix} \text{ et } X = \begin{pmatrix} a \\ b \end{pmatrix}.$$

L'expression qui figure dans cette somme est le carré de la distance des points $(x_i, 1, y_i)$ et $(p_{i1}, p_{i2}, ap_{i1} + bp_{i2})$. Cette distance est minimum lorsque l'on prend la projection orthogonale de $(x_i, 1, y_i)$ sur le plan $\mathcal{P}$ constitué des vecteurs $(u, v, au + bv)$. Cette projection est donnée par

$$p_{i1} = x_i - \frac{a(ax_i + b - y_i)}{1 + a^2 + b^2},$$
$$p_{i2} = 1 - \frac{b(ax_i + b - y_i)}{1 + a^2 + b^2}$$

et le carré de la distance vaut

$$\frac{(ax_i + b - y_i)^2}{1 + a^2 + b^2}.$$

Ainsi, le problème des moindres carrés totaux revient à minimiser

$$\min_{(a,b)\in\mathbb{R}^2} \sum_{i=1}^{N} \frac{(ax_i + b - y_i)^2}{1 + a^2 + b^2}$$

qui est la somme des carrés des distances des points $(x_i, 1, y_i)$ au plan $\mathcal{P}$.

On caractérise les solutions du problème des moindres carrés totaux par une équation de Riccati algébrique comme nous en avons rencontré au Chap. 3.

Théorème 181. *Toute solution du problème*

$$\min_{(P,Q,X)\in\mathcal{V}} \|P - A\|_F^2 + \|Q - B\|_F^2$$

vérifie

$$\begin{aligned} Q &= PX, \\ P &= (A + BX^T)(I_n + XX^T)^{-1}, \\ 0 &= X(B^TA)X + (A^TA)X - X(B^TB) - A^TB. \end{aligned}$$

Preuve Démontrons tout d'abord que $\mathcal{V}$ est une sous-variété différentiable contenue dans $\mathcal{M}_{m,n} \times \mathcal{M}_{m,p} \times \mathcal{M}_{n,p}$. En effet c'est l'ensemble des zéros de la fonction C^∞ suivante

$$\mathcal{F} : \mathcal{M}_{m,n} \times \mathcal{M}_{m,p} \times \mathcal{M}_{n,p} \to \mathcal{M}_{m,p}, \quad \mathcal{F}(P,Q,X) = PX - Q.$$

La dérivée de cette dernière est donnée par

$$D\mathcal{F}(P,Q,X) : \mathcal{M}_{m,n} \times \mathcal{M}_{m,p} \times \mathcal{M}_{n,p} \to \mathcal{M}_{m,p},$$

$$D\mathcal{F}(P,Q,X)(\dot{P},\dot{Q},\dot{X}) = \dot{P}X + P\dot{X} - \dot{Q},$$

elle est toujours surjective ($\dot{Q}$ suffit). Que $\mathcal{V}$ soit une sous-variété résulte de l'exemple 3 donné dans l'appendice « Sous-variétés différentiables ». Son espace tangent est (voir ce même exemple)

$$\begin{aligned} T_{(P,Q,X)}\mathcal{V} &= \ker D\mathcal{F}(P,Q,X) \\ &= \left\{(\dot{P},\dot{Q},\dot{X}) \in \mathcal{M}_{m,n} \times \mathcal{M}_{m,p} \times \mathcal{M}_{n,p} \ : \ \dot{Q} = \dot{P}X + P\dot{X}\right\}. \end{aligned}$$

Nous savons (même appendice, Proposition 196) que si (P,Q,X) réalise le minimum de $\|f(P,Q,X)\|^2$ alors

$$Df(P,Q,X)^* f(P,Q,X) \in \left(T_{(P,Q,X)}\mathcal{V}\right)^\perp.$$

On a ici

$$f : \mathcal{M}_{m,n} \times \mathcal{M}_{m,p} \times \mathcal{M}_{n,p} \to \mathcal{M}_{m,n} \times \mathcal{M}_{m,p}, \quad f(P,Q,X) = (P - A, Q - B)$$

de sorte que

$$Df(P,Q,X)(\dot{P},\dot{Q},\dot{X}) = (\dot{P},\dot{Q})$$

et que

$$Df(P,Q,X)^*(U,V) = (U,V,0).$$

Ainsi

$$Df(P,Q,X)^* f(P,Q,X) = (P - A, Q - B, 0)$$

et la condition d'optimalité $Df(P,Q,X)^* f(P,Q,X) \in \left(T_{(P,Q,X)}\mathcal{V}\right)^\perp$ devient

$$\left\langle (P - A, Q - B, 0)\,, \left(\dot{P}, \dot{P}X + P\dot{X}, \dot{X}\right) \right\rangle = 0$$

pour tout $\dot{P} \in \mathcal{M}_{m,n}$ et $\dot{X} \in \mathcal{M}_{n,p}$. Cette dernière condition est équivalente au système suivant :

$$\begin{aligned} &P - A + (Q - B)X^T = 0,\\ &P^T(Q - B) = 0,\\ &Q = PX. \end{aligned}$$

La première et la dernière équation donnent

$$P - A + (PX - B)X^T = P(I_n + XX^T) - (A + BX^T) = 0$$

de sorte que

$$P = (A + BX^T)(I_n + XX^T)^{-1}.$$

Comme $P^T PX = P^T B$, en remplaçant P par la valeur trouvée ci-dessus, on obtient

$$(A^T + XB^T)(A + BX^T)(I_n + XX^T)^{-1}X = (A^T + XB^T)B.$$

Notons que

$$(I_n + XX^T)^{-1}X = X(I_n + XX^T)^{-1}$$

de sorte que

$$X(B^T A)X + (A^T A)X - X(B^T B) - A^T B = 0$$

qui est l'équation annoncée. □

Un cas particulier important est obtenu lorsque X et B sont des vecteurs. Dans ce cas X est donné par un problème de valeurs propres.

Théorème 182. *Lorsque $A \in \mathcal{M}_{m,n}$, $B \in \mathbb{R}^m$ et $X \in \mathbb{R}^n$, toute solution du problème des moindres carrés totaux*

$$\min_{(P,Q,X)\in\mathcal{V}} \|P - A\|_F^2 + \|Q - B\|_F^2$$

vérifie $P = (A+BX^T)(I_n+XX^T)^{-1}$, $Q = PX$ *où* $\begin{pmatrix} X \\ 1 \end{pmatrix}$ *est un vecteur propre de la matrice* $\mathcal{A} = \begin{pmatrix} A^TA & -A^TB \\ -B^TA & B^TB \end{pmatrix}$. *Si* λ *est la valeur propre correspondante alors*

$$\lambda = \|P - A\|_F^2 + \|Q - B\|_F^2 .$$

Preuve La première assertion est facile et résulte de l'équivalence suivante pour l'équation de Riccati du Théorème 181 :

$$A^TAX - A^TB = -XB^TAX + XB^TB$$

si et seulement si

$$\begin{aligned} A^TAX - A^TB &= X\lambda \\ -B^TAX + B^TB &= \lambda \end{aligned}$$

c'est à dire

$$\begin{pmatrix} A^TA & -A^TB \\ -B^TA & B^TB \end{pmatrix} \begin{pmatrix} X \\ 1 \end{pmatrix} = \begin{pmatrix} X \\ 1 \end{pmatrix} \lambda.$$

Passons à la seconde assertion. Soient P et X vérifiant les équations du Théorème 181 et soit λ la valeur propre correspondante de $\mathcal{A}$. Notons $\nu = \|P - A\|_F^2 + \|PX - B\|_F^2$. Nous allons voir que $\nu = \lambda$. On a

$$\begin{aligned} \nu &= \text{trace } ((A-P)^T(A-P)) + \text{trace } ((B-PX)^T(B-PX)) \\ &= \text{trace } (A^TA + P^TP - 2A^TP) + \text{trace } (B^TB + X^TP^TPX - 2B^TPX). \end{aligned}$$

Rappelons que trace (UV) = trace (VU) quelles que soient les matrices U $r \times s$ et V $s \times r$ de sorte que

$$\text{trace } (X^TP^TPX) = \text{trace } (P^TPXX^T)$$

et que

$$\text{trace } (B^TPX) = \text{trace } (XB^TP).$$

Ceci donne

$$\begin{aligned} \nu = &\text{trace } (A^TA) + \text{trace } (B^TB) + \text{trace } (P^TP(I + XX^T)) \\ &-2\text{trace } ((A^T + XB^T)P). \end{aligned}$$

Comme $P(I + XX^T) = A + BX^T$ on obtient

$$\begin{aligned} \nu &= \text{trace } (A^TA) + \text{trace } (B^TB) + \text{trace } (P^T(A + BX^T)) \\ &\quad -2\text{trace } ((A^T + XB^T)P) \\ &= \text{trace } (A^TA) + \text{trace } (B^TB) - \text{trace } ((A^T + XB^T)P) \end{aligned}$$

en utilisant l'égalité de la trace d'une matrice et celle de sa transposée. En remplaçant P par son expression en X on a

$\nu = \text{trace } (A^TA)+\text{trace } (B^TB)-\text{trace } ((A^T+XB^T)(A+BX^T)(I_n+XX^T)^{-1})$.

L'équation de Riccati donne

$$(A^T + XB^T)B = (A^T + XB^T)AX$$

d'où

$$(A^T + BX^T)(A + BX^T)(I_n + XX^T)^{-1} = (A^T + XB^T)A$$

et donc

$$\begin{aligned}\nu &= \text{trace } (A^TA) + \text{trace } (B^TB) - \text{trace } ((A^T + XB^T)A)\\ &= \text{trace } (B^TB) - \text{trace } (XB^TA)\\ &= \text{trace } (B^TB) - \text{trace } (B^TAX) = \text{trace } (B^TB - B^TAX) = \lambda. \ \square\end{aligned}$$

Les deux théorèmes précédent mis bout à bout donnent le corollaire suivant :

Corollaire 183. *Si tous les vecteurs propres de $\mathcal{A}$ sont du type $\begin{pmatrix} X \\ x_{n+1} \end{pmatrix}$ avec $x_{n+1} \neq 0$, alors $\min_{(P,Q,X)\in\mathcal{V}} \|P - A\|_F^2 + \|Q - B\|_F^2$ est la plus petite des valeurs propres de $\mathcal{A}$.*

L'hypothèse $x_{n+1} \neq 0$ pour tout vecteur propre de $\mathcal{A}$ n'est pas très contraignante. Lorsque le rang de A est égal à n, cette condition est satisfaite pour tout $B \in \mathbb{R}^m$ tel que $\langle AX, B\rangle \neq 0$ pour tout vecteur propre X de A^TA c'est à dire pour tout B pris en dehors de la réunion de n hyperplans de $\mathbb{R}^m$. En effet, $\begin{pmatrix} X \\ 0 \end{pmatrix}$ est un vecteur propre de $\mathcal{A}$ si et seulement si X est un vecteur propre de A^TA et si $-B^TAX = 0$. Pour un tel vecteur propre on a $AX \neq 0$ (A est de rang n) de sorte que la condition $B^TAX = \langle AX, B\rangle = 0$ définit un hyperplan.

5.4.6 Moindres carrés avec contraintes

Une situation assez fréquente est celle d'un système d'équations (prenons le linéaire)

$$Ax = b$$

où l'inconnue x est astreinte à appartenir à un certain ensemble (prenons la sphère unité)

$$\begin{aligned}Ax &= b,\\ \|x\|^2 &= 1.\end{aligned}$$

Lorsqu'un tel système est surdéterminé, c'est à dire si $x \in \mathbb{R}^n$, $b \in \mathbb{R}^m$ et $m > n-1$ (la dimension de la sphère unité) il est sain d'envisager une approche

de type « moindres carrés contraints »c'est à dire de résoudre le problème de minimisation

$$\min_{\|x\|^2=1} \|Ax-b\|_F^2 .$$

Pour caractériser les solutions d'un tel problème nous utilisons la Proposition 196 de l'Appendice « Sous-variétés différentiables ». L'Exemple 5 du même appendice montre que

$$\mathbb{S}^{n-1} = \left\{ x \in \mathbb{R}^n \ : \ \|x\|^2 = 1 \right\}$$

est une sous-variété dans $\mathbb{R}^n$ et son espace tangent est

$$T_x\mathbb{S}^{n-1} = \{ u \in \mathbb{R}^n \ : \ \langle x, u \rangle = 0 \} .$$

Par la proposition précitée, si x est solution de ce problème d'optimisation on a

$$A^T(Ax-b) \in \left(T_x\mathbb{S}^{n-1}\right)^\perp = \{ \lambda x \ : \ \lambda \in \mathbb{R} \}$$

autrement dit, il existe $\lambda \in \mathbb{R}$ tel que

$$\begin{aligned} &A^T(Ax-b) = \lambda x, \\ &\|x\|^2 = 1 \end{aligned}$$

ce qui constitue un système de $n+1$ équations et $n+1$ inconnues. En faisant le produit scalaire de la première équation par x on obtient

$$\lambda = \left\langle A^T(Ax-b), x \right\rangle$$

et x est donné par

$$\begin{aligned} &A^T(Ax-b) = \left\langle A^T(Ax-b), x \right\rangle x, \\ &\|x\|^2 = 1. \end{aligned}$$

6

Appendices

6.1 Calcul différentiel sur les espaces de Banach

Tous les espaces considérés ici sont des espaces de Banach.

6.1.1 Dérivée d'une application

Une application définie sur un ouvert $f : U \subset \mathbb{E} \to \mathbb{F}$ est différentiable en un point $a \in U$ s'il existe une application linéaire et continue $Df(a) : \mathbb{E} \to \mathbb{F}$ telle que,

$$\lim_{u \to 0} \frac{f(a+u) - f(a) - Df(a)u}{\|u\|} = 0.$$

$Df(a)$ est la dérivée de f en a. Cette application est unique et la fonction f est continue en a. Pour tout $u \in \mathbb{E}$ et $\lambda \in \mathbb{R}$ on a

$$\lim_{\lambda \to 0} \frac{f(a+\lambda u) - f(a)}{\lambda} = Df(a)u.$$

On dit que f est dérivable si elle l'est en tout point de U. Si c'est le cas et si $Df : U \subset \mathbb{E} \to \mathbb{F}$ est continue, on dit que f est de classe C^1.

Théorème 184. *Théorème de dérivation des fonctions composées. Soient U un ouvert de $\mathbb{E}$ et V un ouvert de $\mathbb{F}$. Si $f : U \subset \mathbb{E} \to V \subset \mathbb{F}$ est dérivable en $a \in U$ et si $g : V \subset \mathbb{F} \to \mathbb{G}$ est dérivable en $f(a)$ alors $g \circ f : U \subset \mathbb{E} \to \mathbb{G}$ est dérivable en a et*

$$D(g \circ f)(a) = Dg(f(a)) \circ Df(a).$$

Théorème 185. *Théorème d'inversion locale. Soit U un ouvert de $\mathbb{E}$. Si $f : U \subset \mathbb{E} \to \mathbb{F}$ est de classe C^1 et si $Df(a)$ est bijective alors f est une bijection d'un voisinage ouvert V de a contenu dans U sur un voisinage ouvert W de $f(a)$ dans $\mathbb{F}$. De plus $f^{-1} : W \to V$ est de classe C^1 et*

$$Df^{-1}(f(x)) = (Df(x))^{-1}.$$

6.1.2 Dérivée seconde

Notons $\mathcal{L}(\mathbb{E},\mathbb{F})$ l'espace de Banach des applications linéaires et continues de $\mathbb{E}$ dans $\mathbb{F}$. Lorsque $f : U \subset \mathbb{E} \to \mathbb{F}$ est de classe C^1 et que sa dérivée

$$Df : U \subset \mathbb{E} \to \mathcal{L}(\mathbb{E},\mathbb{F})$$

est dérivable en $a \in U$ on dit que f est deux fois dérivable en a. On note $D^2 f(a)$ la dérivée de Df en a et on l'appelle la dérivée seconde de f en a. Cette dérivée seconde est une application linéaire

$$D^2 f(a) : \mathbb{E} \to \mathcal{L}(\mathbb{E},\mathbb{F})$$

que l'on identifie à l'application bilinéaire suivante

$$(u,v) \in \mathbb{E} \times \mathbb{E} \to (D^2 f(a)u)(v) \in \mathbb{F}.$$

Cette application bilinéaire est symétrique c'est à dire que

$$(D^2 f(a)u)(v) = (D^2 f(a)v)(u)$$

pour tout $u, v \in \mathbb{E}$. On note alors

$$D^2 f(a)(u,v) = (D^2 f(a)u)(v)$$

et lorsque $u = v$

$$D^2 f(a)u^2 = D^2 f(a)(u,u).$$

Bien sûr l'expression u^2 n'a aucun sens en dehors de ce contexte sauf lorsque u est un scalaire.

6.1.3 Dérivée d'ordre p

Lorsque f est de classe C^{p-1} sur U, la dérivée d'ordre p de f est définie inductivement par

$$D^p f(a) = D(D^{p-1} f)(a).$$

Elle s'identifie à une application $p-$multilinéaire symétrique

$$D^p f(a) : \mathbb{E} \times \ldots \times \mathbb{E} \to \mathbb{F}$$

et l'on note, comme plus haut pour $p = 2$,

$$D^p f(a)u^p = D^p f(a)(u,\ldots,u).$$

6.1.4 Norme de la dérivée $p-$ième d'une application vectorielle

La norme de la dérivée $p-$ième d'une application de classe C^p et plus généralement la norme d'une application multilinéaire symétrique continue

$$M : \mathbb{E} \times \ldots \times \mathbb{E} \to \mathbb{F}$$

est définie par

$$\|M\| = \sup \frac{\|M(u^1, \ldots, u^p)\|}{\|u^1\| \ldots \|u^p\|}$$

où le supremum est pris pour des vecteurs non nuls $u^1, \ldots, u^p \in \mathbb{E}$. Par cette définition la norme de M est la plus petite constante $K \geq 0$ pour laquelle

$$\|M(u^1, \ldots, u^p)\| \leq K\|u^1\| \ldots \|u^p\|$$

quels que soient les vecteurs $u^1, \ldots, u^p \in \mathbb{E}$.

6.1.5 Inégalité des accroissements finis

Soit C une partie convexe ouverte de $\mathbb{E}$ et soit $f : C \to \mathbb{F}$ une application différentiable. S'il existe une constante $\lambda \geq 0$ telle que $\|Df(x)\| \leq \lambda$ pour tout $x \in C$ alors

$$\|f(x) - f(y)\| \leq \lambda\|x - y\|$$

pour tout x et $y \in C$.

6.1.6 La formule de Taylor : reste de Lagrange

Soit $f : U \subset \mathbb{E} \to \mathbb{R}$ une application de classe C^p à valeurs scalaires. Soient a et $b \in U$ tels que

$$[a, b] = \{a + t(b - a) : 0 \leq t \leq 1\} \subset U.$$

Il existe $z \in]a, b[$ tel que

$$f(b) = f(a) + \sum_{k=1}^{p-1} \frac{D^k f(a)}{k!}(b - a)^k + \frac{D^p f(z)}{p!}(b - a)^p.$$

La démonstration de cette formule consiste à restreindre f au segment $[a, b]$ ce qui en fait une application $t \in [0, 1] \to f(a + t(b - a)) \in \mathbb{R}$ à laquelle on applique la formule de Taylor des lycées et collèges. Cette façon de faire ne se transpose pas à des fonctions à valeurs vectorielles : à chaque application coordonnée f_i correspond un $z_i \in]a, b[$ mais ce ne sont pas nécessairement tous les mêmes. Pour contourner cette difficulté on va utiliser une autre formulation du reste qui se vectorialise bien.

6.1.7 La formule de Taylor : reste intégral

Soit $f : U \subset \mathbb{E} \to \mathbb{F}$ une application de classe C^p définie sur un ouvert $U \subset \mathbb{E}$. Soient a et $b \in U$ tels que

$$[a, b] = \{a + t(b - a) : 0 \leq t \leq 1\} \subset U.$$

Alors

$$f(b) = f(a) + \sum_{k=1}^{p-1} \frac{D^k f(a)}{k!}(b-a)^k + \int_0^1 \frac{(1-t)^{p-1}}{(p-1)!} D^p f(a + t(b-a))(b-a)^p dt.$$

6.2 Calcul différentiel sur les espaces de Hilbert

Il n'y a pas grand chose à ajouter à ce qui vient d'être dit sinon introduire deux nouveaux concepts : le gradient et le hessien. Soit $f : U \subset \mathbb{E} \to \mathbb{R}$ une application à valeurs scalaires dérivable en $a \in U$. Puisque $Df(a) : \mathbb{E} \to \mathbb{R}$ est linéaire et continue, par le théorème de représentation de Rietz, il existe un unique vecteur $\nabla f(a) \in \mathbb{E}$ tel que

$$Df(a)x = \langle x, \nabla f(a) \rangle$$

pour tout $x \in \mathbb{E}$. $\nabla f(a)$ s'appelle le gradient de f en a et se note aussi grad $f(a)$.

Lorsque f est de classe C^1 sur U et si elle est deux fois dérivable en a alors, pour tout $u \in \mathbb{E}$, $D^2 f(a)u$ est un élément de $\mathcal{L}(\mathbb{E}, \mathbb{R})$ que l'on identifie, toujours par le théorème de représentation de Rietz, à un élément de $\mathbb{E}$ que l'on note Hess $(a)u$. On a donc

$$D^2 f(a)(u, v) = \langle v, \text{Hess } (a)u \rangle$$

pour tout u et $v \in \mathbb{E}$. Hess (a) s'appelle la hessienne de f en a,

$$\text{Hess } (a) : \mathbb{E} \to \mathbb{E}$$

c'est une application linéaire symétrique et continue.

6.3 Calcul différentiel sur les espaces euclidiens

Comment exprimer ce qui vient d'être dit en termes de coordonnées ? C'est l'objet de ce nouveau paragraphe.

6.3.1 La structure euclidienne

L'espace $\mathbb{R}^n$ est muni de sa structure euclidienne habituelle donnée par le produit scalaire $\langle x, y\rangle = \sum_{i=1}^n x_i y_i$ et la norme $\|x\|^2 = \sum_{i=1}^n x_i^2$ lorsque x et y sont donnés par leurs coordonnées dans la base canonique :

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \quad \text{et} \quad y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}.$$

Le produit scalaire peut aussi être vu comme un produit de matrices : $\langle x, y\rangle = x^T y$ où A^T est la transposée de la matrice A.

6.3.2 Dérivée d'une application scalaire

Soit $f : \mathbb{R}^n \to \mathbb{R}$ qui est différentiable en un point a. La dérivée de f en a s'exprime à l'aide des dérivées partielles : elle est égale à

$$Df(a)u = \sum_{i=1}^n \frac{\partial f}{\partial x_i}(a) u_i = \langle \nabla f(a), u\rangle$$

et où

$$\nabla f(a) = \begin{pmatrix} \frac{\partial f}{\partial x_1}(a) \\ \vdots \\ \frac{\partial f}{\partial x_n}(a) \end{pmatrix}.$$

Le vecteur $\nabla f(a)$ est le gradient de f en a.

6.3.3 Dérivée d'une application vectorielle

Une application $f : \mathbb{R}^n \to \mathbb{R}^m$ sera notée

$$f = \begin{pmatrix} f_1 \\ \vdots \\ f_m \end{pmatrix}$$

où les f_i sont des applications à valeurs scalaires. La dérivée de f en a est une application linéaire $Df(a) : \mathbb{R}^n \to \mathbb{R}^m$ et, pour $u \in \mathbb{R}^n$, $Df(a)u$ est le vecteur dont les coordonnées sont $Df_i(a)u$. Ainsi

$$\begin{aligned} Df(a)u &= \begin{pmatrix} Df_1(a)u \\ \vdots \\ Df_m(a)u \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n \frac{\partial f_1}{\partial x_i}(a) u_i \\ \vdots \\ \sum_{i=1}^n \frac{\partial f_m}{\partial x_i}(a) u_i \end{pmatrix} \\ &= \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{pmatrix} \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix} = Jf(a)u. \end{aligned}$$

La matrice $Jf(a)$, de taille $m \times n$, est appelée matrice jacobienne de f en a. On peut noter que ses lignes sont les transposées des gradients $\nabla f_j(a)$.

6.3.4 Dérivée $p-$ième d'une application scalaire

Une application scalaire $f : \Omega \subset \mathbb{R}^n \to \mathbb{R}$ est de classe C^p sur l'ensemble ouvert Ω lorsqu'elle possède sur cet ensemble des dérivées partielles jusqu'à l'ordre p qui sont continues. Sous cette hypothèse la dérivée $p-$ième de f en a est la forme $p-$linéaire symétrique suivante

$$D^p f(a) : \mathbb{R}^n \times \ldots \times \mathbb{R}^n \to \mathbb{R},$$

$$D^p f(a)(u^1, \ldots, u^p) = \sum_{i_1=1}^{n} \ldots \sum_{i_p=1}^{n} \frac{\partial^p f}{\partial x_{i_1} \ldots \partial x_{i_p}}(a) u^1_{i_1} \ldots u^p_{i_p},$$

où $u^j_1, \ldots u^j_n$ sont les coordonnées du vecteur u^j dans la base canonique de $\mathbb{R}^n$.

6.3.5 Dérivée $p-$ième d'une application vectorielle

Une application vectorielle $f : \Omega \subset \mathbb{R}^n \to \mathbb{R}^m$ est de classe C^p lorsque c'est le cas pour ses m applications coordonnées $f_1, \ldots f_m$. Dans ce cas la dérivée $p-$ième de f en a est l'application $p-$linéaire symétrique suivante

$$D^p f(a) : \mathbb{R}^n \times \ldots \times \mathbb{R}^n \to \mathbb{R}^m, \;\; D^p f(a)(u^1, \ldots, u^p) = \begin{pmatrix} D^p f_1(a)(u^1, \ldots, u^p) \\ \vdots \\ D^p f_m(a)(u^1, \ldots, u^p) \end{pmatrix}.$$

6.3.6 Dérivées secondes : cas scalaire

Soit $f : \mathbb{R}^n \to \mathbb{R}$ qui soit de classe C^2. En accord avec ce qui vient d'être dit la dérivée seconde de f en a est donnée par

$$D^2 f(a)(u, v) = \sum_{i,j=1}^{n} \frac{\partial^2 f}{\partial x_i \partial x_j}(a) u_i v_j = \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix}^T \text{Hess } (a) \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix}.$$

La matrice Hess (a) est appelée matrice hessienne de f en a, c'est la matrice symétrique $n \times n$ dont les entrées sont les dérivées partielles secondes $\frac{\partial^2 f}{\partial x_i \partial x_j}(a)$.

La norme de cette dérivée seconde est égale à la norme d'opérateur de la matrice hessienne c'est à dire égale à son rayon spectral puisque c'est une matrice réelle symétrique :

$$\|D^2 f(a)\| = \rho(Hf(a)).$$

6.3.7 Dérivées secondes : cas vectoriel

Soit $f : \mathbb{R}^n \to \mathbb{R}^m$ qui soit de classe C^2. La dérivée seconde de $D^2 f(a)$ est donnée par les dérivées secondes des applications coordonnées $D^2 f_j(a)$ que l'on représente par les matrices hessiennes $Hf_j(a)$. La norme de la dérivée seconde vérifie l'inégalité suivante :

$$\|D^2 f(a)\|^2 \leq \sum_{j=1}^{m} \rho(Hf_j(a))^2.$$

6.3.8 Etude d'un exemple : le problème symétrique des valeurs propres

Notons $\mathcal{S}$ l'espace des matrices $n \times n$ symétriques réelles. Une matrice $A \in \mathcal{S}$ possède n valeurs propres réelles auquelles on peut associer des vecteurs propres normalisés. Leur recherche peut être vue comme celle des zéros de l'application

$$V_A : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n \times \mathbb{R}, \quad V_A \begin{pmatrix} x \\ \lambda \end{pmatrix} = \begin{pmatrix} (\lambda I_n - A)x \\ \frac{1}{2}(\|x\|^2 - 1) \end{pmatrix}.$$

C'est une fonction polynomiale de degré 2 en les variables $x_1, \ldots, x_n, \lambda$. La dérivée première de V_A est donnée par

$$DV_A \begin{pmatrix} x \\ \lambda \end{pmatrix} \begin{pmatrix} \dot{x} \\ \dot{\lambda} \end{pmatrix} = \begin{pmatrix} (\lambda I_n - A)\dot{x} + \dot{\lambda}x \\ \langle x, \dot{x} \rangle \end{pmatrix}.$$

La dérivée seconde est égale à

$$D^2 V_A \begin{pmatrix} x \\ \lambda \end{pmatrix} \begin{pmatrix} \dot{x} \\ \dot{\lambda} \end{pmatrix} \begin{pmatrix} \dot{y} \\ \dot{\mu} \end{pmatrix} = \begin{pmatrix} \dot{\mu}\dot{x} + \dot{\lambda}\dot{y} \\ \langle \dot{y}, \dot{x} \rangle \end{pmatrix}.$$

C'est un bon exercice que de montrer que $\|D^2 V_A\| \leq \sqrt{2}$. Les dérivées d'ordre supérieur à 2 sont nulles.

6.4 Fonctions analytiques

Soit $f : U \subset \mathbb{E} \to \mathbb{F}$ définie sur un ouvert $U \subset \mathbb{E}$. Nous dirons que f est de classe C^∞ si elle possède des dérivées de tout ordre en tout point de U. Dans ce cas nous pouvons considérer la série de Taylor en a :

$$\sum_{k=0}^{\infty} \frac{D^k f(a)}{k!}(x-a)^k.$$

Son rayon de convergence est donné par $R = \rho^{-1}$ et

$$\rho = \limsup_{k \to \infty} \left\| \frac{D^k f(a)}{k!} \right\|^{\frac{1}{k}}.$$

Nous dirons que f est analytique lorsque cette série a un rayon de convergence $R > 0$ et que

$$f(x) = \sum_{k=0}^{\infty} \frac{D^k f(a)}{k!}(x-a)^k$$

pour tout x tel que $\|x - a\| < R$.

Notons que la série précédente est absolument convergente et que

$$\|f(x)\| \leq \sum_{k=0}^{\infty} \frac{\|D^k f(a)\|}{k!}\|x-a\|^k.$$

Les règles habituelles de dérivation sous le signe somme sont toujours valides : pour tout $p > 0$ et pour tout x tel que $\|x - a\| < R$ on a

$$D^p f(x) = \sum_{k=0}^{\infty} \frac{D^{p+k} f(a)}{k!}(x-a)^k.$$

Notons qu'ici $D^p f(x)$ et $D^{p+k} f(a)(x-a)^k$ sont des applications $p-$multilinéaires définies sur $\mathbb{E} \times \ldots \times \mathbb{E}$ et à valeurs dans $\mathbb{F}$.

6.5 Sous-variétés différentiables

Définition 186. *On dit qu'une partie V de l'espace de Banach $\mathbb{E}$ est une sous-variété de classe C^r, $1 \leq r \leq \infty$, ou analytique si pour tout point x de V on peut trouver un difféomorphisme ϕ de classe C^r ou analytique d'un voisinage ouvert U de x dans $\mathbb{E}$ sur un ouvert d'un espace de Banach $\mathbb{H} \times \mathbb{K}$ tel que $\phi(U \cap V) = \phi(U) \cap (\mathbb{H} \times \{0\})$.*

Par cette définition une sous-variété apparait localement comme la déformation d'un sous-espace vectoriel par un difféomorphisme. Une autre définition est possible qui décrit un tel ensemble par une équation locale :

Définition 187. *Soit V une partie de l'espace de Banach $\mathbb{E}$ et soit $\mathbb{F}$ un autre espace de Banach. On dit que $F : \mathbb{E} \to \mathbb{F}$ est une équation locale de V en x, de classe C^r ou analytique, s'il existe un ouvert U de $\mathbb{E}$ contenant x sur lequel F est défini et tel que*

1. *$V \cap U = \{y \in U \ : \ F(y) = 0\}$,*
2. *F est de classe C^r ou analytique et $DF(x)$ est surjective,*
3. *$\ker DF(x)$ possède un supplémentaire fermé dans $\mathbb{E}$.*

Remarque 11. Cette dernière condition est toujours satisfaite lorsque $\mathbb{E}$ est un espace de Hilbert ou bien un espace normé de dimension finie.

Définition 188. *On dit qu'une partie V de $\mathbb{E}$ est une sous-variété de classe C^r, $1 \leq r \leq \infty$, ou analytique si pour tout point x de V on peut trouver une équation locale $F : \mathbb{E} \to \mathbb{F}$ de V en x de classe C^r ou analytique.*

Proposition 189. *Les Définitions 186 et 188 sont équivalentes.*

Preuve Partons de la Définition 186. Pour fabriquer une équation locale de V en x il suffit de prendre $F = \Pi_{\mathbb{K}} \circ \phi$ où $\Pi_{\mathbb{K}} : \mathbb{H} \times \mathbb{K} \to \mathbb{K}$ est la projection sur $\mathbb{K}$. Comme $DF(x) = \Pi_{\mathbb{K}} \circ D\phi(x)$ cette dérivée est surjective ; son noyau est $\ker DF(x) = D\phi(x)^{-1}(\mathbb{H} \times \{0\})$ qui admet $D\phi(x)^{-1}(\{0\} \times \mathbb{K})$ pour supplémentaire fermé.

Partons de la Définition 188. On dispose d'une équation locale $F : U \subset \mathbb{E} \to \mathbb{F}$ au point $x \in V$. Nous supposons pour simplifier que $x = 0$. Soit E_s un supplémentaire fermé de $\ker DF(0)$ dans $\mathbb{E}$ et soit

$$\Pi_k : \mathbb{E} \to \ker DF(0)$$

la projection de $\mathbb{E}$ sur $\ker DF(0)$ parallèlement à E_s. Puisque E_s et $\ker DF(0)$ sont fermés, cette projection est continue. On définit

$$\phi : U \subset \mathbb{E} \to \ker DF(0) \times \mathbb{F},\ \phi(u) = (\Pi_k(u), F(u)).$$

Sa dérivée en zéro est $D\phi(0) = (\Pi_k(u), DF(0))$. On vérifie facilement que c'est un isomorphisme. Ceci prouve, par le théorème des fonctions inverses (Théorème 185) que ϕ est un difféomorphisme sur un ouvert U_1 de $\mathbb{E}$ tel que $0 \in U_1 \subset U$. On remarque enfin que

$$\phi(U_1 \cap V) = \phi(U_1) \cap \ker DF(x) \times \{0\}.$$

On obtient ainsi la Définition 186. □

Définition 190. *Soit V une sous-variété de $\mathbb{E}$. On appelle espace tangent à $x \in V$ l'espace T_xV des vecteurs vitesse au point x des courbes dans V passant par x, autrement dit, $v \in T_xV$ s'il existe $\gamma :]-1, 1[\to V$ de classe C^1 telle que $\gamma(0) = x$ et $\dot{\gamma}(0) = v$, où l'on note*

$$\dot{\gamma}(t) = \frac{d\gamma(t)}{dt}.$$

Proposition 191. *Soient V une sous-variété de $\mathbb{E}$, $x \in V$ et F une équation locale de V en x. On a $T_xV = \ker DF(x)$.*

Preuve L'inclusion $\subset$ s'obtient par dérivation de $F(\gamma(t)) = 0$ où $\gamma(t)$ est une courbe dans V telle que $\gamma(0) = x$ et $\dot{\gamma}(0) = v$. On a, par dérivation des fonctions composées,

$$DF(\gamma(t))\dot{\gamma}(t)|_{t=0} = DF(x)v = 0$$

donc $v \in \ker DF(x)$. Pour prouver l'autre inclusion on se donne $v \in \ker DF(x)$ et il faut construire une courbe $\gamma(t)$ qui satisfasse aux trois exigences suivantes : $\gamma(t) \in V$, $\gamma(0) = x$ et $\dot{\gamma}(0) = v$. Pour réaliser ce programme, on suppose que $x = 0$ et on utilise le difféomorphisme $\phi : U_1 \subset \mathbb{E} \to \ker DF(0) \times \mathbb{F}$ décrit au cours de la preuve de la Proposition 189. A l'aide de ϕ on relève la courbe $t \to (tv, 0) \in \ker DF(0) \times \{0\}$ ce qui donne : $\gamma(t) = \phi^{-1}(tv, 0)$. On vérifie facilement que γ répond à la question. □

Corollaire 192. *L'espace tangent T_xV est un sous-espace vectoriel fermé de $\mathbb{E}$.*

Corollaire 193. *Lorsque $\mathbb{E}$ est un espace de dimension finie et que V est une partie connexe de $\mathbb{E}$, la dimension de l'espace l'espace tangent T_xV est indépendante de $x \in V$. On la note $\dim V$ et on l'appelle la dimension de V.*

Preuve Soit $F : U \to \mathbb{R}^m$ une équation locale de V au voisinage de $x \in V$. Quitte à remplacer U par un ouvert plus petit, on peut supposer que $DF(y)$ est surjective pour tout $y \in U$: F est une équation locale de V sur U tout entier. Ainsi $\dim T_yV = \dim \ker DF(y) = n - m$ est constante sur $U \cap V$. Par connexité de V et par continuité de DF cette dimension est partout la même sur V. □

Exemple 1. Soit $\gamma :]-1, 1[\to \mathbb{R}^2$ une courbe plane paramétrée de classe C^r. Supposons que $\frac{d\gamma}{dt}(t) \neq 0$ pour tout $t \in]-1, 1[$, que γ soit injective (c'est-à-dire que la courbe soit sans point double) et que $\gamma^{-1} : \gamma(]-1, 1[) \to]-1, 1[$ soit continue. Sous ces hypothèses l'ensemble $V = \gamma(]-1, 1[) \subset \mathbb{R}^2$ est une sous-variété de dimension 1 de $\mathbb{R}^2$. L'espace tangent à V en un point est la direction portée par le vecteur $\frac{d\gamma}{dt}$ en ce point.

Exemple 2. La sphère, le cylindre, le tore sont des sous-variétés de dimension 2 de $\mathbb{R}^3$. Par contre le cône d'équation $z^2 = x^2 + y^2$ ne l'est pas à cause du sommet. Une fois privé de ce sommet il devient une sous-variété.

Exemple 3. Soient U un ouvert de $\mathbb{E}$, $F : U \subset \mathbb{E} \to \mathbb{F}$ de classe C^r et $V = F^{-1}(0)$. Si, pour tout $x \in V$, $DF(x)$ est surjective alors V est une sous-variété de classe C^r et

$$T_xV = \ker DF(x).$$

Exemple 4. Le graphe d'une application $h : \mathbb{E}_1 \to \mathbb{E}_2$ de classe C^r est une sous-variété de classe C^r contenue dans $\mathbb{E}_1 \times \mathbb{E}_2$. De plus

$$T_{(x,h(x))}\,Graphe\ (h) = Graphe\ (Dh(x)).$$

Exemple 5. La sphère unité d'un espace de Hilbert $\mathbb{E}$:

$$S(\mathbb{E}) = \{x \in \mathbb{E} : \|x\| = 1\}$$

est une sous-variété de classe C^∞ contenue dans $\mathbb{E}$. Une équation de $S(\mathbb{E})$ est donnée par $F(x) = \|x\|^2 - 1 = 0$; F est C^∞ et sa dérivée,

$$DF(x) : \mathbb{E} \to \mathbb{R},\ DF(x)u = \langle x, u \rangle ,$$

est surjective pour tout $x \in S(\mathbb{E})$. L'espace tangent à $S(\mathbb{E})$ en x est donc

$$T_x S(\mathbb{E}) = \{u \in \mathbb{E} : \langle x, u \rangle = 0\} = x^\perp$$

l'orthogonal de x dans $\mathbb{E}$.

Exemple 6. La sphère unité d'un espace de Banach n'est pas nécessairement une sous-variété : penser à $\mathbb{R}^2$ muni de la norme

$$\|(x_1, x_2)\| = \max(|x_1|, |x_2|) .$$

Cette « sphère unité » est ici le carré de sommets $(\pm 1, \pm 1)$, ce n'est pas une sous-variété de $\mathbb{R}^2$.

Exemple 7. Soient $\mathbb{E}$ et $\mathbb{F}$ deux espaces euclidiens, $\dim \mathbb{E} = n$, $\dim \mathbb{F} = m$, U un ouvert de $\mathbb{E}$ et $F : U \to \mathbb{F}$ une application de classe C^s, $s \geq 1$, ou analytique. Supposons que le rang de $DF(x)$ soit constant, égal à r pour tout $x \in U$. Alors, $V = F^{-1}(0)$ est une sous-variété de classe C^s ou analytique de $\mathbb{E}$ et sa dimension est $n - r$.

Prouver l'affirmation contenue dans cet exemple demande un peu de travail. L'argument repose sur une représentation des applications de rang constant : via un changement de variable dans l'espace de départ et un changement de variable dans l'espace d'arrivée, une telle application s'écrit

$$(x_1, \ldots, x_n) \in \mathbb{R}^n \to (x_1, \ldots, x_r, 0, \ldots, 0) \in \mathbb{R}^m.$$

Plus précisément

Proposition 194. *Avec les hypothèses et les notations de l'Exemple 7, pour tout $p \in U$, il existe des difféomorphismes*

$$\xi : U_p \to U_0, \quad \eta : V_{F(p)} \to V_0,$$

définis sur des ouverts $p \in U_p \subset \mathbb{E}$ et $F(p) \in V_{F(p)} \subset \mathbb{F}$ dont les images sont des ouverts $0 \in U_0 \subset \mathbb{R}^n$ et $F(p) \in V_0 \subset \mathbb{R}^m$ et tels que :

$$\xi(p) = 0, \quad \eta(F(p)) = 0$$

et

$$\eta \circ \mathbb{F} \circ \xi^{-1}(x_1, \ldots, x_n) = (x_1, \ldots, x_r, 0, \ldots, 0)$$

pour tout $x \in U_0$.

Preuve On simplifie l'énoncé en supposant que $\mathbb{E} = \mathbb{R}^n$, $\mathbb{F} = \mathbb{R}^m$, $p = 0$ et $F(p) = 0$. Ecrivons

$$F(x) = (F_1(x), \ldots, F_m(x)).$$

Puisque $DF(0)$ est de rang r, quitte à renuméroter équations et inconnues nous supposons que la matrice

$$A = \left(\frac{\partial F_i}{\partial x_j}(0)\right)_{1 \leq i,j \leq r}$$

est inversible. On pose alors

$$\xi(x) = (\xi_1(x), \ldots, \xi_n(x)) = (F_1(x), \ldots, F_r(x), x_{r+1}, \ldots, x_n).$$

Nous voyons que sa dérivée en 0 est inversible puisque

$$\left(\frac{\partial \xi_i}{\partial x_j}(0)\right) = \begin{pmatrix} A & * \\ 0 & I_{n-r} \end{pmatrix}$$

et que A est inversible. Ceci fait de ξ, par le théorème d'inversion locale (Théorème 185), un difféomorphisme d'un voisinage ouvert de 0 dans $\mathbb{R}^n$ dans un autre voisinage ouvert de 0 dans $\mathbb{R}^n$. Notons maintenant

$$\psi(x) = (\psi_1(x), \ldots, \psi_m(x)) = F \circ \xi^{-1}(x).$$

Par construction de ξ on a

$$\psi_i(x) = x_i, \ 1 \leq i \leq r,$$

de sorte que

$$\left(\frac{\partial \psi_i}{\partial x_j}(x)\right) = \begin{pmatrix} I_r & 0 \\ * & B \end{pmatrix}$$

où B est la matrice $(m-r) \times (n-r)$ de terme général $\partial \psi_i / \partial x_j$, $r+1 \leq i \leq m$, $r+1 \leq j \leq n$. Comme cette dérivée est de rang r pour tout x dans un voisinage de zéro, la matrice B est identiquement nulle pour tout x et les fonctions ψ_i, $r+1 \leq i \leq m$, ne dépendent pas des variables x_j, $r+1 \leq j \leq n$:

$$\psi_i(x) = \psi_i(x_1, \ldots, x_r), \ r+1 \leq i \leq m.$$

On introduit enfin

$$\eta(y) = (\eta_1(y), \ldots, \eta_m(y)) = (y_1, \ldots, y_r, y_{r+1} - \psi_{r+1}(y_1, \ldots, y_r), \ldots, y_m - \psi_m(y_1, \ldots, y_r)).$$

Sa dérivée en 0 est égale à

$$\left(\frac{\partial \eta_i}{\partial y_j}(0)\right) = \begin{pmatrix} I_r & 0 \\ * & I_{n-r} \end{pmatrix}$$

ce qui prouve que η est un difféomorphisme sur un voisinage ouvert de 0. On voit enfin que

$$\eta \circ \mathbb{F} \circ \xi^{-1}(x_1, \ldots, x_n) = (x_1, \ldots, x_r, 0, \ldots, 0)$$

d'où la proposition. □

Corollaire 195. *Sous ces hypothèses $V = F^{-1}(0)$ est une sous-variété de dimension $n - r$.*

Preuve La proposition précédente montre que pour tout $x \in V$ il existe un voisinage ouvert U de x dans $\mathbb{E}$ et une application $G : \mathbb{U} \to \mathbb{R}^r$ de même régularité que F, telle que $V \cap U = G^{-1}(0) : G$ est une carte locale. □

Les sous-variétés différentiables constituent un cadre naturel pour les problèmes d'optimisation différentiables avec contraintes de type égalité. L'énoncé suivant donne, dans un tel cadre, des conditions d'optimalité du premier ordre.

Proposition 196. *Soit $f : U_1 \subset E_1 \to E_2$ une application de classe C^1 définie sur un ouvert U_1 d'un espace de Hilbert E_1 et à valeurs dans un espace de Hilbert E_2. Notons $F(x) = \|f(x)\|^2_{E_2}$ et soit V une sous-variété différentiable de E_1, de classe C^1, contenue dans U_1. Pour tout $a \in V$ tel que*

$$F(a) = \min_{x \in V} F(x)$$

on a

$$Df(a)^* f(a) \in N_a V$$

l'espace normal à V en a c'est à dire l'orthogonal de l'espace tangent : $N_a V = (T_a V)^\perp$.

Preuve Suivant la Définition 186, il existe un difféomorphisme ϕ de classe C^1 d'un voisinage ouvert U de a dans E_1 sur un ouvert d'un espace de Banach $\mathbb{H} \times \mathbb{K}$ tel que

$$\phi(U \cap V) = \phi(U) \cap (\mathbb{H} \times \{0\}).$$

Quitte à remplacer U_1 et V par des ensembles plus petits, on peut supposer que $U = U_1$ de sorte que

$$\phi(V) = \phi(U_1) \cap (\mathbb{H} \times \{0\}).$$

Posons $W = \phi(U_1)$. Le difféomorphisme inverse

$$\psi = \phi^{-1} : W \subset \mathbb{H} \times \mathbb{K} \to U_1 \subset E_1$$

vérifie

$$\psi(W \cap (\mathbb{H} \times \{0\})) = V.$$

De plus, en notant $b = \phi(a)$,

$$D\psi(b)(\mathbb{H} \times \{0\}) = T_a V.$$

En effet, soit $u \in T_a V$. Par la Définition 190 il existe une courbe C^1, $\gamma(t) \in V$, telle que $\gamma(0) = a$ et $\dot\gamma(0) = u$. On a $\phi(\gamma(t)) \in W \cap (\mathbb{H} \times \{0\})$ et donc

$$v = \frac{d}{dt}\phi(\gamma(t))|_{t=0} = D\phi(a)u \in \mathbb{H} \times \{0\}.$$

ceci prouve que

$$u = D\phi(a)^{-1}v = D\psi(b)v$$

pour un vecteur $v \in \mathbb{H} \times \{0\}$. Réciproquement, soit $u = D\psi(b)v$ avec $v \in \mathbb{H} \times \{0\}$. Notons

$$\gamma(t) = \psi(b + tv).$$

Il est facile de voir que $\gamma(t) \in V$, $\gamma(0) = a$ et que $\dot{\gamma}(0) = D\psi(b)v$. Ceci prouve l'égalité $D\psi(b)(\mathbb{H} \times \{0\}) = T_aV$.

Revenons à notre minimum. Puisque $F(a) \leq F(x)$ pour tout $x \in V$ on a aussi $F(\psi(b)) \leq F(\psi(y))$ pour tout $y \in W \cap (\mathbb{H} \times \{0\})$ et b réalise le minimum de la fonction C^1 $F \circ \psi$ définie sur un ouvert d'un espace normé. C'est donc un point stationnaire de $F \circ \psi$:

$$D(F \circ \psi)(b) = DF(a) \circ D\psi(b) = 0.$$

Cela signifie que $DF(a)(D\psi(b)v) = 0$ pour tout $v \in \mathbb{H} \times \{0\}$ ou, en d'autres termes, que $DF(a)u = 0$ pour tout $u \in T_aV$. Ainsi

$$\langle Df(a)u, f(a)\rangle = \langle u, Df(a)^* f(a)\rangle = 0$$

pour tout $u \in T_aV$ ce qui signifie bien que $Df(a)^* f(a) \in (T_aV)^\perp = N_aV$. □

6.6 Opérateurs linéaires bornés

Soient $\mathbb{E}$ et $\mathbb{F}$ des espaces de Banach. L'espace $\mathcal{L}(\mathbb{E}, \mathbb{F})$ des opérateurs linéaires continus $L : \mathbb{E} \to \mathbb{F}$ (on dit aussi bornés) est un espace de Banach pour la norme

$$\|L\| = \sup_{\|x\|=1} \|Lx\|.$$

Lorsque $\mathbb{E} = \mathbb{F}$ on note plutôt $\mathcal{L}(\mathbb{E})$.

Théorème 197. Théorème de Banach-Steinhaus. *Pour toute famille $\mathcal{F} \subset \mathcal{L}(\mathbb{E}, \mathbb{F})$ telle que*

$$\sup\{\|Lx\| : L \in \mathcal{F}\} < \infty$$

on a aussi

$$\sup\{\|L\| : L \in \mathcal{F}\} < \infty.$$

Théorème 198. Théorème de l'inverse continu. *Si $L \in \mathcal{L}(\mathbb{E}, \mathbb{F})$ est bijective, son inverse L^{-1} est continu.*

Références

1. Allgower, E., and K. Georg, *Numerical Continuation Methods*, Springer, Berlin Heidelberg New York, (1990)
2. Beyn W.-J., *On smoothness and invariance properties of the Gauss-Newton method*, Num. Funct. Anal. and Opt. **14**, 243–252 (1993)
3. Ben-Israel A., *A Newton-Raphson Method for the Solution of Systems of Equations*, J. Math. Anal. Appl. **15**, 243–252 (1966)
4. Ben-Israel A., T. Greville, *Generalized Inverses Theory and Applications*, J. Wiley and Sons, (1974)
5. Bittanti S., A. Laub, J. C. Willems, *The Riccati equation*, (1991)
6. Blum L., F. Cucker, M. Shub, S. Smale, *Complexity and Real Computation*, (1997)
7. Bollobàs B., *Linear Analysis*, Cambridge University Press, (1990)
8. Bourbaki N., *Eléments de mathématique. Topologie Générale*, 4ème éd., Hermann, (1965)
9. Bourbaki N., *Fonctions d'une variable réelle, Chapitres 1 à 3*, Hermann, Paris, (1958)
10. Brézis H., *Analyse fonctionnelle.*, Masson, (1987)
11. Chabat B., *Introduction à l'analyse complexe, tome 2*, MIR, Moscou, (1990)
12. Chaitin-Chatelin F, V. Frayssé., *Lectures on Finite Precision Computations.* SIAM, Philadelphia, (1996)
13. Dedieu, J.-P. and M.-H. Kim, *Newton's Method for Analytic Systems of Equations with Constant Rank Derivatives.* Journal of Complexity, **18**, 187–209 (2002)
14. Dedieu, J.-P. and M. Shub, *Newton's Method for Overdetermined Systems of Equations.* Mathematics of Computation, **69** 1099–1115 (2000)
15. Demazure M., *Catastrophes et Bifurcations*, Ellipses, (1989)
16. Dennis, J., and R. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equation*, Prentice Hall, (1983)
17. Devaney R., *Chaotic Dynamical Systems*, Addison Wesley, (1989)
18. Dieudonné, J., *Eléments d'Analyse*, Gauthier-Villars, (1968)
19. Francis J. G. F., *The QR transformation I, II,* Comput. J. **1961**, 265–271, 332–345 (1962)

20. Gauss, K.F., *Theoria Motus Corporum Coelestian*, Werke, **7**, 240–254 (1809)
21. Goldstine, H., *A History of Numerical Analysis from the 16th through the 19th Century*, New York, (1977)
22. Hartman P., *Ordinary Differential Equations*, Second Edition, Birkhäuser, (1982)
23. Higham N., *Accuracy and Stability of Numerical Algorithms*, SIAM, Second Edition, (2002)
24. Hirsch H., S. Smale, *Differential Equations, Dynamical Systems, and Linear Algebra*, Academic Press, (1974)
25. Hubbard, J., D. Schleicher and S. Sutherland, *How to Really Find Roots of Polynomials by Newton's Method*, preprint, (1998)
26. Irwin M., *On the stable manifold theorem*, Bull. London Math. Soc. **2**, (1970)
27. Irwin M., *On the smoothness of the composition map*, Quart. J. Math. Oxford, **23**, 113 (1972)
28. Kantorovich, L., *Sur la méthode de Newton*, Travaux de l'Institut des Mathématiques Steklov, XXVIII, 104–144 (1949)
29. Kim, M.-H, *Computational complexity of the Euler type algorithm for the roots of polynomials*, Thesis, (1986)
30. Kim, M.-H, *On approximate zeros and rootfinding algorithms for a complex polynomial*, Mathematics of Computation, **51**, 707–719 (1988)
31. Kim, M.-H and S. Sutherland, *Polynomial Root-Finding Algorithms and Branched Covers*, SIAM J. Computing, **23**, 415–436 (1994)
32. Kublanovskaya V. N., *On some algorithms for the solution of the complete eigenvalue problem,* USSR Comput. Math. Math. Phys. 637–657 (1961)
33. Luenberger, D., *Optimization by Vector Space Methods*, J. Wiley and Sons, (1969)
34. Manning, A., *How to be Sure of Solving a Complex Polynomial using Newton's Method*, Bol. Soc. Bras. Mat., **22**, 157–177 (1992)
35. Malajovich, G., *On Generalized Newton's Methods*, Theoretical Comp. Sci., **133**, 65–84 (1994)
36. Ortega, J., and V. Rheinboldt, *Numerical Solutions of Nonlinear Problems*, SIAM, Philadelphia (1968)
37. Ostrowski, A., *Solutions of Equations in Euclidean and Banach Spaces*, Academic Press, New York, (1976)
38. Pan, V., *Solving a Polynomial Equation : Some History and Recent Progress*, SIAM Review, **39**, 187–220 (1997)
39. Rudin W., *Functional Analysis*, Second edition, McGraw Hill Inc., (1991)
40. Rutishauser H., *Une méthode pour la détermination des valeurs propres d'une matrice,* Comptes Rendus Acad. Sci. Paris, **240**, 34–36 (1955)
41. Shub, M., *Stabilité globale des systèmes dynamiques*, Astérisque 56, Société Mathématique de France, (1978)
42. Shub, M., *Some Remarks on Dynamical Systems and Numerical Analysis*, in : Dynamical Systems and Partial Differential Equations, Proceedings of VII ELAM (L. Lara-Carrero and J. Lewowicz eds.), Equinoccio, Universidad Simon Bolivar, Caracas, (1986)
43. Shub, M. and S. Smale, *Complexity of Bézout's Theorem I : Geometric Aspects*, J. Am. Math. Soc. **6**, 459–501 (1993)

44. Shub, M. and S. Smale, *Complexity of Bézout's Theorem II : Volumes and Probabilities,* in : F. Eyssette, A. Galligo Eds. Computational Algebraic Geometry, Progress in Mathematics, Vol. **109**, Birkhäuser, (1993)
45. Shub, M. and S. Smale, *Complexity of Bézout's Theorem IV : Probability of Success, Extensions*, SIAM J. Numer. Anal. **33**, 128–148 (1996)
46. Shub, M. and S. Smale, *Complexity of Bézout's Theorem V : Polynomial Time*, Theoretical Computer Science, **133**, 141–164 (1994)
47. Shub, M. and A. Vasquez, *Some linearly induced Morse-Smale systems, the QR algorithm and the Toda lattice*, in : The Legacy of Sonia Kovalevskaya, Linda Keen Ed., Contemporary Mathematics, Vol. **64**, AMS, 181–193 (1987)
48. Smale, S., *On the Efficiency of Algorithms of Analysis*, Bull. A.M.S. **13** 87–121 (1985)
49. Smale, S., *Algorithms for Solving Equation*, in : Proceedings of the International Congress of Mathematicians, A.M.S. 172–195 (1986)
50. Smale, S., *Newton's Method Estimates from Data at One Point* in : The Merging of Disciplines : New Directions in Pure, Applied and Computational Mathematics (R. Ewing, K. Gross, and C. Martin eds.), Springer (1986)
51. Stewart G. W., J.-G. Sun, *Matrix perturbation theory*, Academic Press, (1990)
52. Sutherland, S., *Finding Roots of Complex Polynomials with Newton's Method*, Thesis, Boston University, (1989)
53. Wang, X., *Convergence of Newton's method and uniqueness of the solution of equations in Banach spaces*, IMA Journal of Num. Anal., **20**, 123–134 (2000)
54. Wang, X., Han, D., *On dominating sequence method in the point estimate and Smale theorem*, Science in China, Series A, **33**, 135–144 (1990)
55. Wilkinson J. H., *The Algebraic Eigenvalue Problem,* Clarendon Press, Oxford (1965)
56. Yosida K., *Fonctional Analysis*, fifth edition, (1978)
57. Ypma, T., *Historical Development of the Newton-Raphson Method*, SIAM Review, **37**, 531–551 (1995)

Index

Déjà parus dans la même collection

1. T. CAZENAVE, A. HARAUX
Introduction aux problèmes d'évolution semi-linéaires. 1990

2. P. JOLY
Mise en œuvre de la méthode des éléments finis. 1990

3/4. E. GODLEWSKI, P.-A. RAVIART
Hyperbolic systems of conservation laws. 1991

5/6. PH. DESTUYNDER
Modélisation mécanique des milieux continus. 1991

7. J. C. NEDELEC
Notions sur les techniques d'éléments finis. 1992

8. G. ROBIN
Algorithmique et cryptographie. 1992

9. D. LAMBERTON, B. LAPEYRE
Introduction au calcul stochastique appliqué. 1992

10. C. BERNARDI, Y. MADAY
Approximations spectrales de problèmes aux limites elliptiques. 1992

11. V. GENON-CATALOT, D. PICARD
Eléments de statistique asymptotique. 1993

12. P. DEHORNOY
Complexité et décidabilité. 1993

13. O. KAVIAN
Introduction à la théorie des points critiques. 1994

14. A. BOSSAVIT
Électromagnétisme, en vue de la modélisation. 1994

15. R. KH. ZEYTOUNIAN
Modélisation asymptotique en mécanique des fluides Newtoniens. 1994

16. D. BOUCHE, F. MOLINET
Méthodes asymptotiques en électromagnétisme. 1994

17. G. BARLES
Solutions de viscosité des équations de Hamilton-Jacobi. 1994

18. Q. S. NGUYEN
Stabilité des structures élastiques. 1995

19. F. ROBERT
Les systèmes dynamiques discrets. 1995

20. O. PAPINI, J. WOLFMANN
Algèbre discrète et codes correcteurs. 1995

21. D. COLLOMBIER
Plans d'expérience factoriels. 1996

22. G. GAGNEUX, M. MADAUNE-TORT
Analyse mathématique de modèles non linéaires de l'ingénierie pétrolière. 1996

23. M. DUFLO
Algorithmes stochastiques. 1996

24. P. DESTUYNDER, M. SALAUN
Mathematical Analysis of Thin Plate Models. 1996

25. P. ROUGEE
Mécanique des grandes transformations. 1997

26. L. HÖRMANDER
Lectures on Nonlinear Hyperbolic Differential Equations. 1997

27. J. F. BONNANS, J. C. GILBERT, C. LEMARÉCHAL, C. SAGASTIZÁBAL
Optimisation numérique. 1997

28. C. COCOZZA-THIVENT
Processus stochastiques et fiabilité des systèmes. 1997

29. B. LAPEYRE, É. PARDOUX, R. SENTIS
Méthodes de Monte-Carlo pour les équations de transport et de diffusion. 1998

30. P. SAGAUT
Introduction à la simulation des grandes échelles pour les écoulements de fluide incompressible. 1998

Déjà parus dans la même collection

31. E. RIO
Théorie asymptotique des processus aléatoires faiblement dépendants. 1999

32. J. MOREAU, P.-A. DOUDIN, P. CAZES (EDS.)
L'analyse des correspondances et les techniques connexes. 1999

33. B. CHALMOND
Eléments de modélisation pour l'analyse d'images. 1999

34. J. ISTAS
Introduction aux modélisations mathématiques pour les sciences du vivant. 2000

35. P. ROBERT
Réseaux et files d'attente : méthodes probabilistes. 2000

36. A. ERN, J.-L. GUERMOND
Eléments finis : théorie, applications, mise en œuvre. 2001

37. S. SORIN
A First Course on Zero-Sum Repeated Games. 2002

38. J. F. MAURRAS
Programmation linéaire, complexité. 2002

39. B. YCART
Modèles et algorithmes Markoviens. 2002

40. B. BONNARD, M. CHYBA
Singular Trajectories and their Role in Control Theory. 2003

41. A. TSYBAKOV
Introdution à l'estimation non-paramétrique. 2003

42. J. ABDELJAOUED, H. LOMBARDI
Méthodes matricielles – Introduction à la complexité algébrique. 2004

43. U. BOSCAIN, B. PICCOLI
Optimal Syntheses for Control Systems on 2-D Manifolds. 2004

44. L. YOUNES
Invariance, déformations et reconnaissance de formes. 2004

45. C. BERNARDI, Y. MADAY, F. RAPETTI
Discrétisations variationnelles de problèmes aux limites elliptiques. 2004

46. J.-P. FRANÇOISE
Oscillations en biologie : Analyse qualitative et modèles. 2005

47. C. LE BRIS
Systèmes multi-échelles : Modélisation et simulation. 2005

48. A. HENROT, M. PIERRE
Variation et optimisation de formes : Une analyse géometric. 2005

49. B. BIDÉGARAY-FESQUET
Hiérarchie de modèles en optique quantique : De Maxwell-Bloch à Schrödinger non-linéaire. 2005

50. R. DÁGER, E. ZUAZUA
Wave Propagation, Observation and Control in $1-d$ Flexible Multi-Structures. 2005

51. B. BONNARD, L. FAUBOURG, E. TRÉLAT
Mécanique céleste et contrôle des véhicules spatiaux. 2005

52. F. BOYER, P. FABRIE
Eléments d'analyse pour l'étude de quelques modèles d'écoulements de fluides visqueux incompressibles. 2005

53. E. CANCÈS, C. Le BRIS, Y. MADAY
Méthodes mathématiques en chimie quantique. Une introduction. 2006

54. J.-P. DEDIEU
Points fixes, zeros et la methode de Newton. 2006

55. P. LOPEZ, A. S. NOURI
Théorie élémentaire et pratique de la commande par les régimes glissants. 2006

56. J. COUSTEIX, J. MAUSS
Analyse asymptotitque et couche limite. 2006

Druck und Bindung: Strauss GmbH, Mörlenbach