



Licence de Mathématiques - Sorbonne Université
année 2018-2019

UE 3M133

Analyse appliquée

Hervé Le Dret

Laboratoire Jacques-Louis Lions

14 décembre 2018

Table des matières

Introduction	1
1 Prérequis et motivation	3
1.1 Ce qu'il faut savoir avant de commencer	3
1.1.1 Prérequis généraux	3
1.1.2 Topologie et calcul différentiel	3
1.1.3 Culture générale	4
1.2 Motivation	4
1.2.1 Un peu de philosophie : qu'est-ce qu'un nombre réel ?	4
1.2.2 Mais pourquoi ? À quoi ça sert $f(x) = 0$?	5
2 $f(x) = 0$ en dimension 1	7
2.1 Le théorème des valeurs intermédiaires	7
2.2 Une première méthode, la dichotomie	9
2.3 Vitesse de convergence	12
2.4 Quelques autres méthodes en vrac	15
2.5 Itérations de point fixe	18
2.6 La méthode de Newton	20
3 $f(x) = 0$ en dimension quelconque	25
3.1 Topologie	25
3.2 Le théorème de point fixe de Banach	28
3.3 La méthode de Newton dans \mathbb{R}^n	31
3.3.1 Rappels (?) de calcul différentiel	32
3.3.2 $f(x) = 0$ du point de vue théorique	36
3.3.3 La méthode de Newton(-Raphson)	37
3.4 Le théorème de (Newton)-Kantorovich	46
3.5 Une vraie application, la méthode d'Euler implicite	50
3.5.1 Considérations de mise en œuvre pratique	53
3.6 Les méthodes de quasi-Newton	54

Chapitre 1

Prérequis et motivation

1.1 Ce qu'il faut savoir avant de commencer

Les mathématiques sont une discipline cumulative : on y construit des édifices de plus en plus élevés sur des bases larges et solides. Inutile de rappeler ce qu'il advient des édifices construits sur du sable. Les fondations mathématiques sont établies tant bien que mal (par manque de temps) dans les premières années de licence. Tout ce qui a été vu précédemment est donc considéré comme acquis, puisque l'on ne saurait croire à cette légende urbaine de l'existence de gens qui oublient tout une fois l'examen passé. Ce qui suit est une liste plus spécifique de notions antérieures qui seront plus ou moins directement utilisées dans ce cours. Si par malheur, on ne se sent pas complètement au point dessus, il est toujours temps de revoir ce que l'on a vu en L1 et L2, en consultant les notes de cours que l'on n'a pas manqué de conserver soigneusement. Là aussi, c'est sûrement une légende urbaine cette histoire qu'il existe des étudiants qui ne prennent pas de notes en cours, ou qui les mettent à la poubelle le lendemain de l'examen.

1.1.1 Prérequis généraux

- Ensembles, applications.
- \mathbb{R} , \mathbb{C} , propriétés algébriques et topologiques.
- Algèbre linéaire en dimension finie, matrices (on ne saurait trop insister sur combien l'algèbre linéaire est fondamentale).
- Notion d'intégrale (par exemple de Riemann). Une intégrale fonction de sa borne supérieure donne une primitive de l'intégrande (par exemple si celle-ci est continue).

1.1.2 Topologie et calcul différentiel

- Notions de topologie dans \mathbb{R}^n : normes, boules, ouverts, fermés, suites et convergence de suites, suites de Cauchy.
- Fonctions continues sur (une partie de) \mathbb{R}^n : définitions en un point, séquentielle et topologique.
- Différentiabilité et dérivées partielles : différentielle, gradient, matrice jacobienne, fonctions de classe C^1 , inégalité des accroissements finis.
- Dérivation des fonctions composées à plusieurs variables.

— Dérivées partielles d'ordre supérieur : matrice hessienne, théorème de Schwartz, formules et inégalités de Taylor-Lagrange, fonctions de classe C^k .

Il n'est pas impossible que ces prérequis-là soient un peu ambitieux en L3. On en reverra une bonne partie en cours de route, plutôt rapidement et sans démonstration car ce n'est pas l'objet du cours.

1.1.3 Culture générale

Pour avoir une idée de ce à quoi peut bien servir tout ce dont nous allons parler en dehors des mathématiques elles-mêmes, il n'est pas inutile de posséder quelques restes (aussi beaux que possible) de physique, de chimie, de mécanique, etc.

1.2 Motivation

« Analyse appliquée », c'est vague comme titre. C'est surtout très vaste pour une petite UE de 3 ECTS. Nous allons donc nous concentrer sur une petite partie de ce que l'on peut appeler analyse appliquée, à savoir la résolution, autant que faire se peut, d'équations de la forme $f(x) = 0$, où f est une fonction dont on dira plus par la suite.

1.2.1 Un peu de philosophie : qu'est-ce qu'un nombre réel ?

Si f est une fonction de \mathbb{R} dans \mathbb{R} , résoudre l'équation $f(x) = 0$ consiste à déterminer les nombres $x \in \mathbb{R}$ qui annulent f , on dit que ce sont des *racines* de f (en particulier si f est une fonction polynomiale). Mais que signifie « déterminer un nombre réel » ?

Il existe plusieurs constructions du corps des nombres réels à partir au départ de l'ensemble des entiers naturels \mathbb{N} . Ces constructions fournissent plusieurs objets isomorphes entre eux que l'on désigne commodément par un seul symbole \mathbb{R} . Parmi les plus connues, les coupures de Dedekind, qui donnent facilement la propriété de borne supérieure,¹ la construction à l'aide de classes d'équivalence de suites de Cauchy de nombres rationnels, qui donne facilement que toute suite de Cauchy est convergente dans \mathbb{R} ,² ou simplement l'écriture sous forme de développement décimal illimité, qui ne donne pas grand-chose facilement, si ce n'est l'impression (trompeuse) d'avoir une intuition correcte de ce qu'est un nombre réel : $e = 2,718281828459045\dots$

Tout cela est mathématiquement bien fondé, mais il n'en reste pas moins que « un nombre réel » reste quand même un objet bien mystérieux, surtout si l'on souhaite le déterminer.

Dans ce cours, on va adopter le point de vue que déterminer un nombre réel, c'est en fait décrire une façon de l'approcher par une suite de nombres que l'on sait calculer, par exemple des rationnels, éventuellement eux-mêmes approchés par ce que l'on peut calculer en virgule flottante sur ordinateur.³ On sera donc plus proche des suites de Cauchy ou du développement décimal que des coupures de Dedekind. Ce n'est bien sûr pas la seule façon

1. À savoir que toute partie non vide et majorée de \mathbb{R} admet une borne supérieure, c'est-à-dire un plus petit des majorants. Une propriété cruciale.

2. Autre propriété cruciale, équivalente à la précédente.

3. On n'en dira pas plus sur l'aspect implémentation effective de tout ce que l'on va raconter.

d'envisager les choses, mais c'est celle que l'on retiendra ici. C'est en tout cas une façon qui ressort pleinement de l'analyse : en analyse, on passe son temps à majorer, minorer, approcher toutes sortes d'objets. Ici, ce seront les racines de certaines fonctions.

1.2.2 Mais pourquoi ? À quoi ça sert $f(x) = 0$?

En fait, dans des tas de situations, on a besoin de trouver des x tels que $f(x) = 0$, avec $f: \mathbb{R}^n \rightarrow \mathbb{R}^n$ par exemple. Cela peut être parce que le ou les x en question sont des valeurs d'intérêt pour tel ou tel problème. Cela peut également être parce que, à l'intérieur de calculs plus compliqués, on a besoin de racines d'un certain f comme résultats intermédiaires que l'on oubliera in fine.

Prenons un vieil exemple : on veut calculer la longueur de la diagonale d'un carré de côté unité. Pythagore, qu'il ait réellement existé ou non, est avec nous ! Cette longueur x est telle que $x^2 = 1 + 1$. On prend alors $f: \mathbb{R} \rightarrow \mathbb{R}$, $f(x) = x^2 - 2$. Une longueur étant positive, on s'intéresse donc à la racine positive de cette fonction polynomiale, que l'on a coutume de noter $x = \sqrt{2}$. Ce nombre réel est un nombre irrationnel, la suite infinie de ses décimales ne présente aucune périodicité, il est donc impossible de connaître entièrement ce nombre sous forme décimale.⁴ Maintenant, déterminer $\sqrt{2}$ ne signifie pas appuyer sur la touche $\sqrt{}$ d'une calculette, puis sur la touche 2. La calculette doit bien se débrouiller d'une façon ou d'une autre pour afficher un petit nombre de décimales de $\sqrt{2}$ en lesquelles on puisse avoir raisonnablement confiance. Une méthode presque aussi ancienne que cet exemple pour déterminer au sens précédent $\sqrt{2}$ est la *méthode de Héron*. Elle consiste à construire la suite récurrente

$$x_0 = 1, \quad x_{n+1} = \frac{x_n}{2} + \frac{1}{x_n}.$$

Cette suite de nombres rationnels converge, très rapidement, vers $\sqrt{2}$.

$$\begin{aligned} x_1 &= \frac{1}{2} + 1 = \frac{3}{2} = 1,5 \\ x_2 &= \frac{3}{4} + \frac{2}{3} = \frac{17}{12} = 1,4166666\ldots \\ x_3 &= \frac{17}{24} + \frac{12}{17} = \frac{577}{408} = 1,4142156862745098\ldots \\ x_4 &= \frac{577}{816} + \frac{408}{577} = \frac{665857}{470832} = 1,41421356237469\ldots \\ &\vdots \end{aligned}$$

En appuyant sur la touche $\sqrt{}$ puis sur la touche 2, la calculette nous donne

$$\sqrt{2} = 1,414213562373095\ldots$$

et l'on peut avoir confiance dans ces quinze décimales affichées,⁵ sinon on peut espérer que

4. Bien sûr, cela n'empêche nullement de pouvoir manipuler $\sqrt{2}$ symboliquement, tout comme, e , π , γ , etc. Mais cela n'est pas notre propos ici.

5. Au sens où $1,414213562373095 < \sqrt{2} < 1,414213562373096$. Les petits points à droite de l'écriture ne veulent pas dire grand chose. Les inégalités sont strictes car on sait que $\sqrt{2} \notin \mathbb{Q}$. Notons que $x_5 = \frac{886731088897}{627013566048}$ donne (au moins) les mêmes décimales sur cette même calculatrice. Et ainsi de suite pour x_6 , x_7 , etc.

le fabricant se retrouve vite en faillite. On verra que l'antique méthode de Héron est un cas particulier de la *méthode de Newton*.

Plus généralement, on a assez souvent besoin des racines réelles ou complexes d'un polynôme. Le plus souvent, dès que le degré du polynôme est supérieur à 5, ces racines ne peuvent s'exprimer à l'aide d'opérations algébriques et d'extraction de racines n -èmes effectuées sur les coefficients du polynôme, en tout cas pour un polynôme générique. On pourra alors recourir à des suites d'approximations de ces racines (c'est de toutes façons aussi le cas pour extraire les racines n -èmes quand celles-ci peuvent servir).

Au delà des polynômes, on peut également avoir besoin des racines de fonctions faisant intervenir exponentielles, logarithmes, fonctions trigonométriques directes et inverses, etc., qui elles-mêmes ne sont calculables que de façon approchée.

Enfin, tout cela ne se limite pas à la dimension un, mais les mêmes problématiques se posent en dimension quelconque, voire en dimension infinie. On s'en tiendra le plus souvent à la dimension finie.

Chapitre 2

$f(x) = 0$ en dimension 1

Dans ce chapitre, on va systématiquement se placer sur un intervalle fermé borné $[a, b]$ de \mathbb{R} . On considérera également toujours une fonction (au moins) continue $f : [a, b] \rightarrow \mathbb{R}$.

2.1 Le théorème des valeurs intermédiaires

Le résultat fondamental est le théorème des valeurs intermédiaires, bien connu depuis longtemps dans vos études, mais on va le refaire quand même en le formulant dans le contexte qui nous intéresse.

Théorème 2.1.1 *Soit $f : [a, b] \rightarrow \mathbb{R}$ une fonction continue telle que $f(a)f(b) < 0$. Alors il existe $c \in]a, b[$ tel que $f(c) = 0$.*

Démonstration. On se rappelle tout d'abord que si $x \in [a, b]$ est tel que $f(x) \neq 0$, alors par continuité de f , il existe un intervalle ouvert $]x - \delta, x + \delta[$ avec $\delta > 0$ tel que pour tout $y \in [a, b] \cap]x - \delta, x + \delta[$, $f(y)$ est non nul et du même signe que $f(x)$.

Dire que $f(a)f(b) < 0$, c'est exactement dire que $f(a)$ et $f(b)$ sont non nuls et de signes opposés. Sans perte de généralité, on peut supposer que $f(a) < 0$, et donc $f(b) > 0$, sinon on travaillera avec $-f$. On pose $N = \{x \in [a, b]; f(x) < 0\}$. Cette partie N de \mathbb{R} est non vide car $a \in N$, majorée car incluse dans un intervalle borné. Elle admet donc une borne supérieure dans \mathbb{R} . On note $c = \sup N$ cette borne supérieure. Elle est bien sûr supérieure à a puisque c'est un majorant de N et que $a \in N$.

On a ensuite que $c \leq b$ car b est un majorant de N et c est le plus petit des majorants de N . Par conséquent, $c \in [a, b]$ et cela a un sens de parler de $f(c)$. De plus, pour tout $x \in [c, b]$, on a clairement $f(x) \geq 0$, sinon c ne majorerait pas N . D'après le rappel initial, comme $f(a) < 0$ et $f(b) > 0$, on n'a ni $c = a$, ni $c = b$, c'est-à-dire qu'en fait $c \in]a, b[$.

Mais justement, que peut-on dire de $f(c)$?

La définition de borne supérieure implique en particulier qu'il existe une suite $c_n \in N$, c'est-à-dire telle que $f(c_n) < 0$, avec $c_n \rightarrow c$ quand $n \rightarrow +\infty$. Par continuité de f , on en déduit que $f(c_n) \rightarrow f(c)$, ce qui implique que $f(c) \leq 0$.

De l'autre côté, on a que $N \cap]c, b] = \emptyset$. Pour tout $n \in \mathbb{N}$, on a $c + \frac{b-c}{2^n} \in]c, b]$ et donc $f(c + \frac{b-c}{2^n}) \geq 0$. Comme précédemment, $f(c + \frac{b-c}{2^n}) \rightarrow f(c)$ quand $n \rightarrow +\infty$, ce qui implique que $f(c) \geq 0$. D'où finalement $f(c) = 0$. \diamond

Corollaire 2.1.2 Soit $f: [a, b] \rightarrow \mathbb{R}$ une fonction continue. Si il existe α et β dans $[a, b]$ tels que $f(\alpha)f(\beta) < 0$, alors f admet au moins une racine entre α et β .

C'est ce dernier résultat qui va nous servir pour assurer l'existence d'une racine dans un certain intervalle, avant de se lancer à la déterminer. Attention, il n'y a en général pas unicité de cette racine ! On peut en avoir n'importe quel nombre.

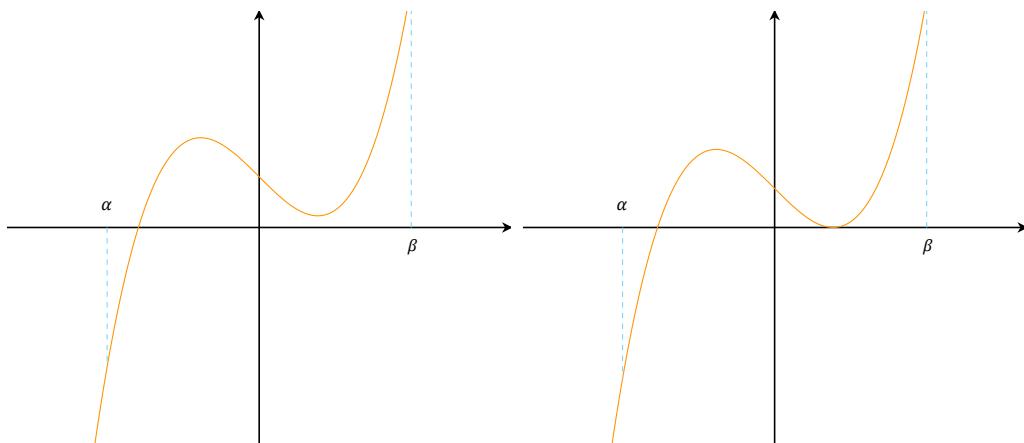


FIGURE 2.1 – Une racine,

deux racines,

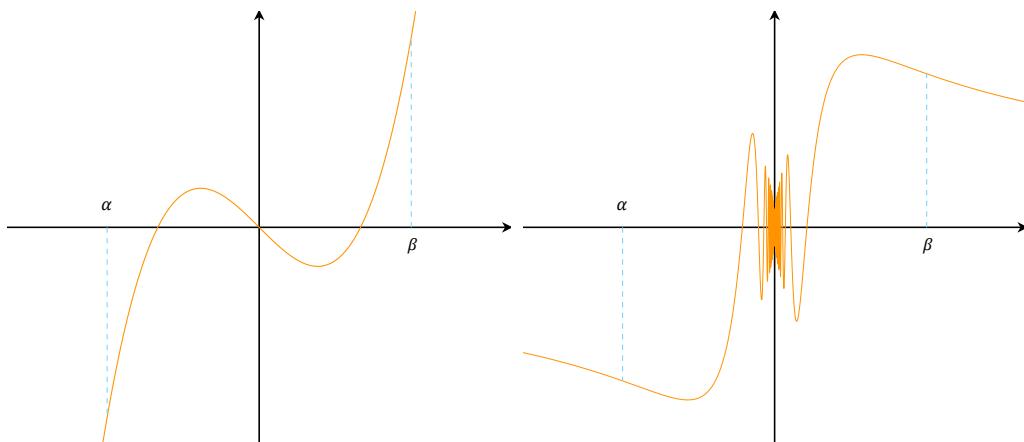


FIGURE 2.2 – trois racines,

une infinité dénombrable de racines,

Mentionnons quand même que l'énoncé traditionnel du théorème des valeurs intermédiaires est qu'une fonction continue sur $[a, b]$ prend toutes les valeurs comprises entre $f(a)$ et $f(b)$. C'est d'ailleurs strictement équivalent à ce qui a été écrit plus haut.

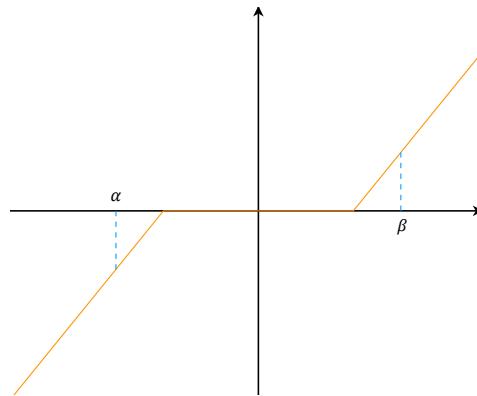


FIGURE 2.3 – une infinité non dénombrable de racines.

2.2 Une première méthode, la dichotomie

On se place dans les hypothèses du Corollaire 2.1.2. On est donc assuré qu'il existe (au moins) une racine de f entre α et β . Supposons pour fixer les idées que $\alpha < \beta$. L'idée de la dichotomie est très simple : diviser pour régner. On va donc regarder ce qui se passe au milieu de l'intervalle $[\alpha, \beta]$, à savoir au point $\frac{\alpha+\beta}{2}$.

Lemme 2.2.1 *Il n'y a que trois cas possibles, qui sont mutuellement exclusifs :*

- i) Soit $f\left(\frac{\alpha+\beta}{2}\right) = 0$,
- ii) soit $f\left(\frac{\alpha+\beta}{2}\right)f(\alpha) < 0$,
- iii) soit $f\left(\frac{\alpha+\beta}{2}\right)f(\beta) < 0$.

Démonstration. Si $f\left(\frac{\alpha+\beta}{2}\right) = 0$, on est dans le cas i). Si $f\left(\frac{\alpha+\beta}{2}\right) \neq 0$, alors ce nombre a un signe, + ou -. Comme $f(\alpha)$ et $f(\beta)$ sont non nuls, de signes opposés, exactement un des deux est du signe opposé à celui de $f\left(\frac{\alpha+\beta}{2}\right)$, c'est-à-dire soit ii), soit iii). \diamond

Cette observation complètement élémentaire est à la base de la méthode de dichotomie. En effet, soit on a eu un coup de chance monstrueux, mais bien improbable, et on a déjà trouvé une racine dans le cas i), soit on est assuré qu'il existe une racine dans $\left]\alpha, \frac{\alpha+\beta}{2}\right]$ dans le cas ii) ou bien dans $\left]\frac{\alpha+\beta}{2}, \beta\right[$ dans le cas iii), ces deux intervalles étant de longueur moitié de l'intervalle de départ. On a donc deux fois mieux localisé une racine qu'au départ.

Attention, si l'on est dans le cas ii) par exemple, cela ne signifie pas qu'il n'existe pas de racine dans $\left]\frac{\alpha+\beta}{2}, \beta\right[$! C'est juste que l'on n'en sait a priori rien puisqu'alors $f\left(\frac{\alpha+\beta}{2}\right)f(\beta) > 0$ et tout peut arriver.

La méthode de dichotomie consiste alors à construire les deux suites de réels a_n et b_n suivantes. On se place dans les hypothèses du Corollaire 2.1.2. On procède par récurrence en posant d'abord

$$a_0 = \alpha, b_0 = \beta.$$

Pour $n \geq 1$, on suppose ensuite déjà construits les nombres $a_i \leq b_i$ pour $i = 0, \dots, n$, tels

que

$$a_i < b_i \text{ et } f(a_i)f(b_i) < 0, \text{ ou bien } a_i = b_i \text{ et } f(a_i) = 0, \quad (2.2.1)$$

$$b_i - a_i \leq \frac{b_0 - a_0}{2^i}. \quad (2.2.2)$$

Ces hypothèses sont bien satisfaites pour $n = 0$. Pour définir a_{n+1} et b_{n+1} , on regarde les différents cas. Si $a_n = b_n$ et $f(a_n) = 0$, alors on pose

$$a_{n+1} = b_{n+1} = a_n.$$

La condition (2.2.1) est bien satisfaite par construction et la condition (2.2.2) également puisque $b_{n+1} - a_{n+1} = 0 \leq \frac{b_0 - a_0}{2^{n+1}}$.

Si par contre $a_n < b_n$ et $f(a_n)f(b_n) < 0$, alors on pose

$$a_{n+1} = a_n, \quad b_{n+1} = \frac{a_n + b_n}{2},$$

dans le cas ii) du Lemme 2.2.1, ou bien

$$a_{n+1} = \frac{a_n + b_n}{2}, \quad b_{n+1} = b_n,$$

dans le cas iii) du Lemme 2.2.1. Dans les deux cas, le Lemme 2.2.1 assure la condition (2.2.1). De plus, dans les deux cas également, on a divisé l'intervalle $[a_n, b_n]$ en deux parties égales, donc

$$b_{n+1} - a_{n+1} = \frac{1}{2}(b_n - a_n) \leq \frac{1}{2} \frac{b_0 - a_0}{2^n} = \frac{b_0 - a_0}{2^{n+1}}.$$

Enfin, dans le cas i), on pose

$$a_{n+1} = b_{n+1} = \frac{a_n + b_n}{2},$$

et là aussi, tout va bien.

On a ainsi défini deux suites $a_n \leq b_n$ par une récurrence qui détermine leurs valeurs pour tout entier n . Ces suites sont de plus très aisément calculables si l'on peut déterminer les signes de $f(a_n)$ et $f(b_n)$.¹

Proposition 2.2.2 *Les suites a_n et b_n sont adjacentes. Elles convergent vers une racine c de f qui est telle que $a_n \leq c \leq b_n$ pour tout n . Enfin, on a l'estimation de vitesse de convergence, ou estimation d'erreur,*

$$\max(|a_n - c|, |b_n - c|) \leq \frac{b_0 - a_0}{2^n}. \quad (2.2.3)$$

Démonstration. À chaque étape de la récurrence, soit $a_{n+1} = a_n$, soit $a_{n+1} = \frac{a_n + b_n}{2} \geq a_n$. La suite a_n est donc croissante. De même, la suite b_n est décroissante. Enfin $b_n - a_n \leq 2^{-n}(b_0 - a_0) \rightarrow 0$ quand $n \rightarrow +\infty$. On a donc bien affaire à deux suites adjacentes, qui sont par conséquent convergentes et de même limite. Notons c leur limite commune. Bien sûr, $c \in [a, b]$.

1. Ce qui peut dans la vraie vie être moins simple qu'il n'y paraît. Notons également qu'il ne peut exister d'algorithme qui décide si un réel donné est nul, cf. le cas i).

Par le Corollaire 2.1.2 et la condition (2.2.1), il existe une racine c_n de f entre a_n et b_n . Comme cette suite de racines est coincée entre deux suites adjacentes, elle converge vers la même limite, $c_n \rightarrow c$ quand $n \rightarrow +\infty$. Comme f est continue, $f(c_n) \rightarrow f(c)$ quand $n \rightarrow +\infty$. Mais $f(c_n) = 0$ pour tout n , donc $f(c) = 0$.

On remarque enfin que les intervalles successifs sont emboîtés par construction : $[a_{n+1}, b_{n+1}] \subset [a_n, b_n]$. Il s'ensuit que, pour tout $k \geq n$, on a $[a_k, b_k] \subset [a_n, b_n]$. Comme $c_k \in [a_k, b_k]$, on en déduit que $c_k \in [a_n, b_n]$, soit $a_n \leq c_k \leq b_n$. Faisant tendre k vers l'infini avec n fixé, on en déduit que $a_n \leq c \leq b_n$. Par conséquent, $\max(|a_n - c|, |b_n - c|) \leq b_n - a_n \leq \frac{b_0 - a_0}{2^n}$, car tout point d'un intervalle est à une distance des extrémités inférieure à la longueur de l'intervalle : $|a_n - c| = c - a_n \leq b_n - a_n$ et de même pour l'autre terme. \diamond

Quels sont les avantages de la dichotomie ? Tout d'abord, elle donne un *encadrement* de la racine qu'elle calcule. C'est-à-dire que l'on est certain que la racine se trouve dans le dernier intervalle calculé, dont on connaît d'ailleurs exactement la longueur : l'estimation d'erreur est *explicite*. On est également certain de sa convergence dans tous les cas, dès que l'on peut la démarrer. Ceci est vrai dès qu'il y a une racine dans l'intervalle de départ, qu'il n'y en ait qu'une ou plusieurs, ou une infinité. En fait, la dichotomie choisit une des racines potentielles de départ. Elle demande par ailleurs très peu de calculs, juste des évaluations de la fonction f et une détermination de signe. Elle est enfin *robuste*, c'est-à-dire relativement peu sensible aux erreurs (d'arrondi) et aux incertitudes éventuelles.

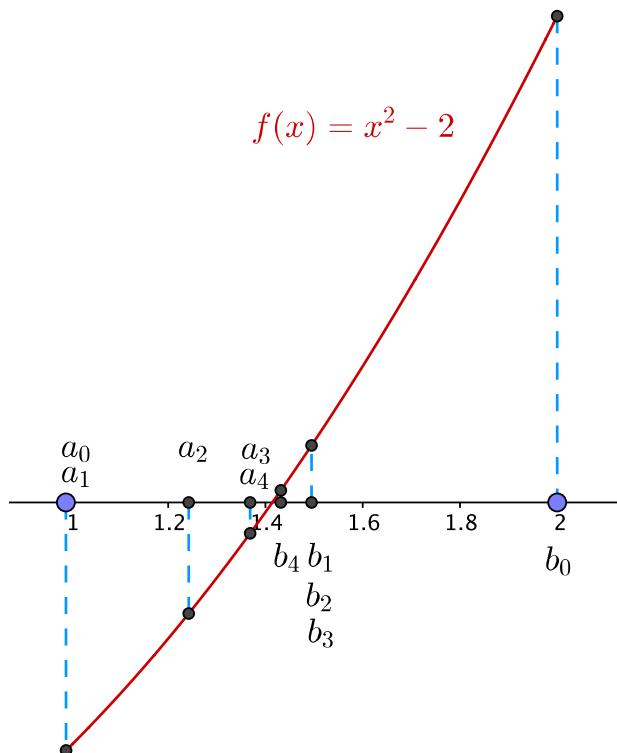


FIGURE 2.4 – Méthode de la dichotomie

Pour ce qui concerne la vitesse de convergence, c'est une autre histoire. Voyons ce que

donne la dichotomie sur la fonction $f(x) = x^2 - 2$ avec $a_0 = 1$ et $b_0 = 2$ (car $(-1) \times 2 < 0$). On donne d'abord les valeurs (tronquées à 7 décimales) de a_n pour $n = 0 \dots 19$, puis celles de b_n pour $n = 0 \dots 19$.

$$\begin{aligned} a_n = 1. & | 1. | 1.25 | 1.375 | 1.375 | 1.40625 | 1.40625 | 1.4140625 | 1.4140625 | 1.4140625 | \\ & 1.4140625 | 1.4140625 | 1.4140625 | 1.4141846 | 1.4141846 | 1.4141846 | 1.4141998 | 1.4142075 \\ & | 1.4142113 | 1.4142132 \end{aligned}$$

$$\begin{aligned} b_n = 2. & | 1.5 | 1.5 | 1.5 | 1.4375 | 1.4375 | 1.421875 | 1.421875 | 1.4179688 | 1.4160156 | \\ & 1.4150391 | 1.4145508 | 1.4143066 | 1.4143066 | 1.4142456 | 1.4142151 | 1.4142151 | 1.4142151 \\ & | 1.4142151 | 1.4142151 \end{aligned}$$

On constate que malgré le terme 2^{-n} prometteur dans l'estimation de vitesse de convergence, la dichotomie est incomparablement moins efficace que la méthode de Héron. Avec 20 termes, on n'a obtenu que 5 ou 6 décimales exactes, alors que la méthode de Héron a déjà 11 décimales exactes rien qu'au cinquième terme. On peut estimer grossièrement que le 20ème terme de la méthode de Héron aura dans les 720900 décimales exactes. De ce point de vue, il n'y a pas photo entre les deux méthodes. Par contre, la méthode de Héron ne donne pas un encadrement de $\sqrt{2}$.

2.3 Vitesse de convergence

On peut quantifier plus précisément ces questions de vitesse de convergence.

Définition 2.3.1 Soit x_n une suite réelle qui converge vers $x \in \mathbb{R}$. On dit que la convergence est d'ordre (au moins) $\alpha \geq 1$ quand il existe $\lambda > 0$, avec $\lambda < 1$ quand $\alpha = 1$, et n_0 tels que pour tout $n \geq n_0$

$$|x_{n+1} - x| \leq \lambda|x_n - x|^\alpha.$$

Quand $\alpha = 1$, on dit que la convergence est linéaire. Quand $\alpha = 2$, on dit qu'elle est quadratique. Enfin, quand

$$\frac{|x_{n+1} - x|}{|x_n - x|} \rightarrow 0 \text{ quand } n \rightarrow +\infty$$

(en supposant $x_n \neq x$), on dit que la convergence est surlinéaire.

Bien sûr, une convergence d'ordre $\alpha > 1$ est automatiquement surlinéaire. On va se limiter ici aux convergences linéaires et quadratiques. Posons systématiquement $e_n = |x_n - x|$ qui représente l'erreur à l'étape n dans le contexte qui nous intéresse.

Proposition 2.3.2 Si la convergence est linéaire, alors on a pour n assez grand

$$e_n \leq C\lambda^n, \tag{2.3.1}$$

pour une certaine constante C (on rappelle que dans ce cas $\lambda < 1$). Si la convergence est quadratique, alors on a pour n assez grand

$$e_n \leq C\mu^{2^n}, \tag{2.3.2}$$

pour une autre constante C et pour un certain $\mu < 1$.

Démonstration. On considère $n \geq n_0$. Traitons d'abord la convergence linéaire. On remarque que $e_{n_0} \leq e_n$, ce qui n'est pas à proprement parler un scoop, mais va nous permettre démarrer un raisonnement par récurrence. On suppose donc en guise d'hypothèse de récurrence que $e_n \leq e_{n_0} \lambda^{n-n_0}$. On en déduit que

$$e_{n+1} \leq \lambda e_n \leq e_{n_0} \lambda^{n+1-n_0}.$$

On a ainsi établi (2.3.1), avec $C = e_{n_0} \lambda^{-n_0}$.

Pour la convergence quadratique, on a par hypothèse que $e_n \rightarrow 0$ quand $n \rightarrow +\infty$. Il existe donc n_1 tel que pour tout $n \geq n_1$, $\lambda e_n < 1$. On pose alors $m_0 = \max(n_0, n_1)$ et on part du même scoop, mais en m_0 . L'hypothèse de récurrence pour $n \geq m_0$ est ici $e_n \leq \frac{1}{\lambda}(\lambda e_{m_0})^{2^{n-m_0}}$, qui est bien satisfaite pour $n = m_0$. On en déduit que

$$e_{n+1} \leq \lambda(e_n)^2 \leq \lambda \left(\frac{1}{\lambda}(\lambda e_{m_0})^{2^{n-m_0}} \right)^2 = \frac{1}{\lambda}(\lambda e_{m_0})^{2 \times 2^{n-m_0}} = \frac{1}{\lambda}(\lambda e_{m_0})^{2^{n+1-m_0}}.$$

On a ainsi établi (2.3.2), avec $\mu = (\lambda e_{m_0})^{2^{-m_0}}$ et $C = \lambda^{-1}$. \diamond

Pour comparer les deux cas, on voit $\lambda < 1$ et $\mu < 1$, mais 2^n est incomparablement plus grand que n dès que n croît un peu. La convergence quadratique est donc beaucoup, beaucoup plus rapide que la convergence linéaire.

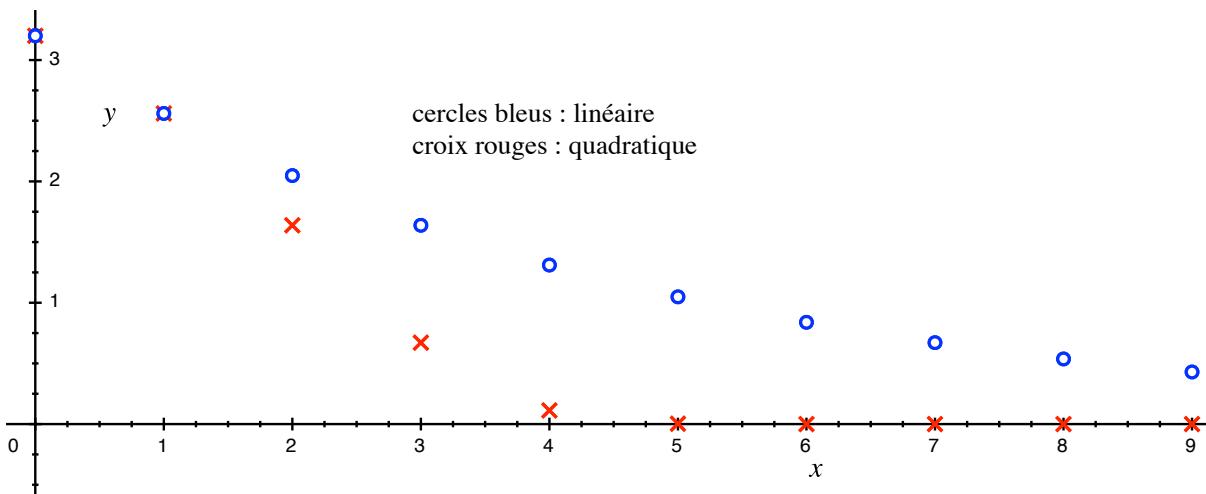


FIGURE 2.5 – Décroissance de l'erreur en fonction de n dans les deux cas.

Il est possible de quantifier un peu plus cela en s'intéressant au nombre de décimales exactes en fonction n . Supposons pour simplifier que l'on a déjà obtenu les chiffres situés à gauche de la virgule, c'est-à-dire que l'erreur e_n est strictement inférieure à 1. Notons que cela finit toujours par arriver pour n assez grand. Le nombre décimales exactes, c'est-à-dire le nombre de décimales de x_n qui coïncident avec celles de x est simplement le nombre de 0

consécutifs à partir et à droite de la virgule dans l'erreur.² Ainsi, pour la méthode de Héron

$$\begin{aligned}x_4 &= 1,41421356237469 \dots \\ \sqrt{2} &= 1,41421356237309 \dots \\ e_4 &= 0,00000000000160 \dots\end{aligned}$$

on a 11 décimales exactes au cinquième terme de la suite. Alors que pour la dichotomie

$$\begin{aligned}a_{19} &= 1,4142132 \dots \\ \sqrt{2} &= 1,4142135 \dots \\ e_{19} &= 0,0000003 \dots\end{aligned}$$

on a 6 décimales exactes au vingtième terme de la suite, et 5 décimales exactes avec b_{19} . Au cinquième terme, on a en tout et pour tout zéro ou une décimale exacte.

Plus précisément, on peut écrire de façon unique $e_n = 10^{-d_n} r_n$ avec $1 \leq r_n < 10$ et alors $D_n = d_n - 1$ est le nombre de décimales exactes. Par exemple, pour le dernier $e_{19} = 10^{-7} \times 3, \dots$. Si l'on prend le logarithme à base 10 de cette égalité, on obtient $\log_{10} e_n = -d_n + \log_{10} r_n$ avec $0 \leq \log_{10} r_n < 1$. On en conclut que $-d_n$ est la partie entière de $\log_{10} e_n$, donc

$$D_n = -[\log_{10} e_n] - 1.$$

Pour simplifier, on va raisonner grossièrement en oubliant la partie entière et le -1 et en considérant que le nombre de décimales exactes est juste $D_n = -\log_{10} e_n$. C'est une approximation qui n'est pas mauvaise. On obtient immédiatement :

Proposition 2.3.3 *Si la convergence est linéaire, alors on a pour n assez grand*

$$D_n \geq an + b, \quad (2.3.3)$$

pour certains a et b . Si la convergence est quadratique, alors on a pour n assez grand

$$D_n \geq a2^n + b, \quad (2.3.4)$$

pour certains a et b .

Si la convergence n'est pas plus que linéaire, on a aussi une majoration $D_n \leq a'n + b'$, c'est-à-dire que le nombre de décimales exactes croît de façon essentiellement affine par rapport à n . On retient que à quelques unités près, le nombre de décimales exactes augmente d'une valeur à peu près constante d'une itération à la suivante (ou aux suivantes si la croissance est faible).

De même, si la convergence n'est pas plus que quadratique, on a aussi une majoration $D_n \leq a'2^n + b'$, c'est-à-dire que le nombre de décimales exactes croît de façon essentiellement exponentielle en puissances de 2 par rapport à n . On retient que à peu de choses près, le nombre de décimales exactes double d'une itération à la suivante. C'est en pratique tellement rapide qu'il n'y a pas vraiment besoin d'aller chercher des méthodes d'ordre 3 ou plus.

2. Ceci n'est pas tout à fait vrai, à cause des bizarries de la numération décimale. Ainsi, si $x = 0,1$ et $x_n = 0,099$, alors $e_n = 0,001$, mais si on prend la phrase « décimales exactes » au pied de la lettre, on a 0 décimale exacte. Par contre, on a en fait trois décimales exactes par rapport au développement décimal impropre de x ... Tout ceci n'a pas un intérêt majeur, ceci dit.

Bien sûr, ce qui précède reste vrai dans toute base de numération, pas seulement en base dix, mais aussi en binaire, en octal, en hexadécimal, en sexagésimal, etc. en utilisant les logarithmes de même base.

Au vu de tout ce qui précède, on se doute bien que

Proposition 2.3.4 *La convergence de la méthode de dichotomie est linéaire.*

Démonstration. Bon, ce n'est pas tout à fait vrai, puisqu'on ne peut exclure *a priori* que la méthode tombe pile sur une racine en un nombre fini d'étapes. Néanmoins, c'est vrai au sens de l'estimation (2.2.3), avec $\lambda = \frac{1}{2}$. \diamond

En général, la dichotomie n'est pas mieux que linéaire. Notons qu'en pratique, on ne tombe jamais pile sur une racine. Par exemple, si celle-ci est irrationnelle et que l'on parte d'un intervalle de départ à extrémités rationnelles,³ les deux suites construites sont rationnelles.

On comprend mieux la comparaison en faveur de la méthode de Héron qui est elle quadratique, comme on le verra plus loin.

2.4 Quelques autres méthodes en vrac

Des méthodes que l'on ne fait que mentionner, sans les étudier particulièrement. La *méthode de fausse position* ou *regula falsi* trouve son origine dans l'Antiquité également pour résoudre des équations linéaires au départ, mais on peut aussi l'utiliser pour le problème $f(x) = 0$ avec f pas nécessairement affine. On part d'un intervalle $[a_0, b_0]$ tel que $f(a_0)f(b_0) < 0$, qui contient donc une racine. On prend la sécante au graphe de f qui joint les points $(a_0, f(a_0))$ et $(b_0, f(b_0))$ et l'on calcule l'abscisse c_0 de son intersection avec l'axe des x . C'est un nombre qui est dans l'intervalle ouvert $]a_0, b_0[$. On calcule ensuite $f(c_0)$. Si $f(a_0)f(c_0) < 0$, on pose $a_1 = a_0$ et $b_1 = c_0$, sinon on pose $a_1 = c_0$ et $b_1 = b_0$. Puis on recommence avec $[a_1, b_1]$ et c_1 , et ainsi de suite.

Il s'agit du même principe que la dichotomie, mais au lieu de prendre le milieu de l'intervalle, on prend l'abscisse d'intersection de la sécante avec l'axe des abscisses, voir Figure 2.6. En conséquence, les intervalles construits sont emboîtés et contiennent toujours une racine, donc en fournissent un encadrement. La méthode de fausse position converge toujours au sens où la suite c_k tend vers une racine.

Voici le calcul de c_k . L'équation de la sécante est celle de la droite du plan qui passe par les deux points indiqués,

$$Y = f(b_k) \frac{X - a_k}{b_k - a_k} + f(a_k) \frac{X - b_k}{a_k - b_k}.$$

Le point recherché c_k satisfait donc

$$0 = f(b_k) \frac{c_k - a_k}{b_k - a_k} + f(a_k) \frac{c_k - b_k}{a_k - b_k}.$$

C'est une équation du premier degré en l'inconnue c_k , qui n'est pas trop dure à résoudre,

$$c_k = \frac{f(a_k)b_k - f(b_k)a_k}{f(a_k) - f(b_k)}.$$

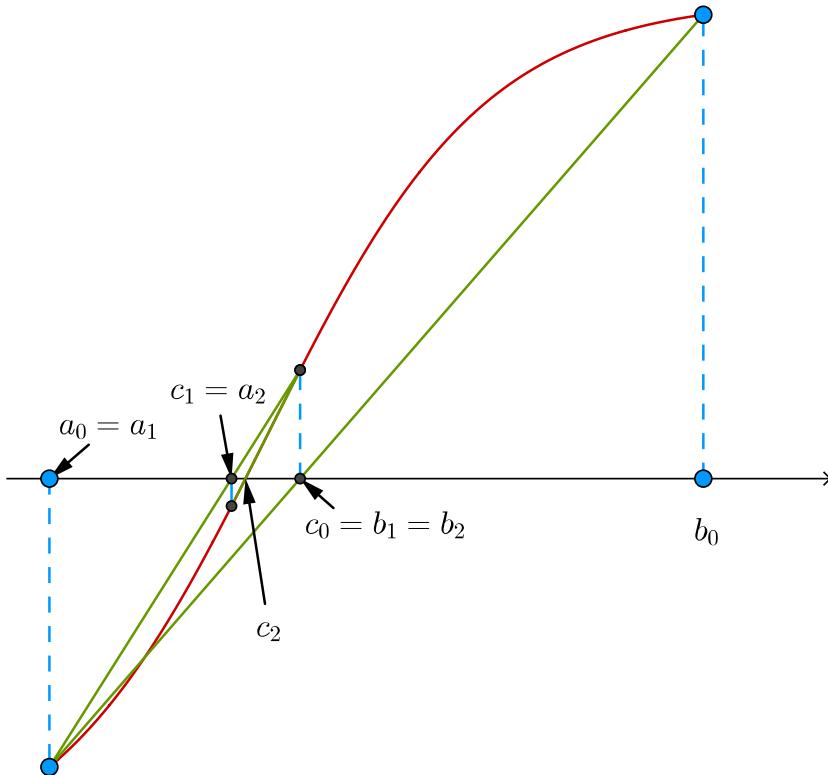


FIGURE 2.6 – Méthode de fausse position (zoomer à l'écran pour y voir de plus près).

En termes de vitesses de convergence, la méthode de fausse position n'est pas si facile que ça à analyser. Il y a des situations dans lesquelles elle est surlinéaire. Il y en a d'autres où elle n'est que linéaire, par exemple si f est concave ou convexe dans l'intervalle de travail, auquel cas, une des extrémités de l'intervalle ne change jamais. Elle ne converge alors pas mieux que la dichotomie, pour plus de calculs. Il y a même des cas où elle converge spectaculairement moins bien que la dichotomie.

Si f est de classe C^2 et que $f''(c) \neq 0$, alors par continuité cela reste vrai dans un voisinage de c , et on est en mauvaise posture pour la fausse position, cf. Figure 2.7. On sent bien que cela risque d'arriver souvent. Au contraire de la dichotomie, la longueur de l'intervalle ne tend alors même pas vers 0. Une seule de ses extrémités tend vers la racine.

Pour éviter cela, il existe une variante de la fausse position appelée *méthode de l'Illinois*. On commence par de la fausse position jusqu'à être coincé dans une situation similaire à celle de la Figure 2.7. Au lieu de continuer comme cela, on remplace $f(a_k)$ (dans le cas de cette figure, $f(b_k)$ si l'on est coincé de l'autre côté) par $\frac{1}{2}f(a_k)$. Cela donne un nouveau c_k ,

$$c_k = \frac{\frac{1}{2}f(a_k)b_k - f(b_k)a_k}{\frac{1}{2}f(a_k) - f(b_k)}.$$

Si ce nouveau c_k passe de l'autre côté de la racine, on le garde comme a_{k+1} et l'on continue

3. Ce qui, en pratique, est obligatoire.

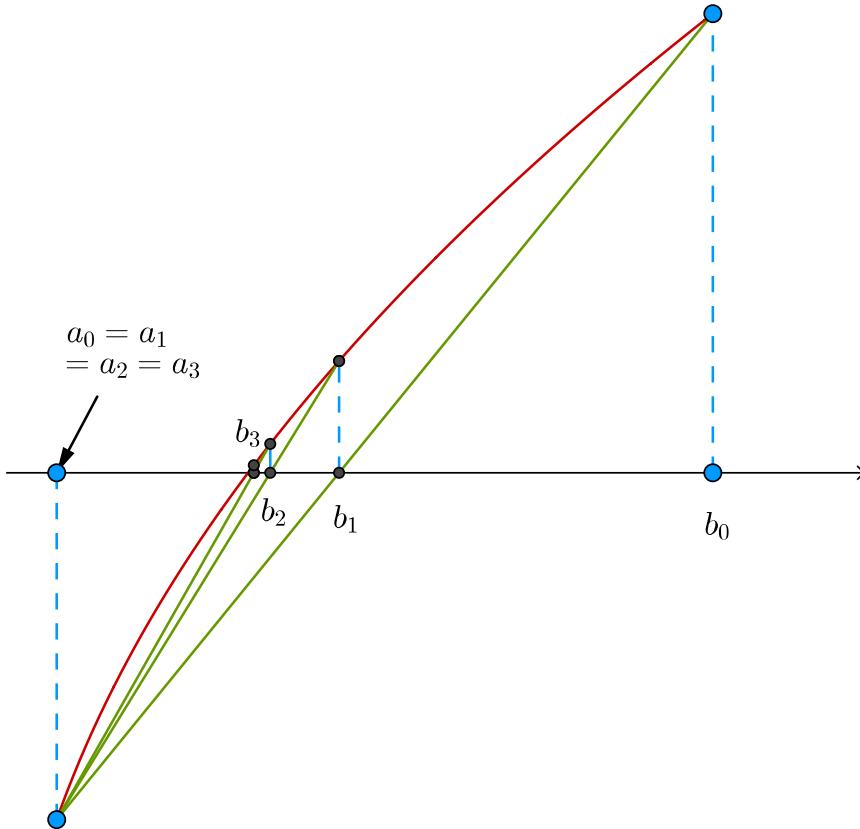


FIGURE 2.7 – Pas mieux que la dichotomie.

la fausse position. Si non, on redivise $\frac{1}{2}f(a_k)$ par 2, c'est-à-dire que l'on prend $\frac{1}{4}f(a_k)$, on calcule le nouveau c_k , et ainsi de suite. Il est à peu près clair sur le dessin que cette procédure s'arrête en un nombre fini d'étapes, voir Figure 2.8. En fait, on peut montrer que la méthode de l'Illinois est surlinéaire, converge toujours et produit des encadrements de la racine dont la longueur tend vers 0. C'est donc une variante très intéressante de la fausse position.

Dans la liste des méthodes en vrac, mentionnons la *méthode de la sécante*. Elle ressemble à la fausse position, sauf que l'on retient systématiquement les deux derniers points calculés comme base de la sécante, voir Figure 2.9. En conséquence de quoi, la méthode ne fournit pas de façon assurée un encadrement de la racine. De plus elle peut diverger complètement car il n'y a pas de condition de signe sur les valeurs de la fonction et l'on peut tomber sur une sécante pratiquement horizontale qui va intersecter l'axe des abscisses très, très loin. Du point de vue du calcul, elle se présente comme une récurrence à deux pas

$$x_k = \frac{f(x_{k-1})x_{k-2} - f(x_{k-2})x_{k-1}}{f(x_{k-1}) - f(x_{k-2})}.$$

On voit que rien n'empêche le dénominateur d'être tout petit, voire de s'annuler, auquel cas la méthode diverge brutalement.

Du point de vue de la vitesse de convergence, sous certaines hypothèses de dérivabilité

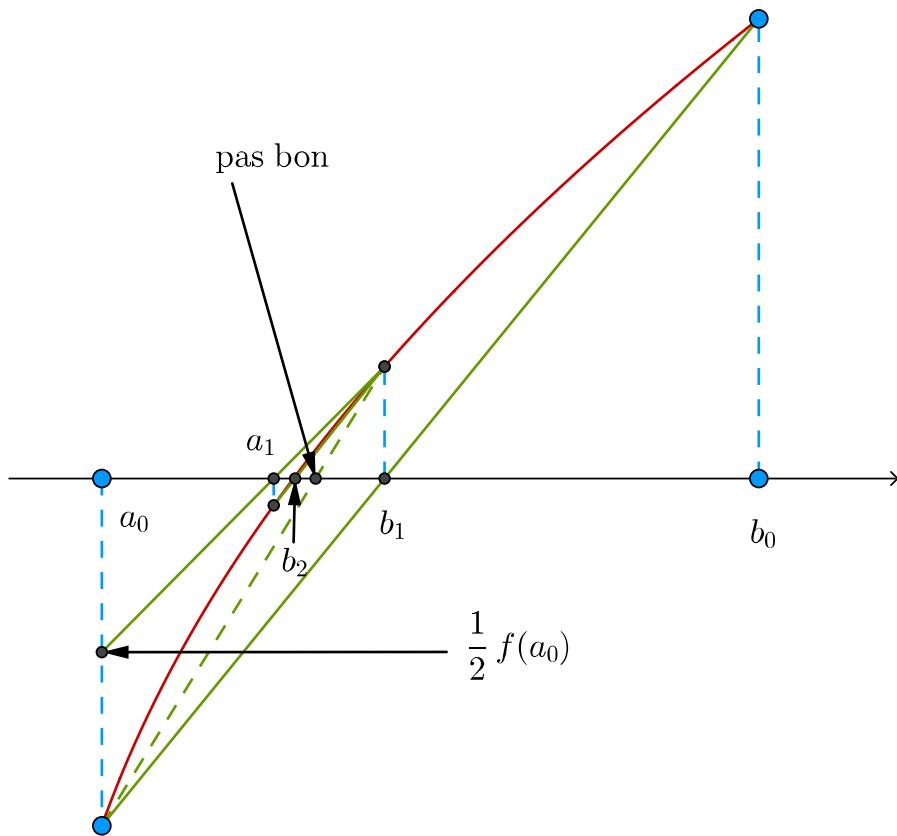


FIGURE 2.8 – La méthode de l'Illinois. Si cela n'avait pas marché avec $\frac{1}{2}f(a_0)$, on aurait recommencé avec $\frac{1}{4}f(a_0)$, $\frac{1}{8}f(a_0)$, etc.

sur f , on montre que si les points initiaux sont assez proches de la racine, alors la méthode est convergente, surlinéaire, d'ordre $\frac{1+\sqrt{5}}{2} \approx 1,618\dots$. Beaucoup plus rapide que linéaire, mais un peu moins que quadratique. Le nombre de décimales exactes est en gros multiplié par 1,6 à chaque itération.

2.5 Itérations de point fixe

Quelques souvenirs du L1 (c'était le bon temps). Tout problème de calcul de racine peut se reformuler en un problème de point fixe. Ainsi, si h est une fonction qui ne s'annule pas sur $[a, b]$, alors résoudre $f(x) = 0$ est équivalent à trouver un point fixe de $g(x) = x + h(x)f(x)$. Réciproquement, trouver un point fixe de g est équivalent à résoudre $f(x) = 0$ pour $f(x) = \frac{g(x)-x}{h(x)}$. Ce sont deux façons différentes de voir le même problème.

Un algorithme simple pour approcher un point fixe est l'algorithme d'itération. On prend un point x_0 dans $[a, b]$ et l'on définit $x_{n+1} = g(x_n)$ pour $n \geq 0$, à condition que $[a, b]$ soit un intervalle invariant par g , c'est-à-dire que $g([a, b]) \subset [a, b]$. Il est évident que si la suite x_n

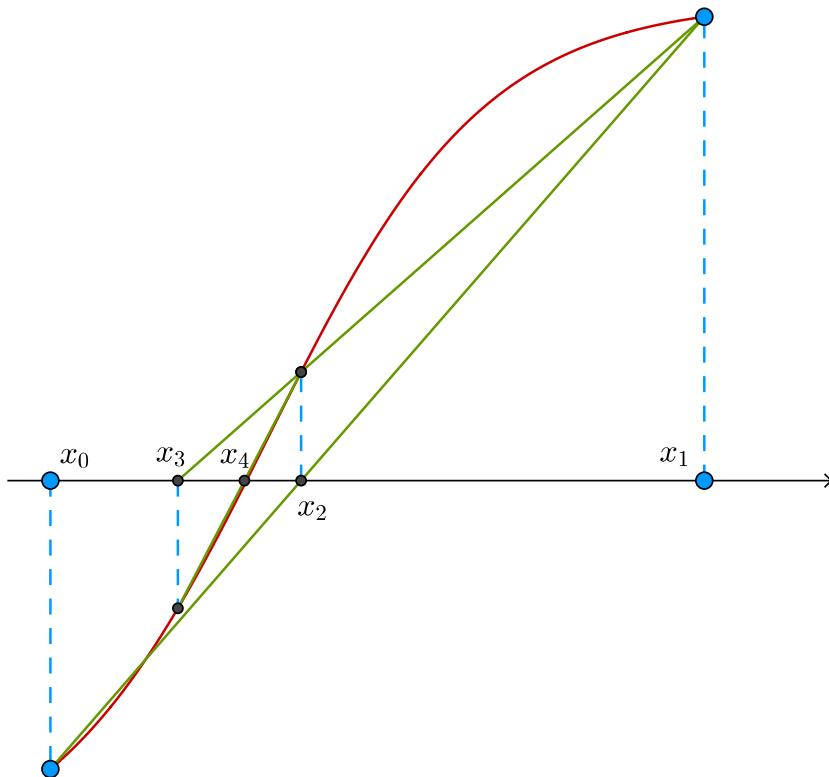


FIGURE 2.9 – Méthode de la sécante (zoomer à l'écran pour y voir de plus près).

converge, elle converge vers un point fixe.⁴

Définition 2.5.1 Soit g est une fonction de classe C^1 sur $[a, b]$, et c un point fixe de g . On dit que

1. c est un point fixe attractif si $|g'(c)| < 1$.
Cas particulier : $g'(c) = 0$. Si g est de classe C^2 , on dit que c est un point fixe super attractif.
2. c est un point fixe répulsif si $|g'(c)| > 1$.
3. on ne dit rien si $|g'(c)| = 1$.

Proposition 2.5.2 Si c est un point fixe attractif et x_0 est suffisamment proche de c , alors $x_n \rightarrow c$ avec convergence linéaire. Si c est super attractif, la convergence est quadratique. Si c est répulsif, une chose est sûre, c'est que $x_n \not\rightarrow c$, sauf si $x_0 = c$.

De plus, on sait que si g est croissante, la convergence a lieu de façon monotone, « en escalier », alors que si g est décroissante, elle a lieu « en spirale », ce qui veut dire en fait que les suites extraites d'ordre pair et impair sont toutes les deux monotones, de sens opposés, voir Figure 2.10.

4. On se rappelle que toutes les fonctions considérées dans ce cours sont au moins continues.

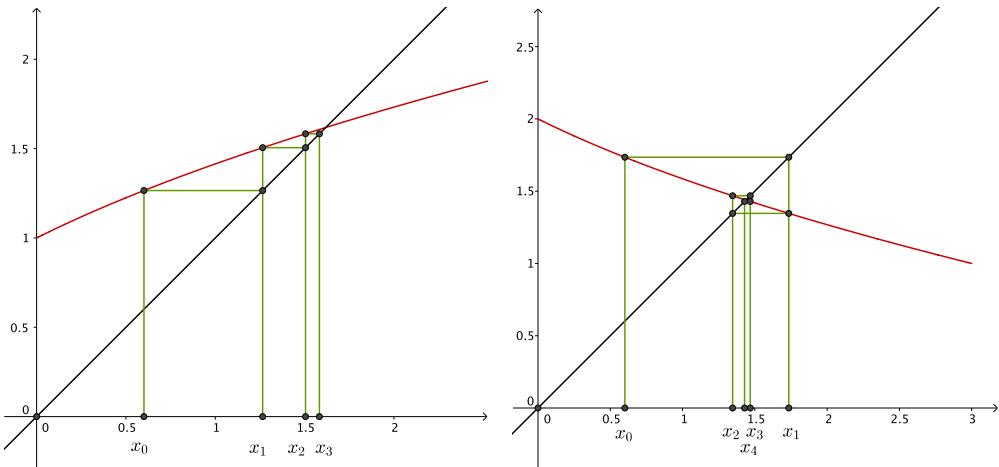


FIGURE 2.10 – Convergence d’itérations de point fixe.

Une façon possible de procéder pour trouver une racine est donc de transformer le problème en un problème de point fixe en s’arrangeant pour que la racine devienne un point fixe super attractif. C’est ce que fait en réalité la méthode de Newton, même si sa définition n’est pas celle d’un algorithme de point fixe.

2.6 La méthode de Newton

Soit c une racine de f , avec f au moins de classe C^1 sur l’intervalle $[a, b]$. On suppose qu’on connaît une valeur approchée x_0 de la racine, d’une façon ou d’une autre. L’idée de la méthode de Newton est de remplacer la courbe représentative de f par sa tangente en x_0 , d’équation

$$Y = f'(x_0)(X - x_0) + f(x_0),$$

et de considérer l’intersection de cette tangente avec l’axe des abscisses $Y = 0$, ce qui donne le point suivant

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}.$$

En général, si x_0 est choisi pas trop loin de la racine c de l’équation, on sent bien que x_1 en est une bien meilleure approximation que x_0 , cf. Figure 2.11. On recommence alors l’opération, ce qui conduit à la suite récurrente

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}.$$

Si cette suite est bien définie, i.e., si on n’a pas divisé par zéro en cours de route et n’est pas sorti de l’intervalle de définition de f , et si elle converge vers une valeur c , alors on a $c = c - \frac{f(c)}{f'(c)}$, soit $f(c) = 0$. Avec les notations précédentes, on a en fait écrit une itération de point fixe pour la fonction $g(x) = x - \frac{f(x)}{f'(x)}$, c’est-à-dire que l’on a choisi $h(x) = -\frac{1}{f'(x)}$.

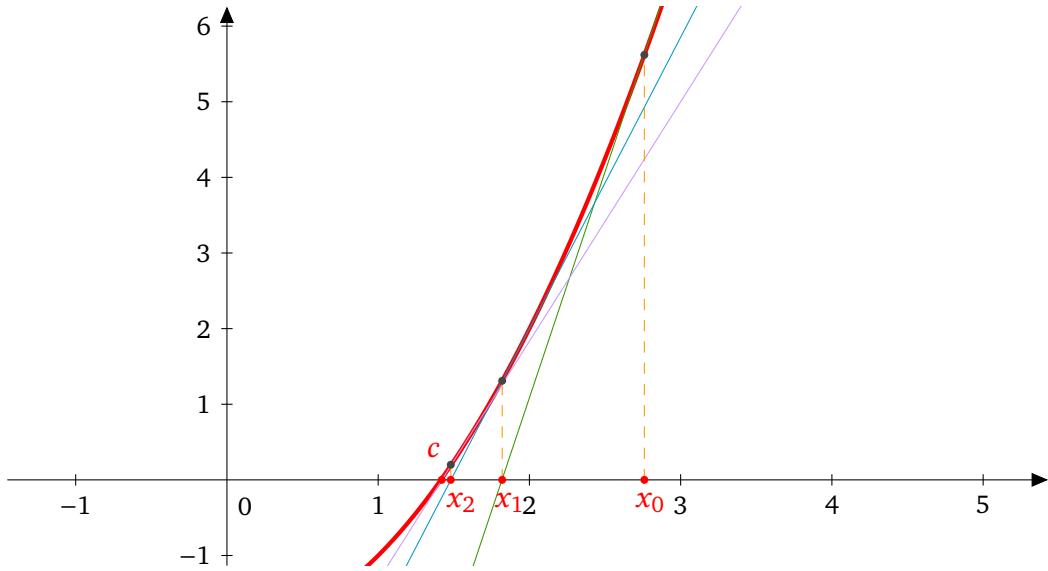


FIGURE 2.11 – La méthode de Newton (zoomer à l'écran pour y voir de plus près).

Comme $g'(x) = 1 - \frac{f'(x)^2 - f(x)f''(x)}{f'(x)^2} = \frac{f(x)f''(x)}{f'(x)^2}$, la racine est bien un point fixe super attractif de g si f' ne s'y annule pas.

Analysons plus précisément la méthode de Newton.

Théorème 2.6.1 *On suppose que f est de classe C^2 sur l'intervalle $I = [c - r, c + r]$, pour un certain $r > 0$ et que f' ne s'annule pas sur I . Soit*

$$M = \max_{x \in I} |f''(x)|, m = \min_{x \in I} |f'(x)| \text{ et } \alpha = \min\left(r, \frac{2m}{M}\right).$$

Alors pour tout point initial $x_0 \in]c - \alpha, c + \alpha[$, la suite de Newton x_n est bien définie pour tout n et converge vers c quand $n \rightarrow +\infty$, avec l'estimation

$$|x_n - c| \leq \frac{1}{K} (K|x_0 - c|)^{2^n}, \quad (2.6.1)$$

où $K = \frac{M}{2m}$.

Démonstration. La fonction f' ne s'annulant pas sur I , on peut y définir une fonction g par $g(x) = x - \frac{f(x)}{f'(x)}$. Par définition de la méthode de Newton, si x_n est bien défini, on a $x_{n+1} = g(x_n)$. La suite de Newton est donc celle des itérées de x_0 par g et il suffit par conséquent de montrer que g admet un intervalle invariant pour montrer que la suite est bien définie.

Pour tout $x \in I$, la formule de Taylor-Lagrange nous dit que

$$0 = f(c) = f(x) + (c - x)f'(x) + \frac{(c - x)^2}{2}f''(\xi),$$

pour un certain ξ situé entre c et x . Divisant par $f'(x)$, qui est non nul sur I , on en déduit que

$$\left(x - \frac{f(x)}{f'(x)}\right) - c = \frac{(c - x)^2}{2} \frac{f''(\xi)}{f'(x)}.$$

On voit donc que

$$|g(x) - c| \leq \frac{M}{2m} |x - c|^2.$$

On pose alors $\alpha = \min(r, \frac{2m}{M})$. Si $x \in]c - \alpha, c + \alpha[\subset I$, on a donc $|x - c| < \alpha$, d'où

$$|g(x) - c| \leq \frac{M}{2m} \alpha^2 \leq \alpha,$$

puisque $\alpha \leq \frac{2m}{M}$. On en déduit que $g(x) \in]c - \alpha, c + \alpha[\subset I$. On a trouvé un intervalle invariant par g et la suite de Newton x_n est donc bien définie si x_0 est pris dans cet intervalle invariant $]c - \alpha, c + \alpha[$.

De plus, on obtient pour tout n

$$|x_{n+1} - c| \leq K|x_n - c|^2,$$

avec $K = \frac{M}{2m}$, d'où l'estimation (2.6.1) par la Proposition 2.3.2 avec $m_0 = 0$ car $K|x_0 - \bar{x}| < 1$.
◊

Non seulement la méthode de Newton converge sous les hypothèses précédentes, mais elle converge quadratiquement. Par contre, la méthode de Newton demande plus de calculs que les précédentes méthodes, en particulier celui de f' , qui n'est pas forcément si simple que ça quand f est elle-même compliquée, et sa convergence n'est pas assurée pour toute valeur initiale de l'itération. Il vaut mieux être suffisamment près de la racine au départ, c'est-à-dire l'avoir déjà assez bien localisée, par exemple par dichotomie. En effet, le Théorème 2.6.1 donne bien un intervalle de convergence, mais celui-ci est centré sur la racine c inconnue. Si son existence est intéressante, l'utilité pratique de cet intervalle est donc discutable.

On voit par ailleurs que la constante K est d'autant plus grande que la dérivée seconde est grande et que la dérivée première est petite. Ce sont des situations défavorables pour la convergence de la méthode de Newton, à la fois en termes d'intervalle et en termes de vitesse de convergence. Il vaut mieux que la dérivée seconde soit petite (moralement, f très proche de son application affine tangente) avec une forte pente.

Gardons quand même à l'esprit qu'en pratique, quand elle marche, la méthode de Newton est terriblement efficace.

Prenons maintenant l'exemple de la fonction $f(x) = x^2 - 2$. On a $f'(x) = 2x$ donc la méthode de Newton consiste à effectuer l'itération

$$x_{n+1} = x_n - \frac{x_n^2 - 2}{2x_n} = \frac{x_n}{2} + \frac{1}{x_n}.$$

Quand on démarre à $x_0 = 1$, c'est exactement la méthode de Héron. Celle-ci est donc effectivement quadratique, de nombreux siècles avant Newton. Bien sûr, l'idée de Héron n'était pas du tout celle de Newton, mais se basait sur des considérations géométriques d'aires de rectangles, voir Figure 2.12.

Plus généralement, pour tout $A > 0$, on peut considérer la fonction $f(x) = x^2 - A$, qui fournit des itérations de Héron

$$x_{n+1} = x_n - \frac{x_n^2 - A}{2x_n} = \frac{x_n}{2} + \frac{A}{2x_n} = \frac{1}{2} \left(x_n + \frac{A}{x_n} \right),$$

lesquelles convergent quadratiquement vers \sqrt{A} .

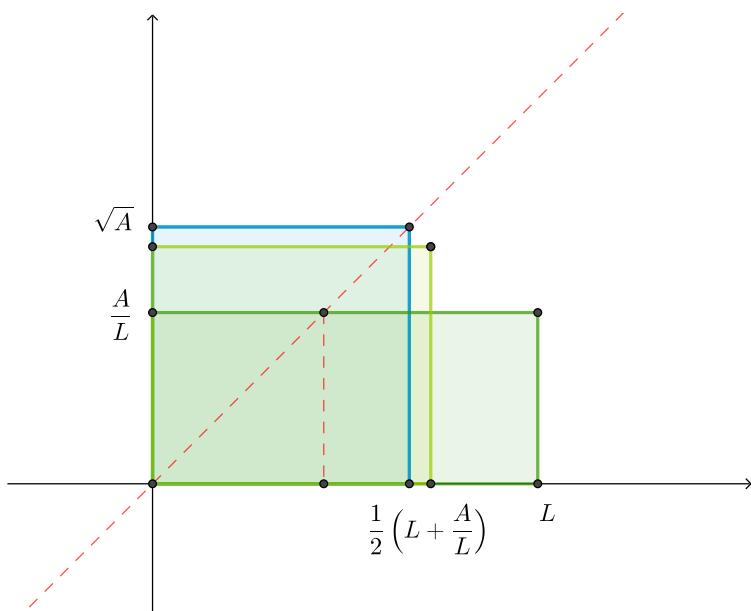


FIGURE 2.12 – La méthode de Héron du point de vue géométrique de Héron : le rectangle vert pâle a la même aire A que le rectangle vert foncé, mais est nettement plus carré, et ainsi de suite, et ainsi de suite...

Chapitre 3

$f(x) = 0$ en dimension quelconque

On va même commencer par se placer dans une généralité telle qu'il n'y a même pas a priori de notion de dimension.

3.1 Topologie

On connaît les espaces \mathbb{R}^n munis de leurs diverses normes usuelles, toutes équivalentes entre elles. Il s'agit de cas particuliers d'une notion topologique bien plus générale, celle des espaces métriques. Ce n'est pas la notion la plus générale en topologie, mais elle nous suffira ici.

Définition 3.1.1 Soit E un ensemble et $d: E \times E \rightarrow \mathbb{R}_+$ une application telle que

- i) $\forall (x, y) \in E^2, d(x, y) = d(y, x)$,
- ii) $d(x, y) = 0$ si et seulement si $x = y$,
- iii) $\forall (x, y, z) \in E^3, d(x, y) \leq d(x, z) + d(z, y)$,

est appelée une distance sur E . Le couple (E, d) est appelé un espace métrique.

Les trois propriétés i), ii) et iii), en particulier la troisième appelée *l'inégalité triangulaire*, sont des abstractions des propriétés de la distance physique de notre expérience de tous les jours, qui est la distance euclidienne dans \mathbb{R}^3 . L'inégalité triangulaire est simplement l'expression du fait qu'il est plus court d'aller directement de x à y que d'y aller en passant par z .

Quelques exemples :

1. La distance dite usuelle sur \mathbb{R} , celle que l'on utilise sans y penser spécialement, est simplement définie par $d(x, y) = |x - y|$.
2. Sur $\mathbb{R}^n, n \geq 1$, on connaît plusieurs normes : $\|x\|_1 = \sum_{i=1}^n |x_i|$, $\|x\|_2 = (\sum_{i=1}^n |x_i|^2)^{1/2}$, $\|x\|_\infty = \max_{1 \leq i \leq n} |x_i|$ (parmi tant d'autres). Chacune de ces normes donne naissance à une distance différente : $d_1(x, y) = \|x - y\|_1$, $d_2(x, y) = \|x - y\|_2$ qui est la distance euclidienne, $d_\infty(x, y) = \|x - y\|_\infty$, lesquelles donnent autant de structures d'espace métrique différentes sur \mathbb{R}^n : (\mathbb{R}^n, d_1) , (\mathbb{R}^n, d_2) , (\mathbb{R}^n, d_∞) . Ces distances sont équivalentes entre elles, c'est-à-dire qu'il existe des constantes $0 < \alpha_{ij} \leq \beta_{ij}$ telles que pour tous x et y dans \mathbb{R}^n , $\alpha_{ij}d_i(x, y) \leq d_j(x, y) \leq \beta_{ij}d_i(x, y)$ avec i et j quelconques dans $\{1, 2, +\infty\}$. Ceci provient de l'équivalence correspondante des normes (qui se lit ci-dessus avec $y = 0$). Toutes les notions topologiques qui suivront (continuité,

suites convergentes, etc.) seront en conséquence identiques. Toutes les normes sur un même espace vectoriel sur \mathbb{R} de dimension finie sont équivalentes.

3. Plus généralement, tout espace vectoriel normé $(E, \|\cdot\|_E)$ est automatiquement doté d'une structure d'espace métrique compatible avec sa norme en posant $d(x, y) = \|x - y\|_E$. C'est une distance qui est invariante par translation : $d(x+z, y+z) = d(x, y)$. Ainsi, l'espace $C^0([0, 1])$ des fonctions continues de l'intervalle $[0, 1]$ à valeurs dans \mathbb{R} muni de la norme naturelle pour cet espace, $\|f\|_{C^0} = \max_{t \in [0, 1]} |f(t)|$, est un espace métrique pour la distance $d(f, g) = \max_{t \in [0, 1]} |f(t) - g(t)|$. Il s'agit d'un espace vectoriel de dimension infinie. Il y a d'autres normes qui peuvent être utiles sur cet espace et qui ne sont pas équivalentes à la norme naturelle, comme $\|f\|_{L^1} = \int_0^1 |f(t)| dt$. La situation est donc plus compliquée qu'en dimension finie.
4. Si $f : \mathbb{R} \rightarrow \mathbb{R}$ est injective, alors $d(x, y) = |f(x) - f(y)|$ est une distance sur \mathbb{R} . Si $f(x) = x$ pour tout x , c'est la distance usuelle, sinon c'en est une autre. Si f n'est pas affine, cette distance ne provient pas d'une norme.
5. Si E est n'importe quel ensemble, $d(x, y) = 0$ si $x = y$, $d(x, y) = 1$ sinon, définit une distance sur E appelée *distance discrète*. En d'autres termes, n'importe quel ensemble peut être doté de cette structure d'espace métrique discret, mais cette structure n'est pas nécessairement très intéressante, sauf cas particulier.
6. Si (E, d) est un espace métrique, alors toute partie X de E peut être munie de la restriction de la distance à $X \times X$ et devenir ainsi automatiquement un autre espace métrique. On dit que c'est la distance *induite* par celle de E ou que (X, d) est un sous-espace métrique¹ de (E, d) .

Attention au fait qu'un espace métrique est bien un *couple* (ensemble, distance sur cet ensemble). Un même ensemble muni de distances différentes correspond à plusieurs espaces métriques différents, cf. l'exemple de \mathbb{R}^n avec différentes distances plus haut.

Dans un espace métrique (E, d) , on peut définir des notions topologiques comme les ouverts — ce sont les réunions quelconques de boules ouvertes $B(x, r) = \{y \in E; d(y, x) < r\}$, les fermés — ce sont les complémentaires des ouverts, les applications continues, etc. Des distances différentes peuvent parfaitement engendrer les mêmes ouverts. C'est le cas dans \mathbb{R}^n muni de ses distances usuelles. On dit alors qu'elles engendrent la même topologie. Mais le contraire peut parfaitement se produire également, comme dans l'exemple 3 plus haut où les deux distances engendrent des topologies différentes.

Voyons de plus près cette notion d'application continue f d'un espace métrique (E, d) à valeurs dans un autre espace métrique (F, δ) . Soit $a \in E$, on dit que f est *continue en a* si

$$\forall \varepsilon > 0, \exists \eta < 0; \forall x \in E, d(x, a) \leq \eta \Rightarrow \delta(f(x), f(a)) \leq \varepsilon.$$

En d'autres termes, on peut être assuré que $f(x)$ est proche de $f(a)$ au sens de la distance δ si l'on prend x suffisamment proche de a au sens de la distance d . C'est très intuitif. Dans le cas où $E = F = \mathbb{R}$ et $d = \delta$ est la distance usuelle, on retrouve exacte la continuité d'une fonction telle que définie en L1.

Bien sûr, on dit que f est continue sur E si f est continue en tout point a de E . C'est équivalent à la propriété que l'image réciproque par f de tout ouvert de F est un ouvert de

1. On ne distingue pas dans la notation d et sa restriction à $X \times X$.

E. La continuité est donc en fait une notion topologique, et pas seulement métrique.² En particulier sur \mathbb{R}^n muni de ses distances usuelles, la continuité ou non d'une application à valeurs dans \mathbb{R}^m muni aussi de ses distances usuelles, ne dépend pas du choix particulier de distance retenu. En conséquence, on prend la distance la plus pratique pour ce que l'on a à faire sur l'instant. Tout ce qui précède s'applique sans modification quand X est une partie de E munie de la distance induite et $f: X \rightarrow F$.

Dans un espace métrique, on peut également définir la *convergence* des suites comme on le fait dans \mathbb{R} muni de sa distance usuelle : une suite x_n de E , c'est-à-dire une application de \mathbb{N} dans E , converge vers un élément $x \in E$ si

$$\forall \varepsilon > 0, \exists n_0 \in \mathbb{N}, \forall n \geq n_0, d(x_n, x) \leq \varepsilon.$$

On appelle bien sûr x la *limite* de la suite x_n au sens de l'espace métrique (E, d) . On note dans ce cas $x_n \rightarrow x$ dans E quand $n \rightarrow +\infty$, ou bien $x = \lim_{n \rightarrow +\infty} x_n$ au sens de E , ou $x = \lim_{n \rightarrow +\infty} x_n$ sans rien préciser quand la distance est sous-entendue.³ Le contenu intuitif est très clair aussi : les valeurs prises par la suite x_n se rapprochent autant qu'on le souhaite de la limite x au sens de la distance d à condition que l'on prenne n assez grand.⁴

Cette limite est unique si elle existe. En effet, si l'on prend deux limites x et \bar{x} de la même suite, on obtient par l'inégalité triangulaire que pour tout $\varepsilon > 0$, $d(x, \bar{x}) \leq 2\varepsilon$, ce qui implique $d(x, \bar{x}) = 0$, ce qui implique $x = \bar{x}$.⁵ Dans le cas de l'espace $C^0([0, 1])$ muni de la distance indiquée un peu plus haut, on reconnaît dans la convergence au sens de cette distance simplement la convergence uniforme d'une suite de fonctions continues sur $[0, 1]$, vers une fonction continue sur $[0, 1]$.⁶

Il n'est pas très difficile de voir qu'une application $f: E \rightarrow F$ est continue en $a \in E$ si et seulement si, pour toute suite $x_n \in E$ qui tend vers a dans E au sens de la distance d , on a que $f(x_n)$ tend vers $f(a)$ dans F au sens de la distance δ , c'est-à-dire que $\delta(f(x_n), f(a)) \rightarrow 0$ pour toute suite x_n telle que $d(x_n, a) \rightarrow 0$, ce qui ramène à des convergences dans \mathbb{R}_+ .

Attention, *a priori* une suite dans un même ensemble peut très bien converger pour une distance et pas pour une autre. Encore pire, elle peut très bien converger vers une limite pour une distance et vers une autre limite différente pour une autre distance ! Tout cela est fondamentalement distance-dépendant.

Dans la pratique bien sûr, quand on a un problème spécifique à résoudre, il y a essentiellement toujours un choix de distance naturel qui s'impose. On ne s'amuse pas à fabriquer exprès des contre-exemples rien que pour le plaisir, mais il faut savoir que de tels contre-exemples existent.

2. Il en va de même pour la continuité en un point $a \in E$ qui s'exprime en termes purement topologiques, sans métrique, en disant que pour tout ouvert V de F contenant $f(a)$, il existe un ouvert de E contenant a dont l'image est incluse dans V . L'interprétation intuitive est exactement la même.

3. Attention quand même, il est loin d'être rare que l'on ait besoin d'avoir plusieurs distances non équivalentes sous la main en même temps sur un même ensemble.

4. La convergence d'une suite est également une notion qui est de nature purement topologique et s'exprime aussi uniquement en termes d'ouverts. Elle n'a pas besoin de distance, mais on se limite ici aux espaces métriques.

5. Plus généralement, un espace métrique est un espace séparé.

6. Notons que si l'on oublie le côté espace vectoriel normé, on sait quand même depuis longtemps que si une suite de fonctions continues converge uniformément vers une fonction, celle-ci est nécessairement continue. On peut remettre ce résultat élémentaire dans le cadre des espaces métriques en considérant l'espace (beaucoup) plus grand $B([0, 1])$ des fonctions bornées sur $[0, 1]$, muni de la même distance avec le max remplacé par un sup. Il dit alors que $C^0([0, 1])$ est un sous-espace vectoriel fermé de $B([0, 1])$.

Dans un espace métrique (E, d) , on dispose également de la notion de *suite de Cauchy*,⁷ exactement comme dans \mathbb{R} muni de sa distance usuelle :

$$\forall \varepsilon > 0, \exists n_0 \in \mathbb{N}, \forall n, m \geq n_0, d(x_n, x_m) \leq \varepsilon.$$

Évidemment, toute suite convergente est de Cauchy, mais la convergence ou non des suites de Cauchy est une propriété qu'un espace métrique possède ou ne possède pas.

Définition 3.1.2 *On dit qu'un espace métrique (E, d) est complet si toute suite de Cauchy y est convergente.*⁸

Les espaces \mathbb{R}^n munis de leurs distances usuelles sont complets. Par contre, \mathbb{Q} muni de la distance usuelle n'est pas complet, la suite des approximations de Héron de $\sqrt{2}$ étant de Cauchy et non convergente (dans \mathbb{Q} , puisque $\sqrt{2} \notin \mathbb{Q}$). L'espace $C^0([0, 1])$ muni de la distance de la convergence uniforme est complet. Par contre, il n'est pas complet si on le munit de la distance $d(f, g) = \int_0^1 |f(t) - g(t)| dt$ associée à la norme L^1 de l'exemple 3. Plus généralement, tout espace vectoriel normé qui est complet pour la distance associée à sa norme est appelé un *espace de Banach*.

Arrêtons là les généralités métriques.

3.2 Le théorème de point fixe de Banach

La notion de point fixe d'une application d'un ensemble dans lui-même a un sens dans n'importe quel ensemble, sans référence à aucune structure supplémentaire. Celle de racine de $f(x) = 0$, un peu moins. Il faudrait déjà qu'il y ait un 0 dans cet ensemble...

Définition 3.2.1 *Soit (E, d) un espace métrique et φ une application de E dans lui-même. On dit que φ est strictement contractante s'il existe $k \in [0, 1[$ tel que pour tout $(x, y) \in E^2$,*

$$d(\varphi(x), \varphi(y)) \leq kd(x, y).$$

Plus généralement, s'il existe un tel k sans la restriction $0 \leq k < 1$, on dit que φ est k -lipschitzienne. Si l'on ne souhaite pas préciser la constante, dite de Lipschitz, on dit lipschitzienne tout court. Attention, cette notion, comme celle d'application lipschitzienne, dépend de la distance. Une même application peut fort bien être une contraction stricte pour une distance et pas pour une autre.

On a le théorème de point fixe de Banach (ou de Picard, mais c'est plutôt Banach dans cette généralité) suivant :

Théorème 3.2.2 *Soit (E, d) un espace métrique complet et φ une application strictement contractante de E dans lui-même. Alors φ admet un point fixe unique x^* . De plus, pour tout $x_0 \in E$, la suite récurrente $(x_n)_{n \in \mathbb{N}}$ définie par $x_{n+1} = \varphi(x_n)$ pour tout $n \geq 0$ converge vers le point fixe x^* .*

7. Là, attention, ce n'est plus une notion exprimable en termes d'ouverts. Elle n'est pas topologique. La distance est bien utile. La raison en est simple : il est facile de construire deux distances sur un même ensemble qui engendrent les mêmes ouverts, mais qui n'ont pas les mêmes suites de Cauchy.

8. Naturellement, la complétude n'est pas non plus une notion topologique. Il faut quelque chose en plus que juste les ouverts, une distance pour ce qui nous concerne et sans aller trop loin.

Démonstration. Montrons d'abord l'unicité du point fixe. Soient $x^* \in E$ et $\tilde{x}^* \in E$ des points fixes de φ . On a donc $\varphi(x^*) = x^*$ et $\varphi(\tilde{x}^*) = \tilde{x}^*$. Comme φ est strictement contractante, on en déduit que $d(x^*, \tilde{x}^*) \leq kd(x^*, \tilde{x}^*)$, soit encore $(1 - k)d(x^*, \tilde{x}^*) \leq 0$. Comme $k < 1$, il s'ensuit que $1 - k > 0$, donc nécessairement $d(x^*, \tilde{x}^*) \leq 0$, c'est-à-dire en fait $d(x^*, \tilde{x}^*) = 0$, d'où $x^* = \tilde{x}^*$.⁹

Montrons ensuite son existence. Soit $x_0 \in E$ un point quelconque et x_n la suite de ses itérés par φ associée. On a alors

$$d(x_{n+1}, x_n) = d(\varphi(x_n), \varphi(x_{n-1})) \leq kd(x_n, x_{n-1}),$$

d'où par une récurrence immédiate (mais à faire quand même en exercice) $d(x_{n+1}, x_n) \leq k^n d(x_1, x_0)$ pour tout n . Pour tout entier $p > n$, il vient par l'inégalité triangulaire

$$d(x_p, x_n) \leq \sum_{i=n}^{p-1} d(x_{i+1}, x_i) \leq \left(\sum_{i=n}^{p-1} k^i \right) d(x_1, x_0).$$

Or

$$\sum_{i=n}^{p-1} k^i \leq \sum_{i=n}^{\infty} k^i = \frac{k^n}{1-k},$$

puisque $0 \leq k < 1$. On a donc finalement

$$d(x_p, x_n) \leq k^n \frac{d(x_1, x_0)}{1-k} \tag{3.2.1}$$

pour tout $p > n$ avec $0 \leq k < 1$, ce qui montre que la suite x_n est de Cauchy. Comme E est complet pour la distance d , la suite x_n converge vers une limite $x^* \in E$. Comme φ est contractante, on a $d(\varphi(x_n), \varphi(x^*)) \leq kd(x_n, x^*)$, donc $\varphi(x_n) \rightarrow \varphi(x^*)$ au sens de (E, d) . En passant à la limite dans l'égalité $x_{n+1} = \varphi(x_n)$ quand $n \rightarrow +\infty$, on obtient $x^* = \varphi(x^*)$ à cause de l'unicité de la limite d'une suite dans un espace métrique. ◇

La condition de contraction stricte, c'est-à-dire $k < 1$, est cruciale. Ainsi, l'application $\mathbb{R} \rightarrow \mathbb{R}$, $x \mapsto \sqrt{1+x^2}$ est contractante, ou encore 1-lipschitzienne pour la distance usuelle sur \mathbb{R} , mais n'a pas de point fixe. Par conséquent, $k \leq 1$ ne suffit pas pour assurer l'existence d'un point fixe sous les autres hypothèses du théorème de Banach.

Corollaire 3.2.3 *On a l'estimation d'erreur*

$$d(x^*, x_n) \leq k^n d(x^*, x_0). \tag{3.2.2}$$

Démonstration. Reprenons la propriété de contraction

$$d(x^*, x_n) = d(\varphi(x^*), \varphi(x_{n-1})) \leq kd(x^*, x_{n-1}),$$

d'où par une récurrence immédiate (mais à faire quand même en exercice) $d(x^*, x_n) \leq k^n d(x^*, x_0)$ pour tout n . ◇

9. La complétude de E ne joue aucun rôle pour l'unicité.

Comme $k < 1$, ceci montre que la convergence des itérations de point fixe de Banach est une convergence linéaire pour la distance d .

Mentionnons un point anecdotique, qui est que l'on peut se faire une idée de combien on a raté le point fixe au départ en regardant les deux premiers termes de la suite. En effet, on a

$$\frac{d(x_1, x_0)}{1+k} \leq d(x^*, x_0) \leq \frac{d(x_1, x_0)}{1-k}.$$

L'inégalité de droite est juste l'estimation (3.2.1) pour $n = 0$ après avoir fait $p \rightarrow +\infty$. En remplaçant dans (3.2.2), on obtient d'ailleurs l'estimation

$$d(x^*, x_n) \leq k^n \frac{d(x_1, x_0)}{1-k}. \quad (3.2.3)$$

dans laquelle tous les termes de droite sont (en principe) calculables. Pour l'inégalité de gauche, il suffit d'appliquer l'inégalité triangulaire $d(x_0, x_1) \leq d(x_0, x^*) + d(x^*, x_1) \leq (1+k)d(x_0, x^*)$.

La convergence est assurée pour tout choix de x_0 , mais elle est bien sûr d'autant plus rapide que x_0 est proche de x^* .

Comme l'estimation (3.2.3), est explicite, on peut quantifier le nombre d'itérations qui permet d'assurer une précision donnée à l'avance au sens de la distance d . En effet, pour être sûr que $d(x^*, x_n) \leq \eta$ pour un certain $\eta > 0$, il suffit de prendre

$$n \geq \frac{\ln(\frac{\eta(1-k)}{d(x_1, x_0)})}{\ln k}.$$

Cela n'a pas grand sens de parler de nombre de décimales exactes dans ce contexte, mais on peut, si l'on veut, compter le nombre de zéros à droite de la virgule dans le terme de droite de (3.2.2).

Mentionnons un corollaire parfois utile du théorème 3.2.2.

Corollaire 3.2.4 Soit (E, d) un espace métrique complet et $\varphi: E \rightarrow E$ une application telle qu'il existe $p \in \mathbb{N}$ pour lequel $\varphi^{\circ p}$ est strictement contractante. Alors φ admet un point fixe unique x^* .

Ici, la notation $\varphi^{\circ p}$ désigne $\varphi \circ \varphi \circ \dots \circ \varphi$, p fois.

Démonstration. On sait que $\varphi^{\circ p}$ admet un point fixe unique x^* . Posons $y^* = \varphi(x^*)$. Il s'ensuit que $\varphi^{\circ p}(y^*) = \varphi^{\circ p}(\varphi(x^*)) = \varphi(\varphi^{\circ p}(x^*)) = \varphi(x^*) = y^*$, lequel donc aussi point fixe de $\varphi^{\circ p}$. Donc $x^* = y^* = \varphi(x^*)$, par unicité de ce point fixe. Naturellement, tout point fixe de φ étant aussi point fixe de $\varphi^{\circ p}$, il y aussi unicité de point fixe pour φ . \diamond

En fait, on a montré un résultat purement ensembliste, sans le moindre gramme de topologie dedans : si $\varphi^{\circ p}$ admet un point fixe *unique*, alors φ admet le même point fixe (unique aussi).

Dans le cas du corollaire, c'est-à-dire quand $\varphi^{\circ p}$ est strictement contractante etc., notons que les itérés de n'importe quel $x_0 \in E$ par φ convergent encore vers x^* . En effet, la suite $x_k^0 = \varphi^{\circ kp}(x_0)$ converge vers x^* . Il en va de même de la suite $x_k^1 = \varphi^{\circ(kp+1)}(x_0)$, qui itère $\varphi^{\circ p}$ à partir de $\varphi(x_0)$. Et ainsi de suite, pour toutes les suites $x_k^m = \varphi^{\circ(kp+m)}(x_0)$, $0 \leq m \leq p-1$.

On a ainsi couvert toutes les classes d'entiers modulo p , soit en fait tous les entiers, et on en déduit immédiatement que $\varphi^{\circ n}(x_0) \rightarrow x^*$ quand $n \rightarrow +\infty$.

Il faut remarquer que si $\varphi^{\circ p}$ est contractante donc continue, φ elle-même n'a aucune raison d'être continue. Ainsi, $\varphi(x) = 1$ pour $-1 \leq x < 0$, $\varphi(x) = 0$ pour $0 \leq x \leq 1$, est telle que $\varphi^{\circ 2}$ est strictement contractante sur $E = [-1, 1]$ muni de la distance usuelle (et peut-être même de n'importe quelle autre distance... 😊).

Il va de soi qu'une application $\varphi: E \rightarrow E$ qui n'a pas de point fixe, ou qui a plusieurs points fixes, ne peut être strictement contractante pour aucune distance d sur E qui rende E complet.

Pour conclure, le théorème de point fixe de Banach, présenté ici hors sol et sans la moindre motivation, se trouve être un outil puissant de démonstration d'existence d'objets intéressants. Par exemple, il permet de démontrer l'existence et l'unicité de la solution du problème de Cauchy pour une équation différentielle du premier ordre générale (linéaire à coefficients variables ou non linéaire), sous des hypothèses appropriées. Il permet aussi de démontrer le théorème d'inversion locale. Mais ce n'est pas notre propos ici.

3.3 La méthode de Newton dans \mathbb{R}^n

On va s'intéresser ici à trouver les racines ou les zéros d'une application $f: \mathbb{R}^n \rightarrow \mathbb{R}^n$, ou d'une partie de \mathbb{R}^n à valeurs dans \mathbb{R}^n , $f(x) = 0$. Cela a un sens, puisque \mathbb{R}^n contient bien un 0. On a déjà vu le cas $n = 1$.

Première remarque, pourquoi $f: \mathbb{R}^n \rightarrow \mathbb{R}^n$ et pas plus généralement $f: \mathbb{R}^n \rightarrow \mathbb{R}^m$ avec m pas nécessairement égal à n ? La raison en est qu'il n'est raisonnable d'espérer trouver des zéros isolés que dans le cas $n = m$. Généralement, si $m < n$ on va trouver soit l'ensemble vide, soit des zéros qui forment une « grosse » partie de \mathbb{R}^n . Par exemple, si $m = 1$ et $n = 2$, on va typiquement tomber sur une courbe, si $m = 1$ et $n = 3$ sur une surface, etc. Par contre, si $m > n$, on va généralement trouver l'ensemble vide.

Pour se convaincre que ceci est plausible, il suffit de regarder le cas où f est affine ou linéaire. On est ici en train de discuter de la résolution d'un système linéaire de m équations à n inconnues. L'ensemble des zéros est soit un singleton, soit un espace affine de dimension supérieure à 1, soit l'ensemble vide, selon le rang, le noyau et l'image de l'application linéaire associée, et le second membre du système linéaire. La situation est assez semblable dans le cas général non linéaire, avec plus de variété¹⁰ naturellement, et il est impossible de conclure de façon générale et universelle comme avec les systèmes linéaires.

Deuxième remarque, \mathbb{R}^n n'est pas naturellement ordonné pour $n > 1$, donc toutes les méthodes fondées sur des questions de signe ou de théorème des valeurs intermédiaires, comme la dichotomie ou la fausse position, n'ont que très peu de chances d'avoir des généralisations en dimension supérieure. Il faut plutôt regarder du côté de la méthode de la sécante ou de la méthode de Newton. La méthode de la sécante se généralise en dimension supérieure, mais pas de façon très simple. Par contre, la méthode de Newton ne change pas tant que ça entre $n = 1$ et $n > 1$.

¹⁰. Private joke liée à la géométrie différentielle.

3.3.1 Rappels (?) de calcul différentiel

Faisons d'abord quelques rappels de calcul différentiel. D'accord, c'est un prérequis, mais une petite révision expresse ne peut pas faire de mal. On se place systématiquement dans des espaces vectoriels E sur \mathbb{R} , avec des notions analogues sur \mathbb{C} ou sur d'autres corps de nombres plus exotiques, munis d'une norme, c'est-à-dire d'une application de E dans \mathbb{R}_+ qui est positivement homogène, satisfait l'inégalité triangulaire et ne s'annule que sur le vecteur nul. Avec cette notion de norme viennent comme on l'a vu des notions de distance, de boules, d'ouverts, de fermés et de continuité. Notons que, pour toute norme, toutes les boules pour la distance associée sont convexes.

Une application f d'un ouvert U d'un espace vectoriel normé E à valeurs dans un autre espace vectoriel normé F est différentiable en un point $x \in U$ s'il existe une application linéaire continue de E dans F , notée df_x telle que l'on puisse écrire

$$f(x + h) = f(x) + df_x h + \|h\|_E \varepsilon(h),$$

pour tout $h \in E$ tel que $x + h \in U$, où ε est une application de E dans F telle que

$$\|\varepsilon(h)\|_F \rightarrow 0 \text{ quand } \|h\|_E \rightarrow 0.$$

L'application linéaire df_x est appelée la *déférentielle* (de Fréchet) de f au point x . Le fait que cette application linéaire soit continue se traduit par l'existence d'une constante C telle que $\|df_x h\|_F \leq C\|h\|_E$ pour tout $h \in E$. On voit que l'accroissement $f(x + h) - f(x)$ se comporte principalement comme le terme linéaire $df_x h$, le terme suivant $\|h\|_E \varepsilon(h)$ se comportant comme un reste négligeable devant le terme linéaire quand $\|h\|_E$ est petit (sauf si le terme linéaire en question est nul...). Il est bien clair que si f est différentiable en x , alors elle est continue en x .

Quand f est différentiable en tout point $x \in U$, on dit qu'elle est différentiable sur U . Quand l'application $U \rightarrow \mathcal{L}(E; F)$, $x \mapsto df_x$, est elle-même continue¹¹, on dit que f est continûment différentiable ou de classe C^1 . Quand $E = F = \mathbb{R}$, alors df_x est l'application linéaire de \mathbb{R} dans \mathbb{R} qui consiste à multiplier par la dérivée de f au point x : $df_x h = f'(x)h$. Ne pas confondre la continuité de la différentielle en x , $h \mapsto df_x h$, qui est automatique ici en dimension finie, et la continuité de $x \mapsto df_x$ de \mathbb{R} dans $\mathcal{L}(\mathbb{R}; \mathbb{R})$, c'est-à-dire la classe C^1 , qui n'est ici autre que la continuité de la fonction f' de \mathbb{R} dans \mathbb{R} . En effet,

$$\|df_x - df_y\|_{\mathcal{L}(\mathbb{R}; \mathbb{R})} = \sup\{|df_x h - df_y h|; |h| \leq 1\} = |f'(x) - f'(y)|.$$

À toutes fins utiles, on rappelle qu'une dérivée partielle n'a rien de bien méchant. C'est juste une dérivée ordinaire par rapport à une des variables quand on fixe toutes les autres.

En dimension finie (de l'espace de départ), toute application linéaire est automatiquement continue, donc il est inutile de se focaliser sur la continuité de df_x , qui est offerte gratuitement dans ce cas. Ce n'est qu'en dimension infinie qu'il faut payer un supplément pour cela. De plus, toujours en dimension finie, après un choix de base dans chacun des deux espaces vectoriels de départ et d'arrivée, la matrice qui représente la différentielle d'une application dans ces bases est appelée sa *matrice jacobienne*. Ses coefficients sont les dérivées partielles des différentes composantes.

¹¹. L'espace $\mathcal{L}(E; F)$ étant lui-même un espace vectoriel normé, ceci a bien un sens. La norme naturelle sur cet espace est $\|g\|_{\mathcal{L}(E; F)} = \sup\{\|g(x)\|_F; \|x\|_E \leq 1\}$.

Plus explicitement, soit $f: U \rightarrow F$ une application de U ouvert d'un espace vectoriel normé E de dimension k à valeurs dans un espace vectoriel normé F de dimension m . On suppose f différentiable au point $x_0 \in U$. Sa différentielle en x_0 est une application linéaire df_{x_0} de E dans F . Si l'on choisit une base $(u_j)_{j=1,\dots,k}$ de E et une base $(v_i)_{i=1,\dots,m}$ de F , et que l'on note (x_j) les coordonnées cartésiennes associées dans E et (y_i) les coordonnées cartésiennes associées dans F , alors l'application f est représentée par m applications coordonnées f_i de l'ouvert de \mathbb{R}^k contenant les coordonnées des points de U , à valeurs dans \mathbb{R} , de telle sorte que

$$f(x) = \sum_{i=1}^m f_i(x_1, x_2, \dots, x_k) v_i, \text{ où } x = \sum_{j=1}^k x_j u_j.$$

La différentielle df_{x_0} de f en x_0 est alors représentée dans ces bases par la matrice jacobienne $\nabla f(x_0)$, matrice $m \times k$ dont les coefficients sont donnés par $(\nabla f(x_0))_{ij} = \frac{\partial f_i}{\partial x_j}(x_0)$, $i = 1, \dots, m$, $j = 1, \dots, k$. Cette représentation a lieu au sens usuel de l'algèbre linéaire, c'est-à-dire que pour tout vecteur $h = \sum_{j=1}^k h_j u_j$ de E , on a

$$df_{x_0} h = \sum_{i=1}^m (df_{x_0} h)_i v_i \text{ avec } (df_{x_0} h)_i = \sum_{j=1}^k (\nabla f(x_0))_{ij} h_j = \sum_{j=1}^k \frac{\partial f_i}{\partial x_j}(x_0) h_j.$$

On reconnaît un simple produit matrice-vecteur. C'est tout-à-fait normal, car si A est la matrice $m \times k$ qui représente l'application linéaire df_{x_0} dans ces bases, on a $A_{ij} = (df_{x_0} u_j)_i$ avec

$$f(x_0 + tu_j) = f(x_0) + t df_{x_0} u_j + |t| \|u_j\|_E \varepsilon(|t|),$$

d'où

$$\begin{aligned} (df_{x_0} u_j)_i &= \left(\frac{f(x_0 + tu_j) - f(x_0)}{t} \right)_i - (\|u_j\|_E \varepsilon(|t|))_i \\ &= \frac{f_i(x_0 + tu_j) - f_i(x_0)}{t} - \|u_j\|_E \varepsilon_i(|t|) \rightarrow \frac{\partial f_i}{\partial x_j}(x_0) \end{aligned}$$

quand $t \rightarrow 0$ par définition de ce qu'est une dérivée partielle. On voit bien sûr que le fait que f soit différentiable en x_0 implique que toutes ces dérivées partielles existent en x_0 .

La composée de deux applications différentiables est différentiable. Si $f: U \rightarrow F$ est différentiable en $x_0 \in U \subset E$ et $g: V \rightarrow G$ est différentiable en $f(x_0) \in V \subset F$, U et V ouverts de leur espace respectif tels que $f(U) \subset V$, alors $g \circ f: U \rightarrow G$ est différentiable en x_0 et sa différentielle est la composée des différentielles de f et g , $d(g \circ f)_{x_0} = dg_{f(x_0)} \circ df_{x_0}$.

Avec des choix de bases dans les trois espaces vectoriels, comme la matrice de la composée de deux applications linéaires est le produit de leurs matrices (dans le même ordre), on en déduit pour les matrices jacobiennes $\nabla(g \circ f)(x_0) = \nabla g(f(x_0)) \nabla f(x_0)$.

En explicitant tout cela avec des dérivées partielles, on obtient la très importante formule de dérivation des fonctions composées de plusieurs variables, qu'il faut absolument savoir appliquer quelles que soient les circonstances, même les plus adverses,

$$\frac{\partial(g \circ f)_l}{\partial x_j}(x_0) = \sum_{i=1}^m \frac{\partial g_l}{\partial y_i}(f(x_0)) \frac{\partial f_i}{\partial x_j}(x_0),$$

pour $j = 1, \dots, k$ et $l = 1, \dots, n$ où n est la dimension de G .¹² C'est une application

¹². En fait, on a seulement besoin de se rappeler du cas $n = 1$, manifestement.

immédiate de la formule générale donnant les coefficients d'un produit matriciel en fonction des coefficients des matrices dont on effectue le produit. C'est de l'algèbre linéaire en fait (dont on ne saurait trop rappeler combien elle est fondamentale).

On a bien sûr la même chose avec la différentiabilité en tout point et avec la classe C^1 pour les fonctions composées.

Quand $\dim E = 1$, on rappelle que le théorème des accroissements finis, et plus généralement la formule de Taylor avec reste de Taylor-Lagrange, sont faux dès que $\dim F > 1$. On les remplace par des inégalités du même nom. L'inégalité des accroissements finis, avec $f : [a, b] \rightarrow F$,

$$\|f(b) - f(a)\|_F \leq \sup_{t \in [a, b]} \|f'(t)\|_F (b - a),$$

et l'inégalité de Taylor-Lagrange,

$$\left\| f(b) - \sum_{i=0}^n \frac{(b-a)^i}{i!} f^{(i)}(a) \right\|_F \leq \sup_{t \in [a, b]} \|f^{(n+1)}(t)\|_F \frac{(b-a)^{n+1}}{(n+1)!},$$

sous les hypothèses adéquates sur f . Les formules avec reste intégral restent par contre vraies

$$f(b) - f(a) = \int_a^b f'(t) dt,$$

et

$$f(b) - \sum_{i=0}^n \frac{(b-a)^i}{i!} f^{(i)}(a) = \int_a^b \frac{(t-a)^n}{n!} f^{(n+1)}(t) dt,$$

avec une notion adéquate d'intégrale à valeurs vectorielles, c'est-à-dire en dimension finie, en intégrant composante par composante.

Quand $\dim E > 1$, on se ramène à la dimension 1 en se plaçant sur des segments, à condition que ces segments restent entièrement dans U . Notons que l'inégalité des accroissements finis nous donne un moyen pratique de vérifier la condition de contraction stricte nécessaire pour pouvoir appliquer le théorème 3.2.2 de point fixe de Banach dans le cas de \mathbb{R}^n , que l'on munit de la norme que l'on préfère.

Proposition 3.3.1 Soit $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$ une application de classe C^1 telle qu'il existe une boule $B(a, r)$ telle que

$$\sup_{x \in \bar{B}(a, r)} \|df_x\|_{\mathcal{L}(\mathbb{R}^n; \mathbb{R}^n)} = k < 1,$$

et

$$kr + \|f(a) - a\| \leq r,$$

alors $f(\bar{B}(a, r)) \subset \bar{B}(a, r)$, f y est strictement contractante.

Ici $\bar{B}(a, r) = \{x \in E; \|x - a\| \leq r\}$ désigne la boule fermée de centre a et de rayon r . On rappelle que

$$\|df_x\|_{\mathcal{L}(\mathbb{R}^n; \mathbb{R}^n)} = \sup_{\substack{u \in \mathbb{R}^n \\ \|u\| \leq 1}} \|df_x u\|.$$

Démonstration. Pour tous $x, y \in \bar{B}(a, r)$, on pose $h: [0, 1] \rightarrow \mathbb{R}^n$, $h(t) = f(x + t(y - x))$ de telle sorte que $h(0) = f(x)$ et $h(1) = f(y)$. Par dérivation des fonctions composées, h est de classe C^1 , avec

$$h'(t) = df_{x+t(y-x)}(y - x).$$

Les propriétés de norme d'application linéaire impliquent que

$$\|h'(t)\| \leq \|df_{x+t(y-x)}\|_{\mathcal{L}(\mathbb{R}^n; \mathbb{R}^n)} \|y - x\|.$$

Le segment $t \mapsto x + t(y - x)$ est entièrement contenu dans la boule qui est convexe, on en déduit donc que

$$\|h'(t)\| \leq k\|y - x\|.$$

L'inégalité des accroissements finis implique alors

$$\|f(y) - f(x)\| = \|h(1) - h(0)\| \leq \sup_{t \in [0,1]} \|h'(t)\|(1 - 0) \leq k\|y - x\|,$$

d'où la contraction stricte.

On utilise alors la deuxième hypothèse. Pour tout $x \in \bar{B}(a, r)$, c'est-à-dire $\|x - a\| \leq r$,

$$\begin{aligned} \|f(x) - a\| &= \|f(x) - f(a) + f(a) - a\| \leq \|f(x) - f(a)\| + \|f(a) - a\| \\ &\leq k\|x - a\| + \|f(a) - a\| \leq kr + \|f(a) - a\| \leq r, \end{aligned}$$

ce qui montre que $f(x) \in \bar{B}(a, r)$. \diamond

Corollaire 3.3.2 *Si f satisfait les hypothèses de la proposition précédente, alors f admet un point fixe unique dans la boule $\bar{B}(a, r)$.*

Démonstration. En effet, \mathbb{R}^n est complet pour toute distance induite par une norme et une boule fermée est un fermé pour la topologie métrique associée à la norme. Elle est donc complète pour la distance en question. Ceci découle du fait que toute suite de Cauchy dans $\bar{B}(a, r)$ est aussi de Cauchy dans \mathbb{R}^n , puisqu'il s'agit de la même distance. Elle a donc une limite dans \mathbb{R}^n , mais comme $\bar{B}(a, r)$ est fermé, cette limite est dans $\bar{B}(a, r)$. Le théorème de point fixe de Banach s'applique donc dans l'espace métrique complet $\bar{B}(a, r)$ avec distance induite par la norme. \diamond

Remarque 3.3.1 1. Un examen rapide des arguments ci-dessus montre que la dimension finie n'y joue aucun rôle. En fait, la proposition est vrai dans exactement les mêmes termes dans un espace vectoriel normé E quelconque, et le corollaire est vrai dans un espace de Banach E quelconque.

2. Dans \mathbb{R}^n , toutes les normes sont équivalentes, donc le choix de la norme n'influe pas sur le caractère C^1 ou pas. Par contre, il influe sur la forme des boules ainsi que sur les valeurs numériques comme $\|df_x\|_{\mathcal{L}(\mathbb{R}^n; \mathbb{R}^n)}$. Suivant ce choix de norme, la quantité $\|df_x\|_{\mathcal{L}(\mathbb{R}^n; \mathbb{R}^n)}$ se calcule plus ou moins facilement, mais on peut très souvent au moins l'estimer. Le fait qu'on demande qu'elle soit majorée sur une boule par une constante strictement inférieure à 1 est à rapprocher de la notion de point fixe attractif vue en dimension $n = 1$. Ainsi, si a est un point fixe de f , alors la deuxième condition est automatiquement satisfaite. \diamond

3.3.2 $f(x) = 0$ du point de vue théorique

Étant donné un ouvert U de \mathbb{R}^n et f une application de classe C^1 de U à valeurs dans \mathbb{R}^n , on cherche à trouver les $x \in U$ tels que $f(x) = 0$.

Tout d'abord, est-ce un but raisonnable ? On sait bien que oui si $n = 1$, mais en général, ce n'est pas si clair. En effet, dès $n = 2$, les dessins qui sont convaincants pour $n = 1$ doivent être faits en dimension $2 \times 2 = 4$, car le graphe de f est alors un sous-ensemble de \mathbb{R}^4 . Du coup, on n'y voit plus grand-chose.

Regardons à nouveau et plus précisément ce qui se passe quand f est une fonction affine et $U = \mathbb{R}^n$, c'est-à-dire que $f(x) = Ax - b$ où A est une matrice $n \times n$ et $b \in \mathbb{R}^n$. Les zéros de f sont donc exactement les solutions de l'équation $Ax = b$. On reconnaît un système linéaire de n équations à n inconnues. La discussion de ce système se fait classiquement selon que A est inversible ou non.

Si A est inversible, alors pour tout $b \in \mathbb{R}^n$, il y a une solution unique, donc isolée, qui est donnée par $x = A^{-1}b$.

Si A n'est pas inversible, alors son noyau n'est pas réduit au vecteur nul. Si b appartient à l'image de A , alors l'ensemble des solutions est de la forme $x_p + \ker A$, où x_p est une solution particulière. C'est un sous-espace affine de dimension supérieure à 1, en particulier les solutions ne sont pas isolées. Si par contre b n'appartient pas à l'image de A , alors l'ensemble des solutions est vide.

Quand on prend une matrice carrée « au hasard » (sans donner un sens précis à cette expression), celle-ci va presque sûrement être inversible et on sera dans la première situation. Dit autrement, il faut vraiment soit jouer de malchance, soit s'y prendre de façon délibérée pour avoir une matrice A $n \times n$ non inversible.¹³

On peut penser que la situation générique avec une fonction f non nécessairement affine sera analogue, c'est-à-dire que sauf malchance, si on a un zéro de f , alors ce zéro va être isolé. Dans toute la suite, on va se donner une norme $\|\cdot\|$ sur \mathbb{R}^n fixée une fois pour toutes. Pour fixer les idées, on va prendre $\|x\| = \max_i |x_i|$, mais cela peut aussi être n'importe quelle autre norme, c'est sans importance.¹⁴ Pour cette norme, les boules sont des (hyper)cubes,



ce qui est aussi sans importance. Elles sont bien convexes.

Proposition 3.3.3 Soit U un ouvert de \mathbb{R}^n et $f: U \rightarrow \mathbb{R}^n$. On se donne $x \in U$ tel que $f(x) = 0$, on suppose que f y est différentiable et que $\nabla f(x)$ est inversible. Alors x est un zéro isolé de f .

Démonstration. Définissons $g: U \rightarrow \mathbb{R}^n$ par $g(y) = (\nabla f(x))^{-1}f(y)$. On a donc $g(x) = 0$ et $\nabla g(x) = I$, la matrice identité, et par définition de la différentiabilité en x , on peut écrire

$$g(y) = g(x) + I(y - x) + \|y - x\|\varepsilon(y - x) = y - x + \|y - x\|\varepsilon(y - x),$$

pour tout $y \in U$. Toujours par définition de la différentiabilité, il existe une boule ouverte $B(x, r)$ telle que $\|\varepsilon(y - x)\| < \frac{1}{2}$ pour tout $y \in B(x, r)$. Par l'inégalité triangulaire, il s'ensuit que

$$\|g(y)\| > \frac{1}{2}\|y - x\|$$

¹³. Par contre, il arrive couramment que l'on tombe sur une matrice inversible, donc bien gentille en théorie, mais qui est telle que le calcul effectif en pratique de la solution est extrêmement difficile, voire impossible avec une précision raisonnable.

¹⁴. Encore que ce choix conduise à des calculs parfois plus simples que d'autres choix.

ce qui montre que le seul endroit où g s'annule dans cette boule est au point x , et il en va de même de $f = (\nabla f(x))g$ puisque le noyau de $\nabla f(x)$ est réduit au vecteur nul. \diamond

Remarque 3.3.2 En fait, on a beaucoup plus fort quand f est de classe C^1 , en faisant appel au *théorème d'inversion locale*, un résultat de calcul différentiel qui nous dit que f est un C^1 -difféomorphisme local, en particulier est localement injective, ce qui implique immédiatement que x est un zéro isolé. Le théorème d'inversion locale est d'ailleurs un autre exemple d'application du théorème de point fixe de Banach.

Bien sûr, si $\nabla f(x)$ n'est pas inversible, on ne peut rien dire. \diamond

3.3.3 La méthode de Newton(-Raphson)

On est maintenant rassuré qu'il n'est pas a priori idiot de chercher des zéros isolés d'une fonction de \mathbb{R}^n dans \mathbb{R}^n . En dimension n quelconque, la méthode de Newton prend parfois le nom de méthode de Newton-Raphson. Le principe est le même qu'en dimension 1.

Nous avons toujours un ouvert U de \mathbb{R}^n et f une application de classe C^2 de U à valeurs dans \mathbb{R}^n . La classe C^2 est ici entendue¹⁵ au sens où toutes les dérivées partielles secondes de toutes les composantes de f sont bien définies et continues sur U . On cherche à déterminer les $x \in U$ tels que $f(x) = 0$.

On suppose disposer d'une valeur approchée $x_0 \in U$ d'une solution x . Comme dans la méthode de Newton scalaire, l'idée est d'approcher f par sa partie linéaire au voisinage de x_0 . En effet, par différentiabilité de f en x_0 , on a

$$0 = f(x) = f(x_0) + \nabla f(x_0)(x - x_0) + \|x - x_0\| \varepsilon(x - x_0),$$

où la fonction ε dépend de x_0 . Mais comme on ne sait rien sur cette dernière, mis à part le fait que $\varepsilon(h) \rightarrow 0$ quand $h \rightarrow 0$, on ne peut pas en faire grand-chose directement.

On se contente donc de chercher à résoudre l'équation d'inconnue x_1 ,

$$f(x_0) + \nabla f(x_0)(x_1 - x_0) = 0, \quad (3.3.1)$$

dans laquelle on a simplement enlevé le reste, et où ne subsiste que la partie linéaire de f , c'est-à-dire l'équivalent de l'équation de la tangente au graphe en dimension 1 (la tangente au graphe est remplacée par un sous-espace affine de \mathbb{R}^{2n} de dimension n). On espère alors, comme en dimension $n = 1$, que x_1 va être une meilleure approximation de x que x_0 (comme le montre clairement le dessin en dimension $2n$ pour ceux qui possèdent des pouvoirs de visualisation en dimension supérieure à 3 hors du commun).

L'équation (3.3.1) en l'inconnue x_1 est en fait un système linéaire de n équations à n inconnues, $\nabla f(x_0)x_1 = \nabla f(x_0)x_0 - f(x_0)$. Si la matrice $\nabla f(x_0)$ est inversible, alors on a une solution unique x_1 simplement donnée par

$$x_1 = x_0 - (\nabla f(x_0))^{-1}f(x_0),$$

¹⁵ Pour éviter de devoir introduire les différentielles d'ordre supérieur à 1, par exemple d'ordre 2. Si les différentielles d'ordre 1 ressortent de l'algèbre linéaire, celles d'ordre supérieur à 1 ressortent de l'algèbre multilinéaire. Ce n'est pas fondamentalement beaucoup plus dur, mais c'est définitivement plus lourd. Ici, on peut s'en tirer avec juste des dérivées partielles et cette notion de classe C^2 à base de dérivées partielles secondes coincide avec celle que l'on aurait si l'on avait la vraie différentielle seconde à notre disposition.

et si jamais $x_1 \in U$ et $\nabla f(x_1)$ est encore inversible, alors on peut itérer le processus pour calculer x_2 , et ainsi de suite. La méthode de Newton-Raphson consiste donc à construire si possible la suite

$$x_{k+1} = x_k - (\nabla f(x_k))^{-1} f(x_k). \quad (3.3.2)$$

Il faut ajouter « si possible », car rien ne garantit a priori que si $x_k \in U$ et $\nabla f(x_k)$ est inversible, alors $x_{k+1} \in U$ et $\nabla f(x_{k+1})$ est inversible. Il est même facile de construire des contre-exemples.

Remarquons qu'en dimension $n = 1$, multiplier $f(x_k) \in \mathbb{R}^n$ à gauche par l'inverse de la matrice $\nabla f(x_k)$ consiste exactement à diviser par $f'(x_k)$.

Commençons par quelques éléments utiles sur l'espace vectoriel des matrices $M_n(\mathbb{R})$, que l'on munit d'une norme (c'est un espace de dimension n^2) et sur le sous-ensemble des matrices inversibles $GL_n(\mathbb{R})$. On rappelle qu'un ouvert de $M_n(\mathbb{R})$ est une réunion de boules ouvertes, et que ceci ne dépend pas du choix de la norme, puisqu'il s'agit encore d'un espace vectoriel sur \mathbb{R} de dimension finie.

On pourrait prendre n'importe quelle norme sur $M_n(\mathbb{R})$, par exemple $|||A||| = \max_{i,j} |a_{ij}|$, mais il est plus agréable pour travailler d'en choisir une qui soit adaptée à la norme que l'on a déjà adoptée sur \mathbb{R}^n . Pour cela, on introduit la notion de *norme matricielle subordonnée*. Cette notion vaut pour n'importe quelle norme sur \mathbb{R}^n . On va voir qu'il s'agit en fait du pendant matriciel de la norme d'application linéaire déjà introduite précédemment.

Définition 3.3.4 Soit $\|\cdot\|$ une norme quelconque sur \mathbb{R}^n . L'application $A \mapsto |||A||| = \sup_{\|x\| \leq 1} \|Ax\|$ est une norme sur $M_n(\mathbb{R})$ appelée norme matricielle subordonnée à la norme sur \mathbb{R}^n .

Proposition 3.3.5 Pour tout $x \in \mathbb{R}^n$, on a

$$\|Ax\| \leq |||A||| \|x\|.$$

De plus, pour tous $A, B \in M_n(\mathbb{R})$, on a

$$|||AB||| \leq |||A||| |||B|||.$$

Enfin, dans le cas de la norme $\|x\| = \max_i |x_i|$, on a

$$|||A||| = \max_{1 \leq i \leq n} \sum_{j=1}^n |a_{ij}|.$$

Démonstration. Pour tout $\lambda \in \mathbb{R}$, $|||\lambda A||| = \sup_{\|x\| \leq 1} \|\lambda Ax\| = |\lambda| \sup_{\|x\| \leq 1} \|Ax\| = |\lambda| |||A|||$, d'où la positivité homogène. De même, $|||A + B||| = \sup_{\|x\| \leq 1} \|(A + B)x\| = \sup_{\|x\| \leq 1} \|Ax + Bx\| \leq \sup_{\|x\| \leq 1} (\|Ax\| + \|Bx\|) \leq \sup_{\|x\| \leq 1} \|Ax\| + \sup_{\|x\| \leq 1} \|Bx\| = |||A||| + |||B|||$, d'où l'inégalité triangulaire. Enfin, si $\sup_{\|x\| \leq 1} \|Ax\| = 0$, c'est bien clairement que $A = 0$. On a affaire à une norme sur l'espace $M_n(\mathbb{R})$.

Pour tout $x \in \mathbb{R}^n$, il existe $u \in \mathbb{R}^n$, $\|u\| = 1$, tel que $x = \|x\|u$. En effet, si $x = 0$, n'importe quel u fait l'affaire et si $x \neq 0$, on prend $u = \frac{x}{\|x\|}$. Donc

$$\|Ax\| = \|x\| \|Au\| \leq |||A||| \|x\|.$$

Ensuite, pour tout $\|x\| \leq 1$, on a

$$\|(AB)x\| = \|A(Bx)\| \leq \|\|A\|\| \|Bx\| \leq \|\|A\|\| \|\|B\|\|,$$

d'où la deuxième inégalité en passant au sup à gauche.

Enfin dans le cas $\|x\| = \max_i |x_i| \leq 1$, on voit que

$$\|Ax\| = \max_{1 \leq i \leq n} \left| \sum_{j=1}^n a_{ij} x_j \right| \leq \max_{1 \leq i \leq n} \sum_{j=1}^n |a_{ij}| |x_j| \leq \max_{1 \leq i \leq n} \sum_{j=1}^n |a_{ij}|,$$

avec égalité en prenant un indice i_0 où le max du terme de droite est atteint et en posant $x_j = \text{signe } a_{i_0,j}$ si $a_{i_0,j} \neq 0$, $x_j = 0$ sinon. \diamond

Remarquons que pour toute norme matricielle subordonnée, on a $\|\|I\|\| = 1$. En fait, la norme matricielle subordonnée d'une matrice n'est rien d'autre que la norme d'application linéaire de l'application linéaire qu'elle définit sur \mathbb{R}^n muni de sa base canonique. Il existe aussi des *normes matricielles*, c'est-à-dire des normes sur $M_n(\mathbb{R})$ qui vérifient l'inégalité de sous-multiplicativité $\|\|AB\|\| \leq \|\|A\|\| \|\|B\|\|$, mais qui ne sont pas nécessairement subordonnées à une norme sur \mathbb{R}^n .

Lemme 3.3.6 *L'ensemble des matrices inversibles $GL_n(\mathbb{R})$ est un ouvert de $M_n(\mathbb{R})$.*

Démonstration. Il suffit de montrer que pour toute matrice inversible $A \in GL_n(\mathbb{R})$, il existe une boule ouverte $B(A, r)$ incluse dans $GL_n(\mathbb{R})$. On traite d'abord le cas $A = I$. Pour tout $R \in M_n(\mathbb{R})$ et $N \in \mathbb{N}$, on a l'identité remarquable¹⁶

$$I - R^{N+1} = (I - R)(I + R + R^2 + \cdots + R^N).$$

D'après la proposition précédente, on a pour tout entier naturel i

$$\|\|R^i\|\| \leq \|\|R\|\|^i.$$

Supposons que $\|\|R\|\| < 1$, ce qui est équivalent à dire que $I - R \in B(I, 1)$. On en déduit deux choses. D'une part, $R^{N+1} \rightarrow 0$ quand $N \rightarrow +\infty$, donc $I - R^{N+1} \rightarrow I$. D'autre part, la suite $S_N = \sum_{i=0}^N R^i$ est une suite de Cauchy dans $M_n(\mathbb{R})$. En effet, par l'inégalité triangulaire et l'estimation ci-dessus

$$\|\|S_{N+m} - S_N\|\| \leq \sum_{i=N+1}^{N+m} \|\|R^i\|\| \leq \sum_{i=N+1}^{N+m} \|\|R\|\|^i = \|\|R\|\|^{N+1} \frac{1 - \|\|R\|\|^m}{1 - \|\|R\|\|} \leq \frac{\|\|R\|\|^{N+1}}{1 - \|\|R\|\|} \rightarrow 0$$

quand $N \rightarrow +\infty$. Or l'espace des matrices $M_n(\mathbb{R})$ est complet pour n'importe quelle norme, donc la suite S_N converge vers une matrice $S \in M_n(\mathbb{R})$. On peut alors passer à la limite quand $N \rightarrow +\infty$ dans l'identité remarquable ci-dessus. On obtient ainsi

$$I = (I - R)S.$$

En effet,

$$\|\|(I - R)S_N - (I - R)S\|\| \leq \|\|I - R\|\| \|\|S_N - S\|\| \rightarrow 0 \text{ quand } N \rightarrow +\infty.$$

¹⁶. Valable dans tout anneau unitaire.

Ceci montre que $I - R$ est inversible (avec $(I - R)^{-1} = S = \sum_{i=0}^{\infty} R^i$). On vient donc de montrer que $B(I, 1) \subset GL_n(\mathbb{R})$.

Revenons au cas général, avec $A \in GL_n(\mathbb{R})$. Pour tout $C \in M_n(\mathbb{R})$, on pose $R = A - C$. Comme A est inversible, en multipliant à gauche par A^{-1} , ceci se réécrit $A^{-1}C = I - A^{-1}R$. D'après l'étape précédente, si $\|A^{-1}R\| < 1$, alors on est assuré que $I - A^{-1}R = A^{-1}C$ est inversible. Ceci implique que $C = A(A^{-1}C)$ est également inversible. Il suffit donc que $\|R\| < \frac{1}{\|A^{-1}\|}$, pour que C soit inversible. En effet, dans ce cas, $\|A^{-1}R\| \leq \|R\| \|A^{-1}\| < 1$. En d'autres termes, on a montré que $B(A, \frac{1}{\|A^{-1}\|}) \subset GL_n(\mathbb{R})$, d'où le résultat avec $r = \frac{1}{\|A^{-1}\|}$. \diamond

Remarque 3.3.3 La preuve ci-dessus fonctionne pour n'importe quelle norme matricielle subordonnée, c'est-à-dire en fait pour n'importe quel choix de norme dans \mathbb{R}^n . Cela marche même pour toute norme matricielle. Donc finalement, la réunion de toutes les boules ouvertes de centre I et de rayon 1 pour toutes les normes matricielles possibles et imaginables est incluse dans $GL_n(\mathbb{R})$, et de même plus généralement pour $B(A, \frac{1}{\|A^{-1}\|})$ dès que A est inversible. \diamond

Proposition 3.3.7 Si f est de classe C^2 sur U et $\nabla f(x)$ est inversible, alors il existe une boule centrée en x telle que pour toute donnée initiale dans cette boule, la suite des itérations de Newton est bien définie et converge quadratiquement vers x :

$$\|x_{k+1} - x\| \leq K \|x_k - x\|^2,$$

pour un certain K .

Démonstration. Comme $\nabla f(x)$ est inversible, il existe une boule $B(\nabla f(x), r)$ dans l'espace des matrices qui ne contient que des matrices inversibles par le Lemme 3.3.6. Comme f est de classe C^1 , il existe une boule $B(x, s)$ dans U cette fois telle que si $y \in B(x, s)$ alors $\nabla f(y)$ appartient à $B(\nabla f(x), r)$, et est donc inversible. On commence par se restreindre à la boule $B(x, s)$.

Dans cette boule, on définit une fonction $g: B(x, s) \rightarrow \mathbb{R}^n$ en posant

$$g(y) = y - (\nabla f(y))^{-1}f(y).$$

Par définition de la méthode de Newton, si x_k est bien défini, on a $x_{k+1} = g(x_k)$. La suite de Newton est donc celle des itérées de x_0 par g et il suffit par conséquent de montrer que g admet une boule invariante pour montrer que cette suite est bien définie pour tout k .

Pour tout $y \in B(x, s)$, on pose $h: [0, 1] \rightarrow \mathbb{R}^n$, $h(t) = f(y + t(x - y))$ de telle sorte que $h(0) = f(y)$ et $h(1) = f(x) = 0$. Par dérivation des fonctions composées, h est de classe C^2 , avec

$$h'(t) = \nabla f(y + t(x - y))(x - y) \text{ et } h''(t) = \nabla^2 f(y + t(x - y))((x - y), (x - y)).$$

Attention, dans le contexte présent, ∇f est la matrice jacobienne dont les composantes sont indexées par deux indices, et $\nabla^2 f$ est une bête à trois indices¹⁷ dont les composantes sont

¹⁷. On appelle ça un *tenseur*.

$(\nabla^2 f)_{ijk} = \frac{\partial^2 f_i}{\partial x_j \partial x_k}$. Quand on écrit tout en composantes, ceci donne

$$(\nabla f(a)b)_i = \sum_{j=1}^n \frac{\partial f_i}{\partial x_j}(a)b_j,$$

et

$$(\nabla^2 f(a)(b, c))_i = \sum_{j,k=1}^n \frac{\partial^2 f_i}{\partial x_j \partial x_k}(a)b_j c_k.$$

Posons

$$M = \sup_{y \in B(x, s)} \left(\max_i \sum_{j,k=1}^n \left| \frac{\partial^2 f_i}{\partial x_j \partial x_k}(y) \right| \right),$$

il vient alors

$$\sup_{y \in B(x, s)} \|\nabla^2 f(y)(b, c)\| \leq M \|b\| \|c\|.$$

Revenant à la fonction h , l'inégalité de Taylor-Lagrange nous dit que

$$\|h(1) - h(0) - h'(0)\| \leq \frac{1}{2} \sup_{t \in [0, 1]} \|h''(t)\|.$$

Quand on réexprime ceci en termes de f , on obtient

$$\begin{aligned} \|f(y) + \nabla f(y)(x - y)\| &\leq \frac{1}{2} \sup_{t \in [0, 1]} \|\nabla^2 f(y + t(x - y))((x - y), (x - y))\| \\ &\leq \frac{1}{2} M \|x - y\|^2, \end{aligned}$$

puisque le segment qui joint x à y est entièrement dans la boule. Comme

$$g(y) = y - (\nabla f(y))^{-1}f(y) = (\nabla f(y))^{-1}(\nabla f(y)y - f(y)),$$

on voit que

$$g(y) - x = -(\nabla f(y))^{-1}(\nabla f(y)(x - y) + f(y)),$$

d'où

$$\begin{aligned} \|g(y) - x\| &\leq \|(\nabla f(y))^{-1}\| \|\nabla f(y)(x - y) + f(y)\| \\ &\leq \frac{1}{2} M \|(\nabla f(y))^{-1}\| \|x - y\|^2 \\ &\leq K \|x - y\|^2, \end{aligned}$$

où l'on a posé $K = \frac{1}{2} M \sup_{y \in B(x, s)} \|(\nabla f(y))^{-1}\|$.

On pose alors $\alpha = \min(s, \frac{1}{K})$. Si $y \in B(x, \alpha) \subset B(x, s)$, on a donc $\|y - x\| < \alpha$, d'où

$$\|g(y) - x\| \leq K \alpha^2 \leq \alpha,$$

puisque $\alpha \leq \frac{1}{K}$. On en déduit que $g(x) \in B(x, \alpha)$, c'est-à-dire que la boule $B(x, \alpha)$ est invariante par g , avec de plus le fait que ∇f est inversible en tout point de cette boule. La suite de Newton x_k est donc bien définie si x_0 est pris dans la boule $B(x, \alpha)$.

De plus, il vient immédiatement que, comme $x_{k+1} = g(x_k)$ pour tout k ,

$$\|x_{k+1} - x\| \leq K \|x_k - x\|^2,$$

d'où la convergence quadratique de la méthode de Newton. \diamond

Remarque 3.3.4 Il s'agit essentiellement de la même démonstration qu'en dimension 1. Les mêmes remarques concernant l'éventuelle divergence de la méthode si x_0 est trop loin de x s'appliquent. \diamond

Le cas particulier $n = 2$ est particulièrement intéressant quand on restreint l'attention aux fonctions holomorphes, c'est-à-dire quand on identifie \mathbb{R}^2 et \mathbb{C} , et que l'on considère les fonctions d'un ouvert de \mathbb{C} à valeurs dans \mathbb{C} qui sont dérivables au sens de \mathbb{C} . Cette condition est beaucoup plus restrictive qu'être juste différentiable de \mathbb{R}^2 dans \mathbb{R}^2 . De façon très étonnante, sans aucune autre hypothèse que leur dérivabilité, même pas l'hypothèse de classe C^1 , les fonctions holomorphes sont en fait automatiquement de classe C^∞ et même analytiques, c'est-à-dire localement égales à leur série de Taylor. De plus, leurs zéros sont forcément isolés. C'est par exemple le cas des fonctions polynomiales (ou de l'exponentielle complexe, mais celle-ci n'a pas beaucoup de zéros). La méthode de Newton va évidemment s'appliquer pour en approcher les zéros, puisqu'il s'agit d'un cas particulier de fonction de \mathbb{R}^2 dans \mathbb{R}^2 .

Comme l'action de la différentielle au point z , considérée comme application linéaire de \mathbb{R}^2 dans \mathbb{R}^2 , sur un vecteur est une similitude, qui se traduit du point de vue complexe par la multiplication complexe par le nombre complexe $f'(z)$, l'inverse de la différentielle correspond à la division complexe par $f'(z)$ quand celui-ci n'est pas nul. On se retrouve donc avec la même formulation que dans le cas réel, $z_{k+1} = z_k - f(z_k)/f'(z_k)$, mais dans le plan complexe.

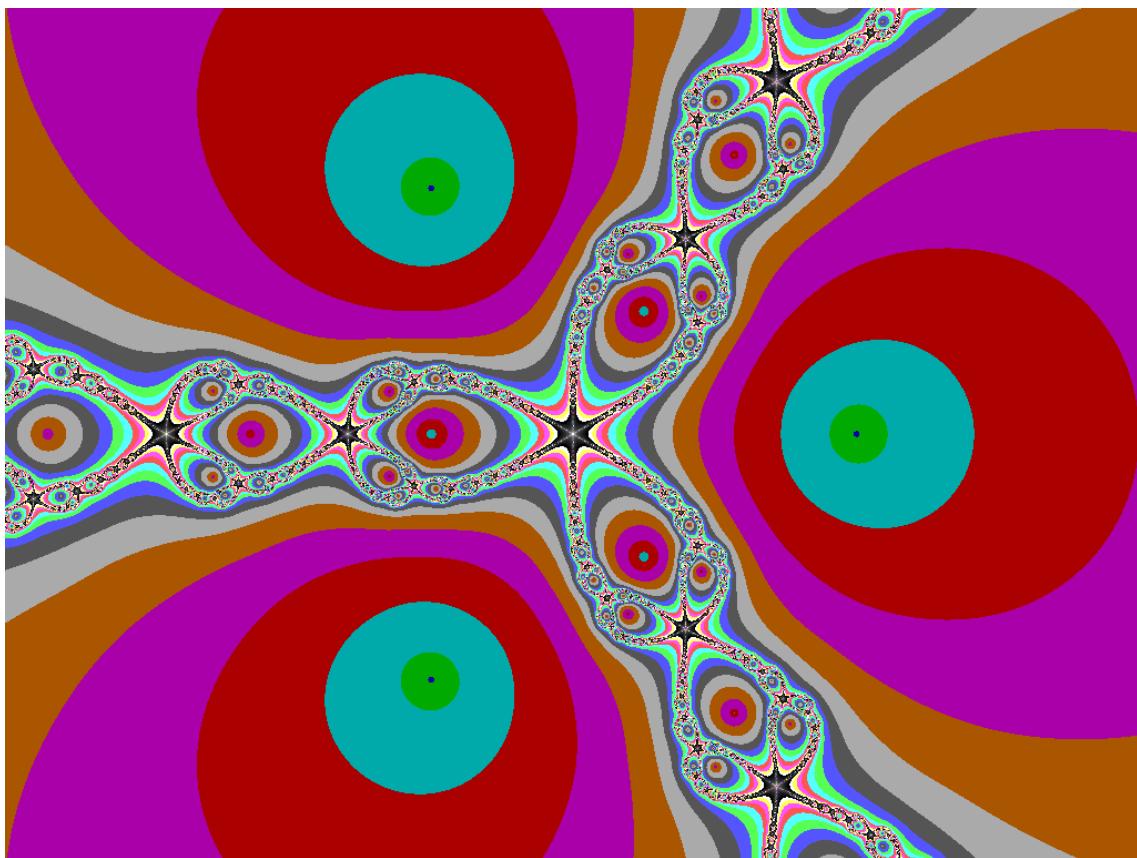
Par la théorie générale, on sait que la méthode de Newton va converger vers une racine quand le point initial z_0 est suffisamment proche de la racine en question. C'est un résultat local. Qu'en est-il plus globalement ? La suite peut converger ou pas et une question qui s'est posée un certain temps, dans le cas des polynômes, est quelle est l'influence du choix de z_0 . En particulier, si la suite de Newton converge partant de z_0 , vers quelle racine du polynôme converge-t-elle ?

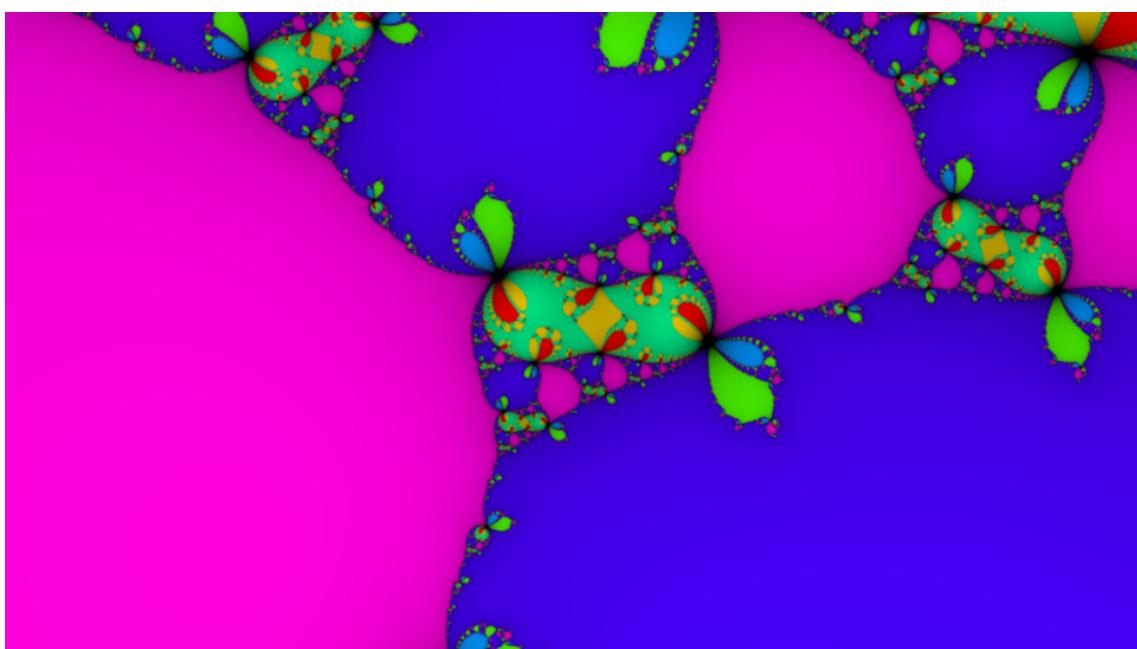
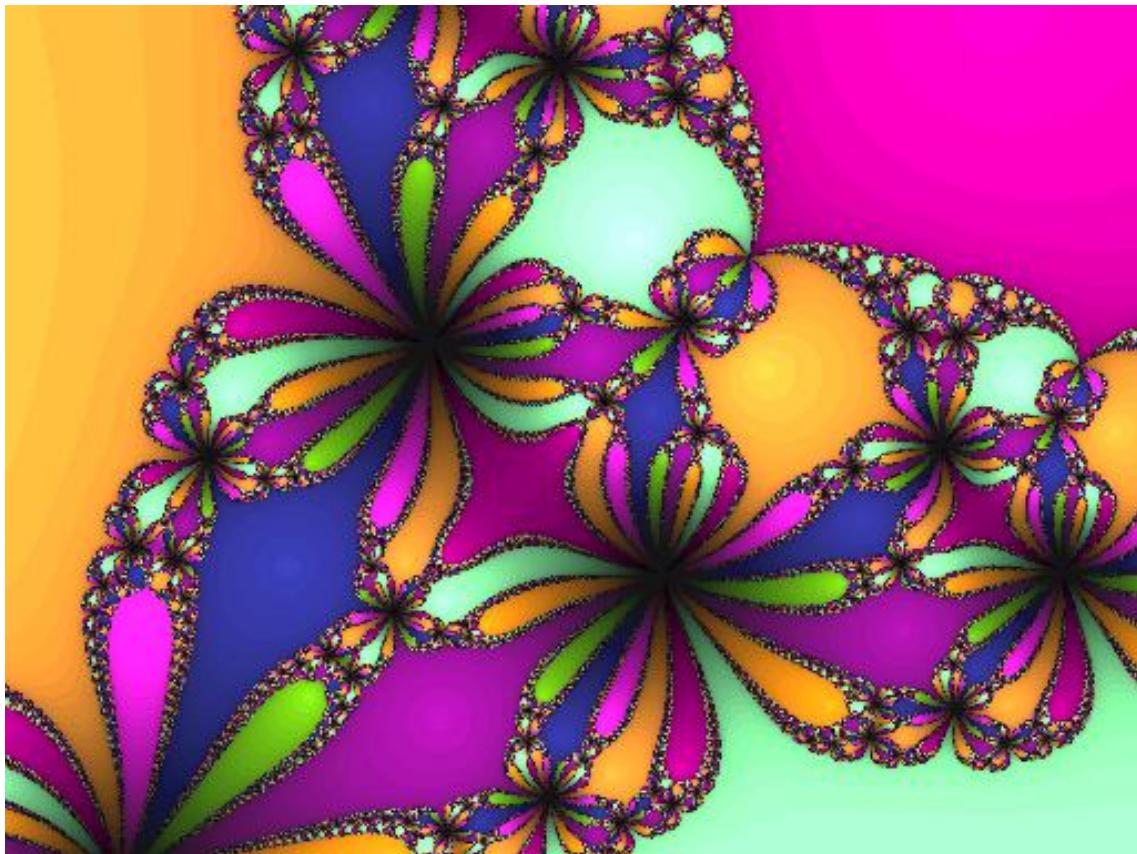
Cayley a montré au 19ème siècle que pour des polynômes de degré deux, cette suite converge vers la racine la plus proche du point de départ. Le plan complexe est donc coupé en deux par une droite, la médiatrice des deux racines (on suppose les racines distinctes), en dehors de laquelle on a convergence vers la racine incluse dans le même demi-plan ouvert. On vérifie par un calcul direct que la médiatrice est invariante par l'itération de Newton. Si le point initial est situé sur cette médiatrice, il s'ensuit que la suite de Newton ne converge pas.

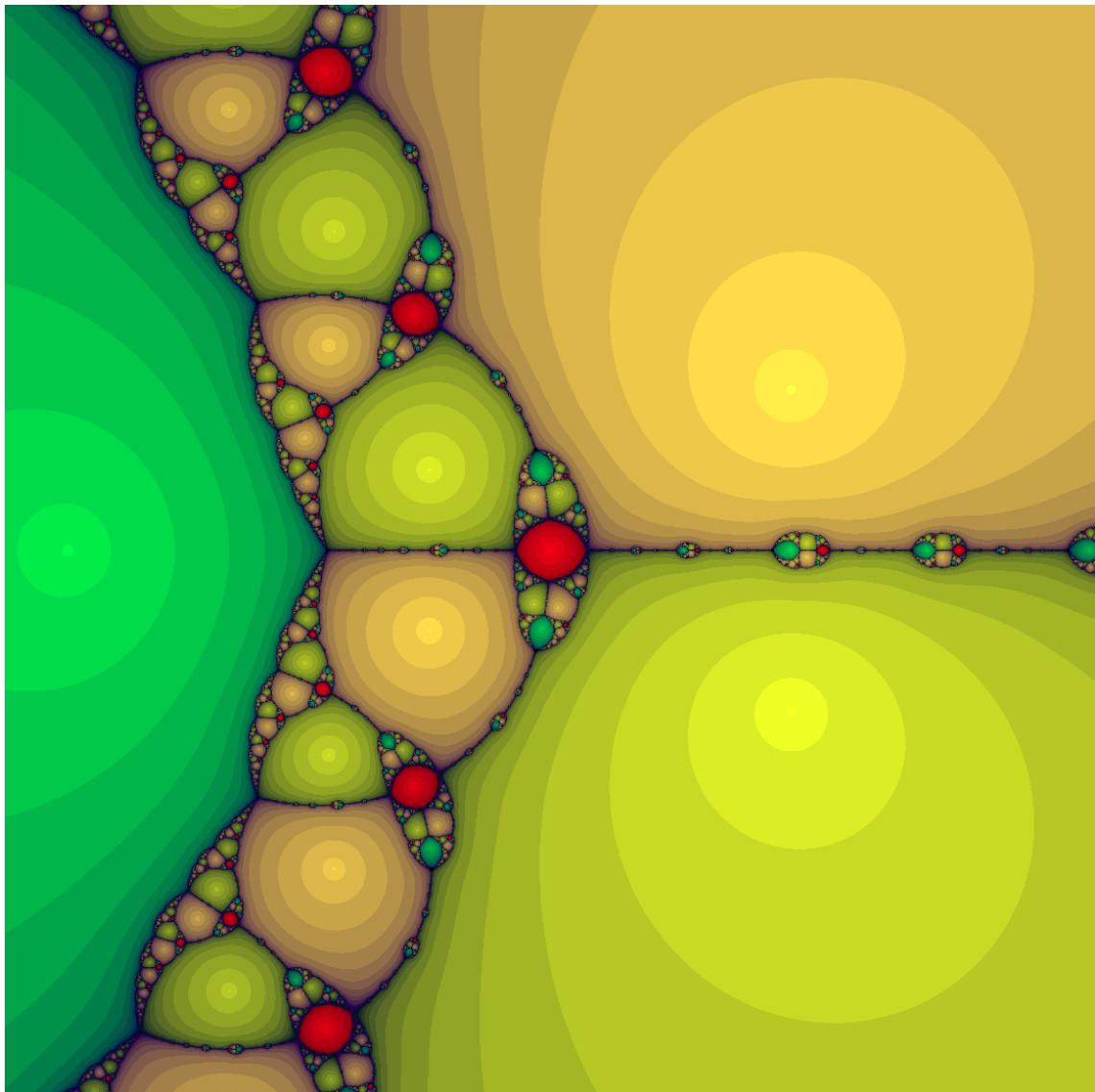
Les choses se complexifient (si l'on ose dire) considérablement à partir du troisième degré. Les bassins d'attraction de chaque racine, c'est-à-dire les régions du plan qui correspondent à une valeur initiale dont la suite des itérées converge vers cette racine, possèdent une structure compliquée. Leur frontière commune a une structure de type fractal. Par exemple, dans le cas d'un polynôme à trois racines distinctes, il peut se faire que chaque point de la frontière soit adhérent en même temps à trois bassins d'attraction différents, ce qui est un

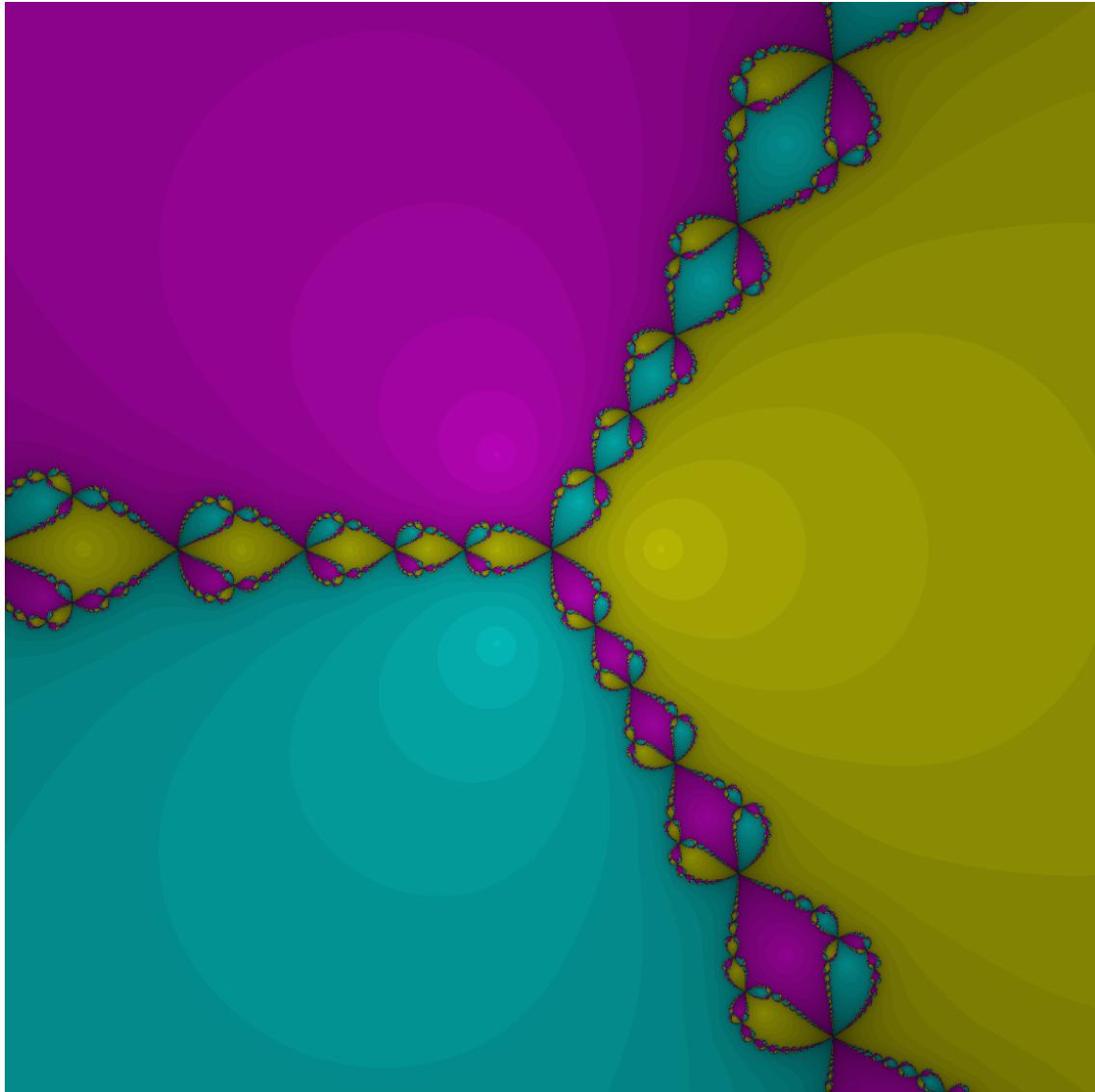
peu dur à visualiser. On en effet tendance à penser qu'une frontière ne sépare en général que deux pays, sauf que là, c'est trois pays qui se côtoient en chacun de ses points.

On trouve sur le web de nombreuses images de « fractales de Newton » qui illustrent ces propriétés de convergence ou de divergence d'ailleurs. On y représente souvent (mais pas toujours, comme dans le premier exemple qui suit) les bassins d'attraction d'une certaine couleur, et l'on peut identifier vers quelle racine on converge sur la base de cette couleur.









Une frontière, trois pays... mais comment la douane s'y prend-elle ?

3.4 Le théorème de (Newton)-Kantorovich

Un théorème évidemment dû à Kantorovich, à propos de la méthode de Newton. En fait, il y a plusieurs versions du théorème de Kantorovich, dont une assez optimale, mais horriblement technique. On va se contenter d'une version moins horrible, mais moins optimale.

Le point principal des théorèmes de type Kantorovich est non seulement de montrer la convergence de la méthode de Newton, mais par la même occasion, de donner des *conditions suffisantes d'existence de racine* que l'on a une petite chance de pouvoir vérifier en pratique ! Ce n'est pas rien. Énonçons le résultat qui va nous occuper dans la suite.

Théorème 3.4.1 (Kantorovich) Soit U un ouvert de \mathbb{R}^n , $x_0 \in U$ et $f: U \rightarrow \mathbb{R}^n$ de classe C^1 .

On suppose que $\nabla f(x_0) \in GL_n(\mathbb{R})$ et qu'il existe un nombre $r > 0$ tel que $\overline{B(x_0, r)} \subset U$ et

$$\|(\nabla f(x_0))^{-1}f(x_0)\| \leq \frac{r}{2}, \quad (3.4.1)$$

$$\forall y, z \in B(x_0, r), \|(\nabla f(x_0))^{-1}(\nabla f(y) - \nabla f(z))\| \leq \frac{1}{r}\|y - z\|. \quad (3.4.2)$$

Alors pour tout $y \in B(x_0, r)$, on a $\nabla f(y) \in GL_n(\mathbb{R})$, les itérations de Newton x_k partant de x_0 sont bien définies pour tout k , restent dans $B(x_0, r)$ et convergent vers $x \in \overline{B(x_0, r)}$, qui est de plus l'unique racine de f dans $\overline{B(x_0, r)}$. On a enfin l'estimation d'erreur

$$\|x_k - x\| \leq \frac{r}{2^k}.$$

Démonstration. On va se simplifier la vie en posant le changement de variable suivant. Si $u \in B(0, 1)$ alors $x = x_0 + ru \in B(x_0, r)$ et réciproquement $u = \frac{x-x_0}{r} \in B(0, 1)$. On introduit donc $h: B(0, 1) \rightarrow \mathbb{R}^n$ par

$$h(u) = \frac{1}{r}(\nabla f(x_0))^{-1}f(x_0 + ru),$$

ou de façon équivalente

$$f(x) = r\nabla f(x_0)h\left(\frac{x-x_0}{r}\right).$$

En ce qui concerne les matrices jacobiniennes, on obtient

$$\nabla h(u) = (\nabla f(x_0))^{-1}\nabla f(x_0 + ru) \text{ et } \nabla f(x) = \nabla f(x_0)\nabla h\left(\frac{x-x_0}{r}\right),$$

par dérivation des fonctions composées. En particulier, $h(0) = \frac{1}{r}(\nabla f(x_0))^{-1}f(x_0)$ et $\nabla h(0) = I$.

Quand on les réécrit en termes du changement de variables u et h , les hypothèses (3.4.1) et (3.4.2) deviennent plus agréables

$$\|h(0)\| \leq \frac{1}{2}, \quad (3.4.3)$$

$$\forall u, v \in B(0, 1), \|\nabla h(u) - \nabla h(v)\| \leq \|u - v\|. \quad (3.4.4)$$

On remarque de plus que les itérations de Newton de f et celles de h se correspondent via le changement de variable. En effet, si $x_k = x_0 + ru_k$ est l'itération de Newton de f , alors

$$\begin{aligned} u_{k+1} &= \frac{x_{k+1} - x_0}{r} = \frac{x_k - x_0}{r} - \frac{\nabla f(x_k)^{-1}f(x_k)}{r} \\ &= u_k - \frac{\nabla h(u_k)^{-1}\nabla f(x_0)^{-1}f(x_k)}{r} = u_k - \nabla h(u_k)^{-1}h(u_k). \end{aligned}$$

Enfin les racines de f et de h dans leurs boules respectives se correspondent manifestement par le changement de variable. On va donc travailler sur h satisfaisant (3.4.3) et (3.4.4) dans la boule unité, et sur ses itérations de Newton, ça sera beaucoup plus confortable.

On prend d'abord $v = 0$ dans la deuxième inégalité (3.4.4). Il vient donc

$$\forall u \in B(0, 1), \|\nabla h(u) - I\| \leq \|u\| < 1.$$

On en déduit que $\nabla h(u)$ est inversible partout dans la boule ouverte, avec l'estimation

$$|||\nabla h(u)^{-1}||| \leq \frac{1}{1 - \|u\|} \quad (3.4.5)$$

qui découle immédiatement de la preuve du Lemme 3.3.6. De plus, on a

$$h(u) - h(v) = \int_0^1 \nabla h(v + t(u - v))(u - v) dt,$$

si bien que

$$\begin{aligned} \|h(u) - h(v) - \nabla h(v)(u - v)\| &= \left\| \int_0^1 (\nabla h(v + t(u - v)) - \nabla h(v))(u - v) dt \right\| \\ &\leq \int_0^1 \|(\nabla h(v + t(u - v)) - \nabla h(v))(u - v)\| dt \\ &\leq \int_0^1 |||\nabla h(v + t(u - v)) - \nabla h(v)||| \|u - v\| dt \\ &\leq \|u - v\|^2 \int_0^1 t dt = \frac{\|u - v\|^2}{2}, \end{aligned} \quad (3.4.6)$$

toujours par l'hypothèse (3.4.4).

Regardons ce que l'on peut dire de la première itération de Newton avec $u_0 = 0$, $u_1 = 0 - Ih(0) = -h(0)$. Bien sûr,

$$\|u_1\| = \|h(0)\| \leq \frac{1}{2}.$$

Par (3.4.5), on a aussi

$$|||\nabla h(u_1)^{-1}||| \leq \frac{1}{1 - \|u_1\|} \leq 2.$$

Enfin, comme $h(u_1) = h(u_1) + u_1 - u_1 + u_0 = h(u_1) - h(u_0) - \nabla h(u_0)(u_1 - u_0)$, on déduit de (3.4.6) que

$$\|h(u_1)\| \leq \frac{\|u_1\|^2}{2} \leq \frac{1}{8} = \frac{1}{2^3}.$$

Ceci incite fortement à poser l'hypothèse de récurrence suivante pour $k \geq 1$,

$$\|u_k - u_{k-1}\| \leq \frac{1}{2^k}, \|u_k\| \leq 1 - \frac{1}{2^k}, |||\nabla h(u_k)^{-1}||| \leq 2^k \text{ et } \|h(u_k)\| \leq \frac{1}{2^{2k+1}}. \quad (3.4.7)$$

Comme $u_k \in B(0, 1)$ par la deuxième inégalité, $u_{k+1} = u_k - \nabla h(u_k)^{-1}h(u_k)$ est bien défini. De plus, $h(u_{k+1}) = h(u_{k+1}) - h(u_k) - \nabla h(u_k)(u_{k+1} - u_k)$. Il vient donc

$$\|u_{k+1} - u_k\| = \|\nabla h(u_k)^{-1}h(u_k)\| \leq |||\nabla h(u_k)^{-1}||| \|h(u_k)\| \leq \frac{2^k}{2^{2k+1}} = \frac{1}{2^{k+1}}.$$

Par l'inégalité triangulaire, il s'ensuit que

$$\|u_{k+1}\| \leq \|u_{k+1} - u_k\| + \|u_k\| \leq \frac{1}{2^{k+1}} + 1 - \frac{1}{2^k} = 1 - \frac{1}{2^{k+1}}.$$

L'estimation (3.4.5) donne alors

$$|||\nabla h(u_{k+1})^{-1}||| \leq \frac{1}{1 - \|u_{k+1}\|} \leq 2^{k+1}.$$

Enfin,

$$\|h(u_{k+1})\| = \|h(u_{k+1}) - h(u_k) - \nabla h(u_k)(u_{k+1} - u_k)\| \leq \frac{\|u_{k+1} - u_k\|^2}{2} \leq \frac{1}{2^{2k+3}}.$$

On a ainsi montré que l'itération de Newton est bien définie pour tout k avec les estimations (3.4.7).

L'estimation $\|u_k - u_{k-1}\| \leq \frac{1}{2^k}$ montre comme on l'a vu déjà bien souvent que la suite u_k est de Cauchy, elle converge donc vers un u dans l'espace complet $\overline{B(0, 1)}$, avec la vitesse de convergence linéaire donnée par l'estimation d'erreur du théorème. Comme $h(u_k) \rightarrow 0$ par la dernière estimation de (3.4.7), on voit que $h(u) = 0$ par continuité.

Il reste à voir qu'il n'y a pas d'autre racine que u dans la boule fermée. Soit v une autre telle racine. On a manifestement $\|u_0 - v\| = \|v\| \leq 1 = \frac{1}{2^0}$. Faisons l'hypothèse de récurrence que $\|u_k - v\| \leq \frac{1}{2^k}$. Il vient

$$\begin{aligned} u_{k+1} - v &= u_k - v - \nabla h(u_k)^{-1}h(u_k) = u_k - v - \nabla h(u_k)^{-1}(h(u_k) - h(v)) \\ &= \nabla h(u_k)^{-1}(\nabla h(u_k)(u_k - v) - h(u_k) + h(v)), \end{aligned}$$

d'où, en utilisant à nouveau (3.4.6),

$$\|u_{k+1} - v\| \leq |||\nabla h(u_k)^{-1}||| \|\nabla h(u_k)(u_k - v) - h(u_k) + h(v)\| \leq 2^k \frac{\|u_k - v\|^2}{2} \leq \frac{1}{2^{k+1}}.$$

On a ainsi montré que $u_k \rightarrow v$, ce qui implique évidemment que $v = u$. \diamond

Remarque 3.4.1 i) La grosse différence avec la proposition 3.3.7 est que l'on ne suppose pas l'existence d'une racine, avec convergence dans une boule que l'on ne connaît pas, mais que l'on montre, si un certain nombre de conditions sont satisfaites, l'existence d'une telle racine avec convergence dans une boule que l'on connaît. Bien sûr, encore faut-il trouver des x_0 et r qui marchent.

ii) Regardons ce que dit le théorème de Kantorovich en dimension 1. Le changement de variable s'écrit $h(u) = f(x_0 + ru)/(rf'(x_0))$, avec h définie sur $[-1, 1]$ et $h'(0) = 1$. Les hypothèses du théorème s'écrivent

$$\begin{aligned} |h(0)| &\leq \frac{1}{2}, \\ |h'(u) - h'(v)| &\leq |u - v|. \end{aligned}$$

Il est assez clair en intégrant les inégalités $1 - u \leq h'(u) \leq 1 + u$ que l'on a alors une seule racine dans $[-1, 1]$, qui est négative si $h(0) \geq 0$ et positive si $h(0) \leq 0$. Dans le cas où h est deux fois dérivable, la condition de Lipschitz sur h' est équivalente au fait que $|h''(u)| \leq 1$ pour tout u , une borne sur la courbure du graphe.

iii) Le théorème est vrai pour n'importe quelle norme sur \mathbb{R}^n en utilisant sa norme matricielle subordonnée. Il reste aussi vrai avec la même démonstration dans un espace

de Banach quelconque, y compris de dimension infinie, en utilisant la norme d'application linéaire, qui se définit exactement comme une norme matricielle subordonnée.

iv) Le théorème ne fournit qu'une convergence linéaire, et non pas quadratique, pour la méthode de Newton. La raison est que l'on ne fait pas l'hypothèse que f est de classe C^2 , mais seulement C^1 avec la matrice jacobienne lipschitzienne (on dit que f est de classe $C^{1,1}$). De plus, la racine trouvée x est dans la boule fermée, et si elle est sur la sphère, rien n'interdit que $\nabla f(x)$ soit non inversible, voir Figure 3.1. On ne peut donc pas espérer plus qu'une convergence linéaire sous ces seules hypothèses. Néanmoins, si f est de classe C^2 et si la racine est dans la boule ouverte, alors on finit tôt ou tard par tomber dans la boule inconnue de la proposition 3.3.7, et la convergence est effectivement quadratique.

v) La version horrible du théorème de Kantorovich donne des conditions vérifiables en pratique et qui assurent la convergence quadratique. En fait, cette version ne suppose pas f de classe C^2 non plus, mais seulement $C^{1,1}$, avec tout un tas d'autres conditions. \diamond

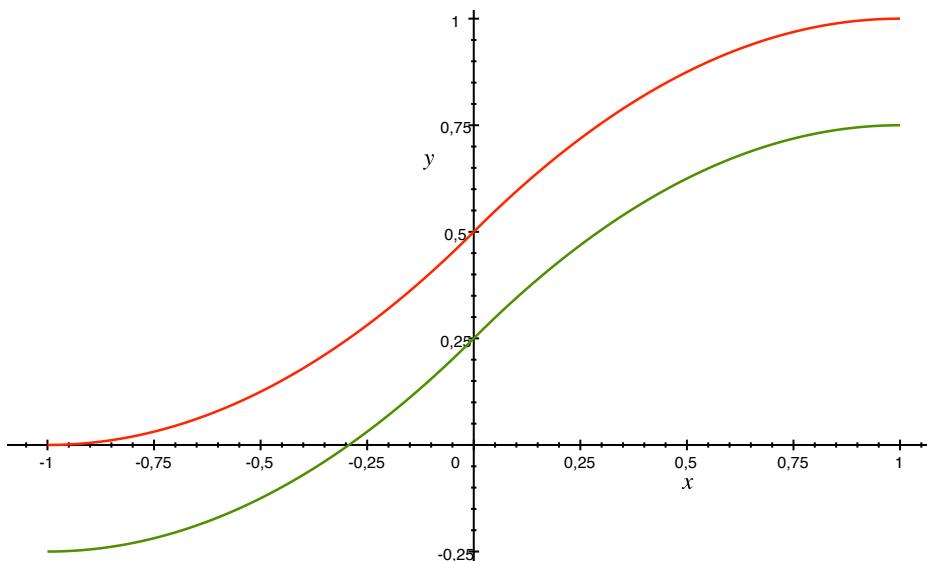


FIGURE 3.1 – Newton-Kantorovich : en vert, convergence quadratique (en partant de 0), en rouge, convergence linéaire.

3.5 Une vraie application, la méthode d'Euler implicite

Dans toutes ces notes, une seule application a été vraiment donnée, celle du calcul de $\sqrt{2}$. C'est un peu court pour de l'analyse dite « appliquée », mais le temps impari l'est également. À titre culturel, je développe rapidement une situation concrète où les méthodes que l'on a décrites sont nécessaires.

On s'intéresse à l'approximation de la solution d'un problème de Cauchy pour une équation différentielle ordinaire

$$\forall t \in [0, T], y'(t) = g(t, y(t)), \quad y(0) = y_0,$$

où $g: [0, T] \times \mathbb{R}^d \rightarrow \mathbb{R}^d$ et $y_0 \in \mathbb{R}^d$ sont donnés et l'on cherche la fonction inconnue $y: [0, T] \rightarrow \mathbb{R}^d$. Si g est continue d'une part et Lipschitzienne par rapport à y , uniformément par rapport à t d'autre part, c'est-à-dire qu'il existe une constante L telle que

$$\forall t \in [0, T], \forall y, z \in \mathbb{R}^d, \|g(t, y) - g(t, z)\| \leq L\|y - z\|,$$

alors ce problème de Cauchy admet une solution et une seule pour tout $y_0 \in \mathbb{R}^d$. C'est le théorème de Cauchy-Lipschitz global. Évidemment, il s'agit d'une question d'intérêt général majeur.

En général, il n'existe pas contre aucune formule qui donne cette solution, et l'on doit l'approcher pour pouvoir en dire quelque chose de quantitatif. Pour cela, on se donne un entier N et l'on introduit un pas de temps $h = \frac{T}{N+1}$, puis on pose $t_n = nh$, $n = 0, 1, \dots, N+1$ qui sont des instants de discrétisation de l'intervalle de temps $[0, T]$.

La méthode d'Euler implicite consiste à construire la suite $y_n \in \mathbb{R}^d$ définie par récurrence par

$$y_0 = y_0, \frac{y_{n+1} - y_n}{h} = g(t_{n+1}, y_{n+1}) \text{ pour } 0 \leq n \leq N.$$

On démontre que cette méthode marche et que y_n est une approximation de $y(t_n)$ d'autant meilleure que h est petit, en un sens précis que l'on ne détaillera pas ici.

Comment calculer y_{n+1} connaissant y_n pour faire progresser la récurrence ? Sauf cas particulier où g est très simple, on doit de fait trouver une racine de la fonction

$$f_n(x) = x - y_n - hg(t_{n+1}, x),$$

d'où le qualificatif « implicite ».

On peut procéder par itérations de point fixe, on peut également appliquer la méthode de Newton. Voyons ce que le théorème de Kantorovich a à nous dire à ce sujet. On a

$$\nabla f_n(x) = I - h\nabla g(t_{n+1}, x),$$

où la matrice jacobienne de g est prise par rapport à x , à $t = t_{n+1}$ fixé. C'est donc bien une matrice $d \times d$. On va supposer que $\nabla g(t_{n+1}, \cdot)$ est lipschitzienne de constante M pour la norme matricielle subordonnée, c'est-à-dire g de classe $C^{1,1}$.

On va évidemment démarrer l'itération de Newton à $x_0 = y_n$ (on est censé approcher y_{n+1} qui est censé approcher $y(t_{n+1})$ qui est proche de $y(t_n)$ qui est lui-même approché censément par y_n , donc c'est raisonnable). On a donc

$$f_n(x_0) = -hg(t_{n+1}, y_n) \text{ et } \nabla f_n(x_0)^{-1} f_n(x_0) = -h(I - h\nabla g(t_{n+1}, y_n))^{-1} g(t_{n+1}, y_n)$$

L'hypothèse de L -lipschitzianité de g se traduit bien sûr par

$$|||\nabla g(t_{n+1}, x)||| \leq L,$$

pour tout x et le caractère Lipschitzien de ∇g par

$$|||\nabla g(t_{n+1}, x) - \nabla g(t_{n+1}, y)||| \leq M\|x - y\|,$$

pour une autre constante de Lipschitz M .

On déduit de la première inégalité que $\nabla f_n(x)$ est inversible dès que $h < \frac{1}{L}$ avec

$$\|\nabla f_n(x)^{-1}\| = \|(I - h\nabla g(t_{n+1}, x))^{-1}\| \leq \frac{1}{1 - hL}.$$

Il vient alors

$$\begin{aligned} \|\nabla f_n(x_0)^{-1} f_n(x_0)\| &= h \|(I - h\nabla g(t_{n+1}, y_n))^{-1} g(t_{n+1}, y_n)\| \\ &\leq \frac{hG}{1 - hL}, \end{aligned}$$

où l'on a posé $G = \|g(t_{n+1}, y_n)\|$ (ou un majorant indépendant de n). Pour satisfaire la première condition du théorème de Kantorovich, il suffit donc que

$$\frac{hG}{1 - hL} \leq \frac{r}{2}. \quad (3.5.1)$$

De même,

$$\|(\nabla f_n(x_0))^{-1} (\nabla f_n(y) - \nabla f_n(z))\| \leq \frac{hM}{1 - hL} \|y - z\|,$$

et pour satisfaire la deuxième condition du théorème de Kantorovich, il suffit donc que

$$\frac{hM}{1 - hL} \leq \frac{1}{r}. \quad (3.5.2)$$

Il existe un tel r si et seulement si

$$\frac{2hG}{1 - hL} \leq \frac{1 - hL}{hM}.$$

Ceci a lieu si et seulement si $0 \leq h \leq \min(h_+, \frac{1}{L})$ où h_+ est la plus petite valeur positive de l'expression $\frac{L \pm \sqrt{2MG}}{L^2 - 2MG}$ (on peut toujours supposer que $L^2 - 2MG \neq 0$). Si l'inégalité de droite est stricte, il y a alors tout un intervalle de valeurs r possibles, ce qui implique que la convergence est quadratique, soit si g est C^2 , soit si l'on en croit la version horrible de Kantorovich.

Bien sûr, il s'agit de la convergence d'une itération de Newton, que l'on doit refaire avec une fonction différente pour chaque valeur de n correspondant à la discréttisation temporelle de l'EDO. En fonction de la difficulté des calculs, les considérations de coût peuvent alors devenir importantes.

3.5.1 Considérations de mise en œuvre pratique

Supposons que nous ayons un vrai problème $f(x) = 0$ de la vraie vie à résoudre par la méthode de Newton. On s'est donc donné un $x_0 \in \mathbb{R}^n$ (en croisant les doigts pour qu'il ne soit pas trop loin de la racine cherchée) et l'on doit calculer un nombre fini de termes de l'itération de Newton

$$x_{k+1} = x_k - (\nabla f(x_k))^{-1} f(x_k).$$

En pratique, « calculer » va bien sûr signifier approcher les valeurs de l'itération en question. Cette approximation est une source d'erreurs supplémentaires qui vont se propager à chaque étape de l'itération. On laisse de côté ici ces erreurs en les supposant négligeables¹⁸. On va donc devoir à chaque itération, à partir du $x_k \in \mathbb{R}^n$ obtenu à la fin de l'itération précédente

1. évaluer $f(x_k) \in \mathbb{R}^n$,
2. évaluer la matrice $\nabla f(x_k) \in M_n(\mathbb{R})$,
3. évaluer le produit matrice-vecteur $\nabla f(x_k)x_k$,
4. évaluer la différence $\nabla f(x_k)x_k - f(x_k)$,
5. résoudre le système linéaire $\nabla f(x_k)x_{k+1} = \nabla f(x_k)x_k - f(x_k)$.

Chacune de ces opérations a un coût, c'est-à-dire va demander plus ou moins d'opérations élémentaires (additions, multiplications, divisions, stockage en mémoire), et donc prendre plus ou moins de temps sur un ordinateur donné. Pour de petites valeurs de n et des fonctions f simples, ce temps peut être tellement bref qu'il semble que le résultat tombe instantanément. Néanmoins, n n'est pas forcément petit, f n'est pas forcément simple¹⁹, et surtout il se peut que le calcul de racine soit en fait inclus dans une ou plusieurs autres boucles itératives, et que l'on doive donc le répéter un très grand nombre de fois sur des valeurs différentes.

Dans ces conditions, la question du temps de calcul peut devenir primordiale²⁰.

Tentons d'évaluer à la louche le nombre d'opérations nécessaires pour chacune des étapes d'une itération.

1. On a n variables scalaires à combiner entre elles pour évaluer chaque composante de f . En général, cela va imposer d'effectuer au moins n opérations élémentaires sur des nombres (probablement beaucoup plus) pour chaque composante, d'où un coût global d'au moins $O(n^2)$ opérations.
2. Même chose, mais avec n^2 composantes, d'où un coût global d'au moins $O(n^3)$ opérations.
3. Là, on sait exactement ce que cela coûte : n^2 multiplications et $n(n - 1)$ additions, en l'absence de structure particulière de la matrice $\nabla f(x_k)$. D'où un coût global de $O(n^2)$ opérations.
4. L'étape la plus économique : n additions.

¹⁸. Que ceci soit justifié ou pas...

¹⁹. Par exemple, il se peut que l'on n'y ait accès qu'à travers un programme informatique lui-même complexe.

²⁰. Être capable de prédire le temps qu'il fera demain après un calcul qui prend un mois ne présente qu'un intérêt météorologique limité. Ceci dit, j'ignore si Météo France utilise la méthode de Newton ou non pour alimenter les bulletins météo des chaînes de télévision.

5. Encore une fois, en l'absence de structure particulière de la matrice $\nabla f(x_k)$, on n'a guère d'autre recours que d'utiliser la méthode de Gauss, ou une de ses déclinaisons plus modernes et sophistiquées, soit $O(n^3)$ opérations.²¹

Les étapes 3 et 4 présentent un coût négligeable par rapport aux étapes 2 et 5. Si f est « simple », il en va de même de l'étape 1. C'est sur les étapes 2 et 5, et en fait surtout 2, qu'il faut porter les efforts d'optimisation si le besoin s'en fait sentir. Cette observation conduit à l'idée des *méthodes de quasi-Newton*.

3.6 Les méthodes de quasi-Newton

Il s'agit simplement de remplacer la matrice jacobienne $\nabla f(x_k)$ par une autre matrice A_k inversible, selon une stratégie à préciser. On se donnera donc x_0 , puis on itérera

$$x_{k+1} = x_k - A_k^{-1} f(x_k). \quad (3.6.1)$$

Pour que ceci soit intéressant, il est préférable que les matrices A_k soient « faciles » à calculer et conduisent à des systèmes linéaires « faciles » à résoudre, sans pour autant détruire complètement les propriétés de convergence de la méthode.

Un exemple extrêmement simple, en dimension 1, consiste à prendre $A_k = f'(x_0) \neq 0$ pour tout k , que l'on ne calcule donc qu'une seule fois. Bien sûr, on est alors en train d'itérer la fonction $g(x) = x - f(x)/f'(x_0)$ dont tout point fixe est clairement racine de f . De fait, cette méthode converge si x_0 est suffisamment proche de x , manifestement d'autant plus vite que x_0 est proche de x .

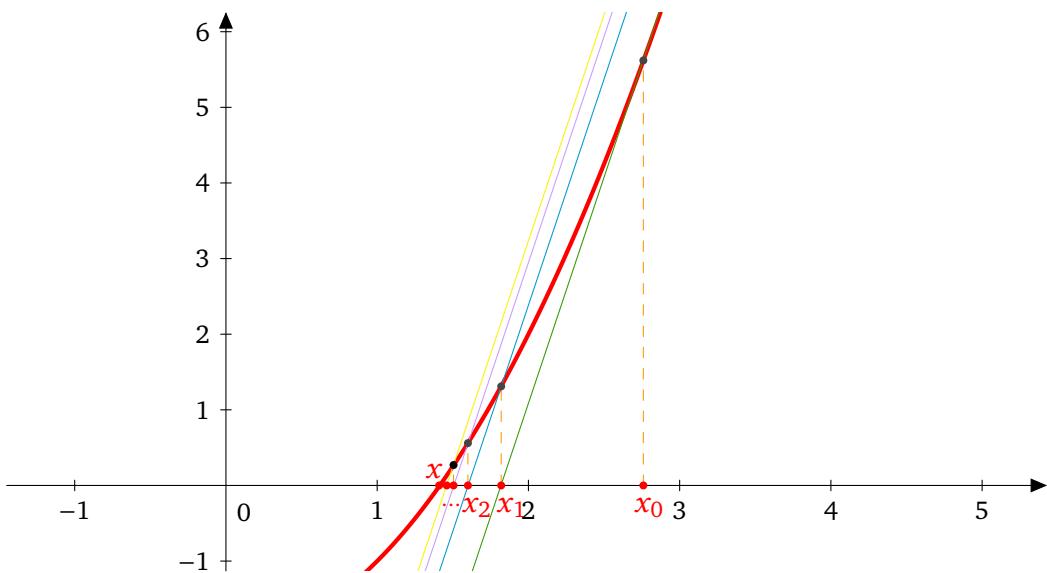


FIGURE 3.2 – Une méthode de quasi-Newton simpliste.

Dans cet exemple, on obtient semble-t-il une convergence linéaire, mais qui peut en pratique se révéler assez satisfaisante. On reviendra plus loin sur d'autres stratégies plus

²¹. Peut-être un poil moins pour certaines méthodes.

intelligentes. Montrons d'abord un résultat de convergence pour ces méthodes, en recopiant plus ou moins ce que l'on a déjà fait pour Newton.

Proposition 3.6.1 Soit f est de classe C^1 sur U , $x \in U$ une racine de f et A_k une suite de matrices de $GL_n(\mathbb{R})$ telle que la suite A_k^{-1} soit bornée. Soit M un majorant de $\{\|A_k^{-1}\|\}; k \in \mathbb{N}\}$. On suppose qu'il existe $r > 0$ et $0 \leq \beta < 1$ tels que pour tout $y \in B(x, r) \cap U$,

$$\|\nabla f(y) - A_k\| \leq \frac{\beta}{M}. \quad (3.6.2)$$

Alors il existe une boule centrée en x telle que pour toute donnée initiale dans cette boule, la suite des itérations de quasi-Newton est bien définie et converge au moins linéairement vers x , qui est de plus l'unique racine dans cette boule.

Démonstration. On pose $g_k(y) = y - A_k^{-1}f(y)$ et $h_k(y) = f(y) - A_k(y - x)$, qui sont définies sur U . En particulier, $x_{k+1} = g_k(x_k)$. Soit $B(x, r)$ la boule sur laquelle l'estimation (3.6.2) est supposée avoir lieu. En diminuant éventuellement r , on peut supposer que son adhérence, c'est-à-dire la boule fermée, est incluse dans U . Pour montrer que la suite est alors bien définie pour tout $x_0 \in B(x, r)$, il suffit de montrer que $g_k(B(x, r)) \subset B(x, r)$ pour tout k .

Comme $h_k(x) = 0$, on peut écrire pour tout $y \in B(x, r)$,

$$\begin{aligned} g_k(y) - x &= y - x - A_k^{-1}f(y) \\ &= A_k^{-1}(A_k(y - x) - f(y)) \\ &= A_k^{-1}(h_k(y) - h_k(x)). \end{aligned}$$

On a $\nabla h_k(y) = \nabla f(y) - A_k$. Par conséquent, par l'inégalité des accroissements finis,

$$\begin{aligned} \|g_k(y) - x\| &\leq \|A_k^{-1}\| \|h_k(y) - h_k(x)\| \\ &\leq \|A_k^{-1}\| \sup_{B(x, r)} \|\nabla h_k(y)\| \|y - x\| \\ &\leq \beta \|y - x\| < r, \end{aligned}$$

et cela pour tout k . On a donc bien $g_k(y) \in B(x, r)$ et la suite est bien définie.

Par un calcul très voisin, on note qu'alors

$$\begin{aligned} f(x_k) &= f(x_k) - f(x_{k-1}) - A_{k-1}(x_k - x_{k-1}) \\ &= h_{k-1}(x_k) - h_{k-1}(x_{k-1}), \end{aligned}$$

pour tout $k \geq 1$. Il vient donc

$$\|f(x_k)\| \leq \frac{\beta}{M} \|x_k - x_{k-1}\|. \quad (3.6.3)$$

Comme par ailleurs,

$$x_{k+1} - x_k = -A_k^{-1}f(x_k)$$

on en déduit que

$$\|x_{k+1} - x\| \leq \|A_k^{-1}\| \|f(x_k)\| \leq \beta \|x_k - x_{k-1}\|$$

pour tout $k \geq 1$. Il s'ensuit, comme dans la preuve du théorème de point fixe de Banach, que la suite x_k est de Cauchy. Elle converge donc vers un certain \bar{x} dans l'espace complet $B(x, r)$, et la convergence est linéaire, encore une fois comme dans la preuve du théorème de point fixe de Banach. Par l'estimation (3.6.3) et en utilisant le fait que f est continue, on en déduit que $f(\bar{x}) = 0$.

Pour conclure, on remarque que

$$x - \bar{x} = -A_0^{-1}(f(x) - f(\bar{x}) - A_0(x - \bar{x})) = -A_0^{-1}(h_0(x) - h_0(\bar{x})),$$

d'où à nouveau

$$\|x - \bar{x}\| \leq \beta \|x - \bar{x}\|,$$

ce qui implique que $\bar{x} = x$ puisque $\beta < 1$. \diamond

Remarque 3.6.1 i) La méthode de Newton consiste à prendre $A_k = \nabla f(x_k)$. L'estimation (3.6.2) est alors assurée sur une boule assez petite par continuité uniforme de ∇f sur toute boule fermée. On établit donc ici la convergence de la méthode de Newton pour f de classe C^1 , mais pas nécessairement C^2 . Par contre, dans ce cas en l'absence de régularité au delà de C^1 , on ne peut guère espérer récupérer la convergence quadratique.

ii) Cette condition dit plus généralement que, d'une certaine façon, A_k ne doit pas être trop éloigné des valeurs prises par ∇f dans la boule.

iii) Il existe une version plus quantitative du résultat, plus dans l'esprit du théorème de Kantorovich, théorème que l'on devrait avoir le temps de voir plus tard.

iv) Les méthodes de quasi-Newton sont très utilisées dans le contexte de l'optimisation. Soit F une fonction de \mathbb{R}^n à valeurs dans \mathbb{R} que l'on cherche à maximiser ou minimiser. Si F est de classe C^1 , une condition nécessaire pour avoir un point x de minimum ou de maximum local est que $\nabla F(x) = 0$. Une stratégie d'optimisation peut donc consister à chercher les racines de la fonction $f(x) = \nabla F(x)$ qui est bien une fonction de \mathbb{R}^n à valeurs dans \mathbb{R}^n . Si F est de classe C^2 , alors on pourra penser utiliser la méthode de Newton (dont la convergence quadratique est assurée localement si F est de classe C^3) ou bien des méthodes de quasi-Newton.

Dans ce cas, $\nabla f = \nabla^2 F$ est la *hessienne* de F , qui peut être très difficile à calculer pour des F compliquées, et conduire à des systèmes linéaires difficiles à résoudre. Les méthodes de quasi-Newton prennent tout leur intérêt ici, d'où leur popularité en optimisation. En effet, on dispose de très nombreuses façons de se donner directement les matrices $B_k = A_k^{-1}$ à partir des quantités antérieurement calculées, ce qui élimine totalement l'étape de résolution de système linéaire, qui est alors remplacée par un produit matrice-vecteur. Il arrive souvent que la perte de vitesse de convergence par rapport à la méthode de Newton soit plus que compensée par la simplification des autres calculs, ce qui rend *in fine* les méthodes de quasi-Newton plus performantes dans ce contexte que la méthode de Newton.

v) Pour revenir au contexte général de l'équation $f(x) = 0$, une stratégie possible consiste à prendre $A_{k+p} = \nabla f(x_k)$ pour $1 \leq p \leq m$, puis de passer à la matrice jacobienne suivante $A_{k+m+p} = \nabla f(x_{k+m})$ et ainsi de suite. L'intérêt de cette version de quasi-Newton est de quand même adapter la matrice de l'itération à la matrice jacobienne, mais seulement de temps à autre. Entre les deux, on ne recalcule pas la matrice, et on peut de plus en utiliser une factorisation de type *LU* pour que toutes les résolutions de systèmes linéaires intermédiaires

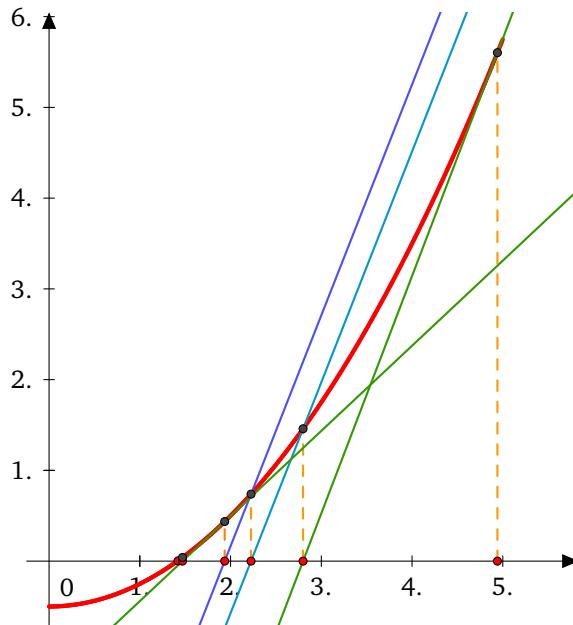


FIGURE 3.3 – Illustration de la dernière méthode de quasi-Newton.

se fassent en seulement $O(n^2)$ opérations. En choisissant bien m , voire en le faisant varier, on peut aboutir à des compromis performance/coût intéressants.