

*To explain all nature is too difficult a task for any one man  
or even for any one age. 'Tis much better to do a little with certainty,  
and leave the rest for others that come after you, than to explain all things.*  
Isaac Newton (1642-1727)

## CHAPITRE XVII

# Méthodes itératives pour la résolution d'équations

**Objectif.** Les méthodes itératives figurent parmi les méthodes numériques les plus courantes et le plus puissantes. L'idée est de partir d'une valeur approchée (souvent grossière) de la solution, puis d'augmenter la précision par l'application itérée d'un algorithme bien choisi.

Dans ce chapitre nous discutons deux méthodes itératives classiques : la méthode du point fixe pour résoudre une équation du type  $f(x) = x$  où  $f$  est une fonction contractante, puis son raffinement, la méthode de Newton pour résoudre  $f(x) = 0$  où  $f$  est une fonction dérivable. Pour des compléments voir le livre de J.-P. Demailly, *Analyse numérique et équations différentielles*, EDP Sciences, 1996.

### Sommaire

- 1. La méthode du point fixe.** 1.1. Dynamique autour d'un point fixe. 1.2. Espaces métriques. 1.3. Fonctions contractantes. 1.4. Le théorème du point fixe. 1.5. Quelques applications.
- 2. La méthode de Newton.** 2.1. Vitesse de convergence. 2.2. Itération de Newton. 2.3. Exemples. 2.4. Bassin d'attraction. 2.5. Version quantitative. 2.6. Critères pratiques.
- 3. Application aux polynômes complexes.** 3.1. Le théorème de Gauss-d'Alembert. 3.2. Relation entre racines et coefficients. 3.3. Instabilité des racines mal conditionnées.

### Retour sur la méthode dichotomique

Commençons par la méthode « par tâtonnement » que l'on appelle plus savamment « la méthode dichotomique ». Cette méthode a le mérite d'être élémentaire, mais elle a deux inconvénients : d'abord elle ne converge que lentement, puis implémentée naïvement (avec des arrondis aléatoires) elle peut être inutilisable à cause des phénomènes de bruit, déjà discutés au chapitre XV, §5.6.

**Exemple 0.1.** Le programme `dichotomie.cc` résout le problème de bruit par un calcul fiable basé sur l'arithmétique d'intervalles. (Le lire puis le tester.) La difficulté principale est d'implémenter soigneusement la fonction donnée  $f: \mathbb{R} \rightarrow \mathbb{R}$  selon l'arithmétique d'intervalles arrondie en une fonction

```
Interval f( const Interval& x ) .
```

D'une part il faut garantir l'encadrement de l'image exacte, et d'autre part cet encadrement doit être assez précis, c'est-à-dire on veut éviter des sur-encadrements grossiers. Discuter l'importance de ces prérequis pour la correction et l'efficacité de la méthode dichotomique. Puis analyser brièvement comment satisfaire ces exigences pour un polynôme comme  $f(x) = x^6 - 9x^5 + 30x^4 - 40x^3 + 48x - 32$ .

**Exemple 0.2.** La méthode dichotomique se généralise du cas unidimensionnel  $f: \mathbb{R} \rightarrow \mathbb{R}$  aux fonctions  $f: \mathbb{R}^m \rightarrow \mathbb{R}^n$ , disons avec  $m = n = 2$  pour fixer les idées. Si cela vous intéresse, vous pouvez regarder le film [www-sop.inria.fr/coprin/logiciels/ALIAS/Movie/movie\\_undergraduate.mpg](http://www-sop.inria.fr/coprin/logiciels/ALIAS/Movie/movie_undergraduate.mpg) puis analyser la méthode proposée. Essayez d'abord de décrire l'algorithme plus en détail : comme avant la principale difficulté est de bien implémenter  $f$  selon l'arithmétique d'intervalles arrondie. Étant donnée une telle implémentation de  $f$ , formulez l'algorithme dichotomique puis prouver sa correction. (Si vous êtes courageux, vous pouvez l'implémenter en partant du programme `dichotomie.cc`.)

Dans ce chapitre nous chercherons à obtenir des méthodes itératives qui convergent plus rapidement et qui seront plus faciles à implémenter que la méthode dichotomique. Nous développons l'essentiel de la théorie, le théorème du point fixe et la méthode de Newton, sous une forme prête à programmer. Par contre, la problématique des calculs fiables sera (temporairement) négligé afin de ne pas encombrer la première présentation. Ceci est partiellement justifié par le fait qu'une fonction contractante soit numériquement stable, et lors des itérations les erreurs accumulées restent bornées (et on espère même négligeables).

## 1. La méthode du point fixe

**1.1. Dynamique autour d'un point fixe.** Nous allons nous intéresser aux itérations d'une fonction  $f: E \rightarrow E$ , où  $(E, d)$  est un espace métrique. Typiquement  $E$  est une partie de  $\mathbb{R}^m$  et  $d: E \times E \rightarrow \mathbb{R}$  est la distance euclidienne  $d(x, y) = \sqrt{\sum_i (x_i - y_i)^2}$ . La fonction  $f$  peut être compliquée, d'autant plus ses itérées  $f^2 = f \circ f$ ,  $f^3 = f \circ f \circ f$ ,  $f^4 = f \circ f \circ f \circ f$ , ... Nous ne calculons jamais ces itérées toutes entières. Nous supposons seulement qu'il est facile à évaluer  $f$  en un point donné, et nous allons suivre sa trajectoire :

**Définition 1.1.** Étant donnée  $f: E \rightarrow E$  et une valeur initiale  $x_0$  nous construisons la suite itérative  $(x_n)_{n \in \mathbb{N}}$  partant de  $x_0$  par l'application itérée de la fonction  $f$ , de sorte que  $x_{n+1} = f(x_n)$  pour tout  $n \in \mathbb{N}$ .

Le cas trivial est celui d'un point fixe de  $f$ , c'est-à-dire d'un point  $a \in E$  tel que  $f(a) = a$ . Par contre, le comportement de  $f$  dans un voisinage d'un point fixe  $a$  peut nous donner des renseignements utiles :

**Exemple 1.2.** L'exemple le plus simple est une fonction linéaire  $f: \mathbb{R} \rightarrow \mathbb{R}$ ,  $f(x) = kx$  avec une constante  $k \in \mathbb{R}$ . Elle admet  $a = 0$  pour point fixe. Pour  $x_0 \neq 0$  deux phénomènes peuvent se produire :

- (1) Si  $|k| < 1$ , par exemple  $k = \frac{1}{2}$ , alors les images successives  $x_n = f^n(x_0)$  s'approchent de  $a$ , car  $|f^n(x_0) - a| = |k|^n \cdot |x_0 - a| \rightarrow 0$ . On dit que  $a$  est un point fixe attractif (ou stable).
- (2) Si  $|k| > 1$ , par exemple  $k = 2$ , alors les images successives  $x_n = f^n(x_0)$  s'éloignent de  $a$ , car  $|f^n(x_0) - a| = |k|^n \cdot |x_0 - a| \rightarrow \infty$ . On dit que  $a$  est un point fixe répulsif (ou instable).

**Exemple 1.3.** Plus généralement, étudions une fonction  $f: \mathbb{R} \rightarrow \mathbb{R}$  que l'on suppose continûment dérivable. Comme avant nous supposons qu'elle admet un point fixe  $a = f(a)$ .

- (1) Supposons que  $|f'(a)| < 1$ . Dans ce cas on peut choisir n'importe quelle constante  $k$  telle que  $|f'(a)| < k < 1$ , et la continuité de  $f'$  nous assure l'existence d'un voisinage  $V = [a - \varepsilon, a + \varepsilon]$  tel que  $|f'(\xi)| \leq k$  pour tout  $\xi \in V$ . Pour tout  $x \in V$  nous pouvons ainsi majorer les accroissements : il existe  $\xi$  entre  $a$  et  $x$  tel que  $f(x) - f(a) = f'(\xi)(x - a)$ , et par conséquent

$$|f(x) - a| = |f(x) - f(a)| = |f'(\xi)(x - a)| \leq k|x - a|.$$

Ainsi les images itérées de  $x \in V$  sont de plus en plus proches de  $a$ , plus précisément :

$$|f^n(x) - a| \leq k^n |x - a| \quad \text{pour tout } n \in \mathbb{N}.$$

Autrement dit,  $a$  est un point fixe attractif.

- (2) Réciproquement l'inégalité  $|f'(a)| > 1$  implique que  $|f'| \geq k > 1$  dans un voisinage  $V$  de  $a$  : toute valeur initiale  $x \in V \setminus \{a\}$  s'éloigne de  $a$ , dans le sens que  $|f(x) - a| \geq k|x - a|$ . Autrement dit,  $a$  est un point fixe répulsif.

☞ *Astuce.* — Dans ce cas on peut inverser  $f$ , au moins localement : la restriction  $f|_V: V \rightarrow U$  est une bijection de  $V$  sur  $U = f(V)$  ; son inverse  $g = f|_V^{-1}$  est une fonction continûment dérivable avec  $g(a) = a$  et  $g'(a) = \frac{1}{f'(a)}$ . (Le prouver.) On peut ainsi se ramener au premier cas.

- (3) Dans le cas douteux  $|f'(a)| = 1$ , une analyse plus fine s'impose. Pour  $x \mapsto x - x^3$  le point fixe  $a = 0$  est attractif, pour  $x \mapsto x + x^3$  il est répulsif. (Le détailler.) Pour  $x \mapsto x + x^2$  il est attractif à gauche mais répulsif à droite, pour  $x \mapsto x - x^2$  c'est l'inverse. (Le détailler.)

**Remarque 1.4.** En dimension  $\geq 2$  la situation peut être plus complexe. Considérons une application linéaire  $\begin{pmatrix} x \\ y \end{pmatrix} \mapsto \begin{pmatrix} \lambda & 0 \\ 0 & \mu \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$ . Si  $|\lambda|, |\mu| < 1$ , le point fixe  $a = 0$  est attractif. Si  $|\lambda|, |\mu| > 1$ , il est répulsif. Si  $|\lambda| < 1 < |\mu|$ , il existe une direction stable et une direction instable.

**1.2. Espaces métriques.** Étant donné un ensemble  $E$  et une suite  $(x_n)_{n \in \mathbb{N}}$  dans  $E$ , comment définir la notion de convergence ? La façon la plus commode serait de disposer d'une notion de distance  $d(x, y)$  entre deux points  $x, y \in E$ , c'est-à-dire une application  $d: E \times E \rightarrow \mathbb{R}$ . Pour que la distance se comporte comme on le souhaite, nous exigeons les trois propriétés suivantes :

**Positivité:**  $d(x, y) \geq 0$  pour tout  $x, y \in E$ , avec  $d(x, y) = 0$  si et seulement si  $x = y$ .

**Symétrie:**  $d(x, y) = d(y, x)$  pour tout  $x, y \in E$ .

**Inégalité triangulaire:**  $d(x, z) \leq d(x, y) + d(y, z)$  pour tout  $x, y, z \in E$ .

Si  $d: E \times E \rightarrow \mathbb{R}$  satisfait à ces axiomes, on appelle  $d$  une *métrique* sur  $E$ , et la paire  $(E, d)$  est appelée un *espace métrique*. (Souvent on appelle  $E$  un espace métrique si  $d$  est sous-entendue sans équivoque.)

**Exemple 1.5.** Tout ensemble  $E$  peut être muni de la *métrique discrète*  $d: E \times E \rightarrow \mathbb{R}$  définie par  $d(x, y) = 0$  si  $x = y$ , et  $d(x, y) = 1$  si  $x \neq y$ . (Exercice.) C'est gratuit et peu intéressant.

**Exemple 1.6.** Sur  $\mathbb{R}$  on a bien sûr la métrique usuelle  $d(x, y) = |x - y|$ . L'ensemble  $\mathbb{R}^m$  peut être muni de la métrique euclidienne  $d_2(x, y) = \sqrt{\sum_k (x_k - y_k)^2}$ . Plus généralement, pour tout  $p \geq 1$ , on obtient une métrique  $d_p(x, y) = (\sum_k |x_k - y_k|^p)^{1/p}$ . Pour  $p = \infty$  on pose  $d_\infty(x, y) = \sup_k |x_k - y_k|$ .

(Pour ces métriques la positivité et la symétrie sont immédiates, mais l'inégalité triangulaire est délicate. C'est un bon exercice de révision si vous l'avez déjà vu.)

**Exemple 1.7.** Si  $(E, d)$  est un espace métrique et  $F \subset E$  est une partie, alors  $(F, d_F)$  est un espace métrique muni de la métrique *induite*  $d_F := d|_{F \times F}: F \times F \rightarrow \mathbb{R}$  obtenue par restriction. Ainsi toute partie de  $\mathbb{R}^m$  est munie de la métrique  $d_p$  héritée de  $\mathbb{R}^m$ .

**Définition 1.8.** Soit  $(E, d)$  un espace métrique. On dit qu'une suite  $(x_n)_{n \in \mathbb{N}}$  dans  $E$  *converge* vers  $a \in E$  (pour la métrique  $d$ ) si la distance  $d(x_n, a)$  converge vers 0. Plus explicitement : pour tout  $\varepsilon > 0$  il existe  $N \in \mathbb{N}$  tel que  $d(x_n, a) < \varepsilon$  pour tout  $n \geq N$ .

*Exercice 1.9.* Vérifier que sur  $\mathbb{R}^1$  toutes les métriques  $d_p$  coïncident. Dans le plan  $\mathbb{R}^2$  dessiner la « boule »  $B(0, 1) = \{x \in \mathbb{R}^2 \mid d(x, 0) \leq 1\}$  de rayon 1 autour de 0 pour les métriques  $d_1, d_2, d_\infty$ . Bien que ces métriques soient distinctes, la notion de convergence est la même : une suite  $x_n$  dans  $\mathbb{R}^2$  converge vers  $a$  pour la métrique  $d_p$  si et seulement si  $x_n$  converge vers  $a$  pour la métrique  $d_q$ . Il existe des constantes  $\alpha, \beta > 0$  telles que  $\alpha d_p(x, y) \leq d_q(x, y) \leq \beta d_p(x, y)$ . On dit que ces métriques sont *équivalentes*.

**1.3. Fonctions contractantes.** Souvent l'équation à résoudre se présente (ou peut se reformuler) comme un problème de point fixe  $f(x) = x$ . Cette approche se prête à une vaste généralité qui s'avère très utile. Soulignons d'abord le caractère purement *métrique* :

**Définition 1.10.** Soit  $(E, d)$  un espace métrique, soit  $f: E \rightarrow E$  une application, et soit  $k < 1$  une constante. On dit que  $f$  est *contractante* de rapport  $k$  si  $d(f(x), f(y)) \leq kd(x, y)$  pour tout  $x, y \in E$ . Autrement dit, l'application  $f$  rapproche les points au moins de rapport  $k$  (toujours avec  $k < 1$  fixé).



**Remarque 1.11.** Soit  $I \subset \mathbb{R}$  un intervalle et soit  $f: I \rightarrow \mathbb{R}$  une fonction dérivable. On pose  $k := \sup_I |f'|$ . Alors  $f$  est contractante si  $k < 1$ . Effectivement, pour tout  $x, y \in I$  le théorème des accroissements finis nous assure qu'il existe  $\xi$  entre  $x$  et  $y$  tel que  $f(x) - f(y) = f'(\xi)(x - y)$ , donc  $|f(x) - f(y)| \leq k|x - y|$ .

*Exercice 1.12.* Prouver un critère analogue pour  $f: I \rightarrow \mathbb{R}^m$  sur une partie convexe  $I \subset \mathbb{R}^m$ .  
*Attention.* — Il ne suffit pas de supposer  $I \subset \mathbb{R}^m$  connexe. (Esquisser un contre-exemple.)

*Exercice 1.13.* Comme on a vu, la convergence d'une suite  $x_n$  vers un point  $x$  est une propriété topologique : elle ne change pas lorsqu'on remplace la métrique  $d$  par une métrique équivalente. Par contre la propriété d'une fonction d'être contractante dépend de la métrique. En s'inspirant de la figure ci-dessus, regardons la fonction affine  $f: \mathbb{R}^2 \rightarrow \mathbb{R}^2$  donnée par  $f \begin{pmatrix} x \\ y \end{pmatrix} = k \begin{pmatrix} \cos \alpha & -\sin \alpha \\ \sin \alpha & \cos \alpha \end{pmatrix} \cdot \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} a \\ b \end{pmatrix}$ . Vérifier que pour la métrique  $d_2$  cette application est contractante si et seulement si  $k < 1$ . Pour  $\alpha = \frac{\pi}{4}$  et  $k$  proche de 1, l'application  $f$  n'est pas contractante par rapport à la métrique  $d_1$ , ni par rapport à la métrique  $d_\infty$ .

**1.4. Le théorème du point fixe.** Intuitivement, une fonction contractante  $f: E \rightarrow E$  se contracte sur un point, l'unique point fixe de  $f$ . Le but de ce paragraphe est justement d'établir ce théorème, dit de point fixe. Malheureusement l'intuition trompe facilement : le point crucial est l'*existence* d'un point fixe.

**Exemple 1.14.** Comme on a vu au chapitre XV, l'itération de la fonction  $f(x) = \frac{1}{2}(x + \frac{2}{x})$  permet d'approcher la valeur de  $\sqrt{2}$ , l'unique point fixe de  $f$ . Restreint à l'intervalle  $I = [1, +\infty[$  il s'agit d'une fonction contractante car la dérivée  $f'(x) = \frac{1}{2} - \frac{1}{x^2}$  est à valeur dans  $[-\frac{1}{2}, \frac{1}{2}[$ .

Dans la pratique on ne calcule qu'avec des nombres rationnels, et sur  $E = [1, +\infty[ \cap \mathbb{Q}$  la restriction  $h = f|_E: E \rightarrow E$  est toujours contractante. Mais, malheureusement, elle n'a plus de point fixe dans  $E$  : le problème est que l'espace  $E$  contienne des « trous », il n'est pas ce que l'on appelle *complet*.

**Définition 1.15.** Une suite  $(x_n)_{n \in \mathbb{N}}$  dans un espace métrique  $(E, d)$  est dite *de Cauchy* si pour tout  $\varepsilon > 0$  il existe  $N \in \mathbb{N}$  tel que  $d(x_n, x_m) < \varepsilon$  pour tout  $n, m \geq N$ .

Toute suite convergente est de Cauchy mais la réciproque est fautive en général : regarder une suite dans  $\mathbb{Q}$  qui converge vers  $\sqrt{2} \in \mathbb{R}$ . On arrive ainsi à la définition suivante :

**Définition 1.16.** Un espace métrique  $(E, d)$  est *complet* si toute suite de Cauchy converge.

**Remarque 1.17.** L'espace  $\mathbb{R}$  avec la métrique usuelle est complet. L'espace  $\mathbb{R}^m$  avec la métrique  $d_p$  est complet, quelque soit  $p \in [1, \infty[$ . Il en est de même pour tout sous-ensemble *fermé*. (Le montrer.) Par exemple, tout intervalle fermé  $[a, b] \subset \mathbb{R}$  avec la métrique induite est un espace métrique complet. Par contre  $\mathbb{Q} \subset \mathbb{R}$  n'est pas complet.

Les espaces métriques complets sont très importants en analyse parce qu'ils permettent de *construire* certains objets comme limites de suites de Cauchy, l'existence étant assurée par l'hypothèse de complétude. En voici un exemple fondamental aussi bien pour la théorie que pour le calcul numérique :

**Théorème 1.18** (le théorème du point fixe). Soit  $(E, d)$  un espace métrique complet et soit  $f : E \rightarrow E$  une application contractante de rapport  $k < 1$ . Alors :

- (1) Il existe un et un seul point  $a \in E$  vérifiant  $f(a) = a$ .
- (2) Pour tout  $x_0 \in E$  la suite itérative  $x_{n+1} = f(x_n)$  converge vers  $a$ , vérifiant  $d(x_n, a) \leq k^n d(x_0, a)$ . La convergence est donc au moins aussi rapide que celle de la suite géométrique  $k^n \rightarrow 0$ .
- (3) Pour le calcul concret on a la majoration d'erreur  $d(x_n, a) \leq \frac{k}{1-k} d(x_n, x_{n-1})$ .

*Remarque.* — Pour le calcul concret on suppose, comme dans les exemples précédents, que l'on sache calculer  $x_{n+1} = f(x_n)$  à partir de  $x_n$ . Par contre on ignore typiquement la valeur limite  $a$ . Dans cette situation l'inégalité  $d(x_n, a) \leq \frac{k}{1-k} d(x_n, x_{n-1})$  nous est très utile car elle permet de majorer la distance de la valeur approchée  $x_n$  à la valeur inconnue  $a$  en fonction de  $x_n$  et  $x_{n-1}$  seulement. Contrairement à  $d(x_n, a)$ , la quantité  $\frac{k}{1-k} d(x_n, x_{n-1})$  est en général très facile à calculer.

**DÉMONSTRATION.** *Unicité.* — Si  $a, b \in E$  sont deux points fixes d'une fonction contractante de rapport  $k < 1$ , alors  $d(a, b) = d(f(a), f(b)) \leq kd(a, b)$  entraîne  $d(a, b) = 0$  donc  $a = b$ .

*Existence.* — Une récurrence facile montre  $d(x_{n+1}, x_n) \leq k^n d(x_1, x_0)$  pour tout  $n \in \mathbb{N}$ , puis

$$\begin{aligned} d(x_{n+p}, x_n) &\leq d(x_{n+p}, x_{n+p-1}) + \dots + d(x_{n+2}, x_{n+1}) + d(x_{n+1}, x_n) \\ &\leq (k^{p-1} + \dots + k^1 + k^0) d(x_{n+1}, x_n) \leq \frac{k^n}{1-k} d(x_1, x_0) \end{aligned}$$

pour tout  $n, p \in \mathbb{N}$ . La suite  $(x_n)_{n \in \mathbb{N}}$  est donc de Cauchy et converge puisque  $E$  est complet. Notons  $a$  sa limite, et vérifions qu'il s'agit d'un point fixe : l'application  $f$ , étant contractante, est continue. L'équation de récurrence  $x_{n+1} = f(x_n)$  donne donc  $a = \lim x_{n+1} = \lim f(x_n) = f(\lim x_n) = f(a)$ .

*Vitesse de convergence.* — On a  $d(x_n, a) \leq k^n d(x_0, a)$ , donc pour toute valeur initiale  $x_0 \in E$  la suite itérative  $x_{n+1} = f(x_n)$  converge vers le point fixe  $a$ . Pour estimer la distance de  $x_n$  à  $a$ , on a la majoration  $d(x_{n+p}, x_n) \leq \frac{k}{1-k} d(x_n, x_{n-1})$ . Le passage à la limite  $p \rightarrow \infty$  donne l'inégalité cherchée.  $\square$

**Exemple 1.19.** Dans notre exemple préféré la fonction  $f : [1, 2] \rightarrow [1, 2]$ ,  $f(x) = \frac{1}{2}(x + \frac{2}{x})$  est contractante de rapport  $k = \frac{1}{2}$ . Le théorème affirme donc que pour toute valeur initiale  $x_0 \in [1, 2]$  la suite itérative  $x_{n+1} = f(x_n)$  converge vers l'unique point fixe  $a = \sqrt{2}$ , et que  $|x_n - a| \leq (\frac{1}{2})^n |x_0 - a|$ . (La convergence est en fait beaucoup plus rapide que celle de la suite géométrique  $(\frac{1}{2})^n \rightarrow 0$ , voir plus bas.)

**Remarque 1.20.** La majoration  $d(x_n, a) \leq k^n d(x_0, a)$ , donne  $-\log d(x_n, a) \geq -\log d(x_0, a) - n \log(k)$ . Dans le cas  $E = \mathbb{R}$  le nombre de décimales exactes de  $x_n$  comme approximation de  $a$  est donc minoré par une fonction affine de  $n$ . On parle d'une *convergence linéaire*. On remarquera que la constante de contraction  $k < 1$  influence sensiblement la vitesse de convergence.

**Exemple 1.21.** Soulignons que la majoration  $d(x_n, x_{n-1}) < \varepsilon$  n'implique pas forcément que  $d(x_n, a) < \varepsilon$ . Il faut tenir compte du facteur  $\frac{k}{1-k}$  : appliquons la condition d'arrêt  $|x_n - x_{n-1}| < 10^{-4}$  à l'itération de  $f : \mathbb{R} \rightarrow \mathbb{R}$ ,  $f(x) = 0,999999 \cdot x$  et la valeur initiale  $x_0 = 100$ . Cette fonction est contractante, mais seulement très faiblement :  $x_0 = 100$  et  $x_1 = 99,9999$  sont proches, mais encore loin du point fixe  $a = 0$ .

**Exemple 1.22.** Pour appliquer le théorème il faut soigneusement vérifier les hypothèses. À titre d'avertissement, voici quelques exemples voisins où le théorème ne s'applique pas :

- (1)  $f: [0, 1] \rightarrow \mathbb{R}$ ,  $f(x) = 1 + \frac{1}{2}x$  est contractante mais n'admet pas de point fixe. Pourquoi ?
- (2)  $g: ]0, 1[ \rightarrow ]0, 1[$ ,  $g(x) = \frac{1}{2}x$  est contractante mais n'admet pas de point fixe. Pourquoi ?
- (3) Sur  $E = [1, 2] \cap \mathbb{Q}$  la fonction  $h(x) = \frac{1}{2}(x + \frac{2}{x})$  n'admet pas de point fixe. Pourquoi ?
- (4) La fonction  $f: \mathbb{R} \rightarrow \mathbb{R}$ ,  $f(x) = x + \frac{1}{1+e^x}$  vérifie  $0 < f'(x) < 1$ . Est-elle contractante ? sur un compact  $[a, b]$  ? Admet-elle un point fixe ? Pourquoi le théorème ne s'applique-t-il pas ?

**1.5. Quelques applications.** Pour  $f(a) = a$  et  $|f'(a)| < 1$  nous avons vu qu'il existe un voisinage  $V = [a - \varepsilon, a + \varepsilon]$  sur lequel la fonction  $f$  est contractante de rapport  $k = \sup_V |f'| < 1$ . Pour tout  $x \in V$  nous avons  $|f(x) - a| \leq k|x - a|$ , donc  $f(V) \subset V$ , et nous pouvons appliquer le théorème : pour tout  $x_0 \in V$  on a convergence  $x_n \rightarrow a$  de vitesse (au moins) linéaire  $|x_n - a| \leq k^n |x_0 - a|$ .

**Remarque 1.23.** Dans la pratique, on se donne souvent une fonction  $f: \mathbb{R} \rightarrow \mathbb{R}$ . Il faut d'abord expliciter un fermé  $V \subset \mathbb{R}$  tel que  $f(V) \subset V$  et une constante  $k$  telle que  $\sup_V |f'| \leq k < 1$ . Ceci correspond à localiser le point fixe  $a$  en explicitant un voisinage  $V$  qui soit suffisamment précis pour appliquer le théorème.

☞ Plus la constante  $k$  est petite, plus la convergence  $k^n \rightarrow 0$  est rapide. Comme  $k$  est minorée par  $|f'(a)|$ , on a intérêt à choisir plutôt un *petit* voisinage  $V$  de  $a$ , si possible, afin de minimiser  $k$ .

☞ Si  $\sup_V |f'| \geq 1$ , alors on a choisi  $V$  trop grand et il faut recommencer avec un fermé mieux adapté. Si l'on sait d'avance que  $V$  ne contient pas de point fixe, ou si le point fixe  $a$  vérifie  $|f'(a)| \geq 1$ , il est inutile d'insister : le théorème ne pourra pas s'appliquer.

**Exemple 1.24.** On se propose d'approcher l'unique solution  $x \in \mathbb{R}$  de l'équation  $x = \cos x$ .

*Exercice préparatoire.* — Tracer sommairement le graphe de  $\cos x$  afin de localiser la solution. (Il n'y a qu'une seule.) Déterminer un intervalle  $[a, b]$  sur lequel le théorème du point fixe s'applique. Quelle constante de contraction obtenez-vous ?

*Calcul numérique.* — Le programme `iter1.cc` calcule la suite itérative  $x_{n+1} = \cos x_n$  à partir de  $x_0 = 0$ . La suite converge numériquement, mais cette observation empirique ne remplace pas l'analyse mathématique précédente. Étant donnée la constante de contraction  $k$ , le programme peut déterminer la marge d'erreur  $\frac{k}{1-k} |x_n - x_{n-1}|$ . Ainsi on itère jusqu'à la précision  $\varepsilon$  souhaitée, et on peut conclure que  $|x_n - x| \leq \varepsilon$ .

**Exercice 1.25.** Encadrer les solutions réelles de l'équation  $\exp(x) = x^3$  à  $10^{-8}$  près.

*Indication.* — Reformuler le problème sous forme de point fixe,  $f(x) = x$ . Tracer sommairement le graphe de  $f$  ; il y a deux solutions. Les points fixes sont-ils attractifs ou répulsifs ? Dans le cas attractif, appliquer le théorème du point fixe. Dans le cas répulsif passer à la fonction inverse  $f^{-1}$ .

*Exercice/M 1.26.* On se propose d'analyser l'application  $f: \mathbb{R}^2 \rightarrow \mathbb{R}^2$ ,  $f\left(\begin{smallmatrix} x_1 \\ x_2 \end{smallmatrix}\right) = \left(\begin{smallmatrix} -x_1 + \frac{3}{2}x_2 + \frac{5}{4} \\ -\frac{1}{2}x_1 + x_2 + \frac{3}{4} \end{smallmatrix}\right)$ .

Déterminer les points fixes de  $f$ , puis calculer la dérivée  $\left(\frac{\partial f}{\partial x_j}\right)$  dans ces points. Quelles sont les valeurs propres ? Les points fixes sont-ils attractifs ? Soit  $x$  le point fixe non attractif de  $f$ . Montrer que  $f$  est localement inversible autour de  $x$ , c'est-à-dire il existe un voisinage  $V$  de  $x$  tel que  $f|_V: V \rightarrow U$  soit une bijection sur  $U := f(V)$ , et que l'inverse  $g = f|_V^{-1}: U \rightarrow V$  soit une application continûment dérivable. Est-ce que le point fixe  $x$  est attractif pour  $g$  ?

*Exercice/M 1.27.* Quels sont les points fixes de l'application  $f: \mathbb{R}_+^2 \rightarrow \mathbb{R}_+^2$ ,  $(x, y) \mapsto \left(\frac{x+y}{2}, \sqrt{xy}\right)$  ? Écrire un programme qui calcule une valeur approchée de la limite  $\lim_{n \rightarrow \infty} f^n(x, y)$  et admirer la vitesse.

*Remarque.* — La limite  $AGM(x, y) := \lim f^n(x, y)$  est un compromis astucieux, inventé par Gauss, entre la moyenne arithmétique  $\frac{x+y}{2}$  et la moyenne géométrique  $\sqrt{xy}$ . Pour cette raison on appelle la valeur  $AGM(x, y)$  la *moyenne arithmético-géométrique*, ou *arithmetic-geometric mean* en anglais.

*Quelques pistes de réflexion.* — Comment expliquer la convergence super-rapide ? La fonction  $f$  est-elle contractante ? Est-ce que la suite itérative  $f^n(x, y)$  converge pour tout  $(x, y) \in \mathbb{R}_+^2$  ? Dans quel sens est-ce que la fonction  $f$  « contracte vers la diagonale » ? Pour un point  $(x, x)$  de la diagonale calculer la dérivée dans la direction orthogonale  $(1, -1)$ . Dans quel sens la diagonale est-elle « super-attractive » ? (Voir la suite.)



## 2. La méthode de Newton

**2.1. Vitesse de convergence.** Regardons à nouveau la dynamique autour d'un point fixe attractif : soit  $\phi : \mathbb{R} \rightarrow \mathbb{R}$  une fonction continûment dérivable, et soit  $a = \phi(a)$  un point fixe vérifiant  $|\phi'(a)| < 1$ . On a vu que dans un voisinage  $V = [a - \varepsilon, a + \varepsilon]$ , la fonction  $\phi$  est contractante de rapport  $k = \sup_V |\phi'| < 1$ , et pour tout  $x_0 \in V$  on a convergence linéaire :  $|\phi^n(x_0) - a| \leq k^n |x_0 - a|$ . Le cas  $\phi'(a) = 0$  est particulier : quand  $\phi^n(x_0)$  approche  $a$ , la fonction  $\phi$  contracte de plus en plus fortement, ce qui accélère la convergence :

**Remarque 2.1.** Soit  $\phi : \mathbb{R} \rightarrow \mathbb{R}$  de classe  $C^2$  telle que  $\phi(a) = a$  et  $\phi'(a) = 0$ . Soit  $V = [a - \varepsilon, a + \varepsilon]$  et  $M := \max_V |\phi''|$ . D'après le développement de Taylor, pour tout  $x$  il existe  $\xi$  entre  $a$  et  $x$  tel que

$$\phi(x) = \phi(a) + \phi'(a)(x-a) + \frac{1}{2}\phi''(\xi)(x-a)^2.$$

Pour tout  $x_0 \in V$  ceci implique  $|\phi(x_0) - a| \leq \frac{M}{2}|x_0 - a|^2$ . On conclut que pour toute valeur initiale  $x_0 \in V$  vérifiant  $|x_0 - a| < \frac{2}{M}$  on a une convergence particulièrement rapide :

$$\frac{M}{2}|\phi^n(x_0) - a| \leq \left(\frac{M}{2}|x_0 - a|\right)^{2^n}.$$

Ceci veut dire que le nombre de décimales exactes de  $x_n = \phi^n(x_0)$  comme approximation de  $a$  double à chaque itération ! C'est cette convergence super-rapide qui fait de cette observation un outil très puissant : si vous avez calculé une approximation  $x_n$  avec un écart  $\frac{M}{2}|x_n - a| \leq 2^{-1}$  près, disons, l'itération suivante ne laissera qu'un écart  $\leq 2^{-2}$ , celle d'après  $\leq 2^{-4}$ , puis  $\leq 2^{-8}$ , puis  $\leq 2^{-16}$ , puis  $\leq 2^{-32}$  etc.

**Définition 2.2.** Dans un espace métrique  $(E, d)$  soit  $(x_n)_{n \in \mathbb{N}}$  une suite convergente de limite  $a$ . On note  $e_n = d(x_n, a)$  l'écart entre  $x_n$  et la limite cherchée  $a$ , et on suppose  $e_n > 0$  pour tout  $n$ .

- (1) On dit que la convergence est asymptotiquement (au moins) *linéaire* si  $\limsup \frac{e_{n+1}}{e_n} < 1$ . Ceci équivaut à dire qu'il existe  $k < 1$  et  $n_0 \in \mathbb{N}$  tels que  $e_{n+1} \leq ke_n$  pour tout  $n \geq n_0$ . Dans le cas  $E = \mathbb{R}$ , le nombre de décimales exactes est minoré par une fonction affine de  $n$ .
- (2) On dit que la convergence est asymptotiquement (au moins) *quadratique* si  $\limsup \frac{e_{n+1}}{e_n^2} < +\infty$ . Ceci équivaut à dire qu'il existe  $K \in \mathbb{R}_+$  tel que  $e_{n+1} \leq Ke_n^2$  pour tout  $n \in \mathbb{N}$ . Dans le cas  $E = \mathbb{R}$ , le nombre de décimales exactes double à peu près à chaque itération.

**Exemple 2.3.** Le théorème du point fixe promet une convergence linéaire. Nous avons vu la convergence quadratique de la méthode de Newton-Héron au chapitre XV, §3.1.

**Proposition 2.4.** Soit  $U \subset \mathbb{R}^m$  un ouvert et  $\phi : U \rightarrow \mathbb{R}^m$  une application de classe  $C^2$ . Si  $\phi(a) = a$  et  $\phi'(a) = 0$  alors existe un voisinage  $V$  de  $a$  tel que  $\phi|_V$  soit contractante de rapport  $\frac{1}{2}$ . Par conséquent :

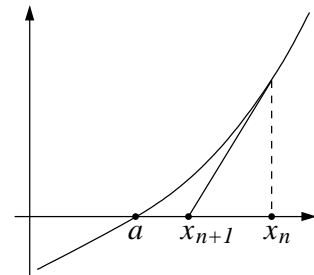
- (1) On a  $\phi(V) \subset V$  et pour tout  $x_0 \in V$  la suite itérative  $x_n = \phi^n(x_0)$  converge vers  $a$ .
- (2) La vitesse de convergence est au moins linéaire,  $|x_n - a| \leq 2^{-n}|x_0 - a|$ .
- (3) Finalement la convergence devient quadratique,  $|x_{n+1} - a| \leq \frac{M}{2}|x_n - a|^2$ .

**Exercice/M 2.5.** Prouver cette proposition dans le cas  $\mathbb{R}^m$  en généralisant les arguments donnés pour  $\mathbb{R}$ . (Cf. Demailly, §IV.2.3.) Ce résultat motive la définition suivante :

**Définition 2.6.** Soit  $\phi : \mathbb{R}^m \supset U \rightarrow \mathbb{R}^m$  de classe  $C^1$ . Un point fixe  $a = \phi(a)$  est *super-attractif* si  $\phi'(a) = 0$ . Un voisinage  $V$  de  $a$  est dit *newtonien* pour  $\phi$  s'il satisfait aux conditions (1), (2), (3) ci-dessus.

**2.2. Itération de Newton.** On cherche à résoudre une équation  $f(x) = 0$  dans  $\mathbb{R}^m$  dont on connaît une solution approchée  $x_n$ . La méthode de Newton permet de transformer l'équation  $f(x) = 0$  en un problème de point fixe  $\phi(x) = x$ . Son intérêt réside dans le fait que les zéros de  $f$  deviennent des points fixes *super-attractifs* pour  $\phi$ , ce qui nous permettra des calculs extrêmement efficaces.

Au lieu d'une approche purement métrique à la méthode du point fixe, on veut tirer profit du calcul différentiel ! On approche donc la fonction  $f$  par la tangente en  $x_n$ , c'est-à-dire  $t(x) := f(x_n) + f'(x_n)(x - x_n)$ . C'est l'approximation de Taylor d'ordre 1.



Pour  $x_{n+1}$  on prendra l'unique solution de l'équation affine  $t(x) = 0$ , à savoir  $x_{n+1} = x_n - f'(x_n)^{-1}f(x_n)$ . Autrement dit, pour approcher les zéros de  $f$  on itère l'application  $\phi(x) = x - f'(x)^{-1}f(x)$ .

**Théorème 2.7** (méthode de Newton, version qualitative). Soit  $U \subset \mathbb{R}^m$  un ouvert et soit  $f: U \rightarrow \mathbb{R}^m$  une fonction de classe  $C^2$  telle que  $f'(x)$  soit inversible pour tout  $x \in U$ . Alors la fonction  $\phi: U \rightarrow \mathbb{R}^m$  définie par  $\phi(x) = x - f'(x)^{-1}f(x)$  est de classe  $C^1$ , les points fixes de  $\phi$  sont exactement les zéros de  $f$ , et en tout point fixe  $a = \phi(a)$  la dérivée s'annule :  $\phi'(a) = 0$ .

Les zéros de  $f$  deviennent ainsi des points fixes super-attractifs de  $\phi$ , comme souhaité, et dans un voisinage convenable on a convergence quadratique :

**Corollaire 2.8.** Soit  $f: \mathbb{R}^m \supset U \rightarrow \mathbb{R}^m$  une fonction de classe  $C^3$  et soit  $a \in U$  une racine simple de  $f$ , c'est-à-dire  $f(a) = 0$  à dérivée  $f'(a)$  inversible. Alors il existe un rayon  $\delta > 0$  de sorte que la boule  $B(a, \delta) = \{x \in \mathbb{R}^m \mid |x - a| \leq \delta\}$  soit contenue dans  $U$  et soit newtonienne pour  $\phi$ .

**Exercice/M 2.9.** Prouver ce théorème et son corollaire (pour  $m = 1$  si vous voulez simplifier). Qu'obtient-on quand on l'applique à  $f: \mathbb{R}_+ \rightarrow \mathbb{R}$ ,  $f(x) = x^n - a$ ? Pour quelles valeurs initiales  $x_0$  l'itération de Newton converge-t-elle? Comparer avec le résultat du chapitre XV, §3.1 : en quoi notre formulation du théorème de Newton-Héron est-elle plus précise ou plus générale que le résultat qualitatif précédent?



*Heuristique : Si  $f(a) = 0$  et l'on dispose d'une valeur initiale  $x_0 \approx a$  suffisamment proche de  $a$ , alors l'itération de Newton  $x_n = \phi^n(x_0)$  converge très rapidement vers  $a$ .*



**2.3. Exemples.** En vue d'une implémentation sur ordinateur nous sous-entendons ici qu'il est raisonnablement facile d'évaluer  $f$  et  $f'$  en  $x_n$ , c'est-à-dire on sait implémenter des approximations de  $f(x_n)$  et de  $f'(x_n)$  avec une précision et une efficacité satisfaisantes. C'est le cas pour tous nos exemples, mais pour un problème réaliste cette étape peut en elle-même nécessiter une analyse approfondie. Nous supposons ce problème résolu par une bibliothèque convenable.

**Exemple 2.10.** On reprend l'équation  $\cos(x) = x$  qui a été résolue par le programme `iter1.cc` en utilisant la méthode du point fixe. Testez le programme `iter2.cc` qui résout cette équation par la méthode de Newton, et admirez la vitesse de convergence.

**Exercice/P 2.11.** Si vous voulez, vous pouvez reprendre l'équation  $\exp(x) = x^3$ , résolue plus haut par la méthode du point fixe, et approcher ses deux solutions par la méthode de Newton.

**Remarque 2.12.** Avant d'appliquer la méthode de Newton il faut bien vérifier que l'on est en présence d'un zéro simple, c'est-à-dire  $f(a) = 0$  mais  $f'(a)$  inversible. Dans ce cas il existe un voisinage  $U$  tel que  $f'(x)$  reste inversible pour tout  $x \in U$ , et on peut appliquer le théorème précédent.

**2.4. Bassin d'attraction.** La convergence de la méthode de Newton n'est assurée que pour une valeur initiale proche d'un zéro. Nos résultats précédents motivent la définition suivante :

**Définition 2.13** (Rayon de Newton). Soit  $f: \mathbb{R}^m \supset U \rightarrow \mathbb{R}^m$  une fonction de classe  $C^2$  et soit  $a \in U$  une racine simple de  $f$ , c'est-à-dire  $f(a) = 0$  à dérivée  $f'(a)$  inversible. Dans ce cas on appelle

$$\rho_a := \sup\{\delta \in \mathbb{R}_+ \mid \text{la boule } B(a, \delta) \subset U \text{ est newtonienne pour } \phi\}$$

le rayon de Newton de la racine  $a$ .

**Remarque 2.14.** Le corollaire 2.8 assure que  $\rho_a > 0$  (pourvu que  $f$  est de classe  $C^3$ , mais  $C^2$  suffit d'après le théorème 2.20 plus bas). Une fois on dispose d'une valeur approchée  $x_0$  de la racine  $a$  à une marge  $\leq \rho_a$  près, on peut approcher  $a$  aussi précisément que l'on le souhaite. Il suffit d'itérer la méthode de Newton : la convergence est assurée et on peut même garantir une bonne vitesse de convergence !

**Remarque 2.15.** Comment majorer la marge d'erreur entre l'approximation calculée  $x_n$  (mais erronée) et la valeur exacte cherchée  $a$  (mais inconnue)? La formulation ci-dessus fait semblant de connaître  $a$ , mais dans la pratique on ignore la valeur exacte et on ne dispose que de la valeur approchée. Une fois on est sûr d'être dans un voisinage newtonien, le théorème du point fixe nous dit que  $|x_n - a| \leq |x_n - x_{n-1}|$ .

**Remarque 2.16.** Dans un cours d'analyse on se contente souvent de la formulation *qualitative* du théorème 2.7 donnée ci-dessus. Quant à l'application *pratique* avouons qu'elle laisse encore à désirer :

- (1) Comment déterminer (ou au moins minorer) le rayon de Newton  $\rho_a$  d'une racine  $a$  ? La formulation qualitative ne donne pas de réponse explicite, elle assure seulement que  $\rho_a > 0$ . Une version quantitative sera donnée par le théorème 2.20 plus bas.
- (2) Comment savoir pour une valeur initiale  $x_0$  si l'on est dans le rayon de Newton d'une racine  $a$  ? Normalement on ne connaît pas les racines, donc on veut un critère en fonction de  $x_0$  seulement. Un tel critère sera donné par le théorème 2.25.
- (3) Comment trouver une approximation initiale  $x_0$  telle que l'on puisse garantir la convergence rapide  $x_n = \phi^n(x_0) \rightarrow a$  ? Certes, il suffit de localiser la racine  $a$  à  $\rho_a$  près, mais comment faire ? En général, partant juste de la fonction  $f$  sans aucune connaissance au préalable de ses racines, c'est un problème délicat qu'il faut analyser au cas par cas.

*L'algorithme de Newton est une méthode locale et non globale : il faut de bonnes valeurs initiales pour assurer sa convergence et son efficacité.*

**Exercice 2.17.** À titre d'exemple, analyser  $f: \mathbb{R} \rightarrow \mathbb{R}$ ,  $f(x) = \arctan(x)$  afin d'illustrer la convergence ou divergence de la méthode de Newton. L'unique zéro est  $a = 0$ . Tracer le graphe de  $f$  et déterminer pour quelles valeurs initiales  $x_0$  l'itération de Newton  $\phi^n(x_0)$  converge, puis pour lesquelles elle diverge. Que se passe-t-il pour les cas critiques intermédiaires ? (L'équation  $\arctan(x) = 2x/(1+x^2)$  peut elle-même être résolue par la méthode de Newton.) Construire ainsi une suite itérative de Newton qui converge, mais dont la convergence est initialement aussi lente que l'on veut. Déterminer finalement le rayon de Newton  $\rho_a$ .

**Exercice 2.18.** Pour les fonctions unidimensionnelles  $f: \mathbb{R} \rightarrow \mathbb{R}$  il est très commode de donner des critères « géométriques » de convergence pour de l'itération Newton  $x_n = \phi^n(x_0)$  :

- (1) Supposons que  $f(a) = 0$  ainsi que  $f' > 0$  et  $f'' \geq 0$  sur  $I = [a, a + \varepsilon]$ . Alors  $\phi(I) \subset I$  et pour toute valeur initiale  $x_0 \in I$  on obtient une suite décroissante  $x_0 \geq x_1 \geq x_2 \geq \dots$  qui converge vers  $a$ .
- (2) Supposons que  $f(a) = 0$  ainsi que  $f' > 0$  et  $f'' \leq 0$  sur  $I = [a - \varepsilon, a]$ . Alors  $\phi(I) \subset I$  et pour toute valeur initiale  $x_0 \in I$  on obtient une suite croissante  $x_0 \leq x_1 \leq x_2 \leq \dots$  qui converge vers  $a$ .

Dans les deux cas on obtient la majoration d'erreur

$$|x_n - a| \leq \left| \frac{f(x_n)}{f'(a)} \right|.$$

En particulier, pour  $f'(a) > 0$  et  $f''(a) > 0$  il suffit de commencer l'itération légèrement à droite de  $a$  afin de garantir  $x_n \rightarrow a$ . Pour  $f'(a) > 0$  et  $f''(a) < 0$  il suffit de commencer légèrement à gauche de  $a$ . La taille exacte de l'intervalle peut être déterminé en étudiant  $f''$ . Que faire dans le cas  $f'(a) < 0$  ? Dessin !

**Exercice 2.19.** On considère la fonction polynomiale

$$f: \mathbb{R} \rightarrow \mathbb{R}, \quad f(x) = 3x^3 - 10x^2 - 4x + 8.$$

- (1) Montrer que  $f$  admet trois racines réelles distinctes, que l'on notera  $a, b, c \in \mathbb{R}$  avec  $a < b < c$ . Encadrer ces racines par trois intervalles de la forme  $]k, k + 1[$  avec  $k \in \mathbb{Z}$ . Dans la suite on veut appliquer la méthode de Newton pour approcher les racines  $a, b, c$  avec plus de précision.
- (2) Sur quels intervalles la fonction  $f$  est-elle croissante/décroissante ? Sur quels intervalles est-elle concave/convexe ? En déduire une valeur initiale  $c_0 \in \mathbb{Z}$  pour laquelle vous pouvez garantir que la méthode de Newton donne une suite récurrente  $c_n = \phi^n(c_0)$  qui converge vers la racine  $c$ . (Justifiez votre choix.) Même question pour  $a_0$  tel que  $a_n \rightarrow a$ , puis  $b_0$  tel que  $b_n \rightarrow b$ .
- (3) Calculer les valeurs approchées  $a_3, b_3, c_3$  des trois racines  $a, b, c$  par la méthode de Newton, comme préparée ci-dessus, en effectuant 3 itérations pour chacune. Majorer l'écart  $|a_3 - a|, |b_3 - b|, |c_3 - c|$ . Vérifier les approximations en développant  $f_3(x) = 3(x - a_3)(x - b_3)(x - c_3)$ . Obtiennent-ils  $f(x)$  exactement ? Que se passe-t-il pour cinq puis dix itérations ?
- (4) Exhiber une valeur initiale  $u_0$  pour laquelle la méthode de Newton donne une suite  $(u_n)_{n \in \mathbb{N}}$  qui ne converge pas. Pourquoi ceci ne met pas en cause le théorème 2.7 ?



**2.5. Version quantitative.** Le théorème 2.7 reste seulement qualitatif dans le sens qu'il n'explicite pas de voisinage sur lequel il garantit une convergence. Les exercices précédents montrent des critères géométriques pour les fonctions convexes/concaves, ce qui fait intervenir  $f''$ . Le théorème suivant donne une version quantitative qui explicite un intervalle sur lequel on peut garantir la convergence :

**Théorème 2.20** (méthode de Newton, version quantitative). *Soit  $f: \mathbb{R} \rightarrow \mathbb{R}$  de classe  $C^2$ . Supposons que  $f(a) = 0$  ainsi que  $|f'| \geq m > 0$  et  $|f''| \leq M$  sur un intervalle  $I = [a - \varepsilon, a + \varepsilon]$  avec  $\varepsilon > 0$ . Nous posons  $\delta := \min(\varepsilon, m/M)$  et  $V := [a - \delta, a + \delta]$ . Alors pour toute valeur initiale  $x_0 \in V$  la suite  $x_n = \phi^n(x_0)$  vérifie*

$$|x_n - a| \leq \left(\frac{1}{2}\right)^{2^n-1} |x_0 - a|.$$

En particulier  $V$  est un voisinage newtonien de  $a$  dans le sens de la définition 2.6.

DÉMONSTRATION. Par translation nous pouvons supposer que  $a = 0$ . Nous avons

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

D'autre part le développement de Taylor en  $x_n$  nous donne

$$0 = f(0) = f(x_n) - f'(x_n)x_n + \frac{1}{2}f''(\xi_n)x_n^2$$

pour un  $\xi_n \in [0, x_n]$  convenable. De ces deux équations on déduit

$$x_{n+1} = \frac{1}{2} f'(x_n)^{-1} f''(\xi_n) x_n^2.$$

Avec  $|f'| \geq m > 0$  et  $|f''| \leq M$  ceci donne

$$|x_{n+1}| \leq \frac{M}{2m} |x_n|^2$$

ou encore  $\frac{M}{2m} |x_{n+1}| \leq \left(\frac{M}{2m} |x_n|\right)^2$ . Par récurrence on obtient

$$\frac{M}{2m} |x_n| \leq \left(\frac{M}{2m} |x_0|\right)^{2^n}.$$

Pour  $x_0 \in V$  on a  $|x_0| \leq \delta \leq \frac{m}{M}$ , donc  $\frac{M}{2m} |x_0| \leq \frac{1}{2}$ . On conclut que  $|x_n| \leq \left(\frac{1}{2}\right)^{2^n-1} |x_0|$ .  $\square$

**Exercice/M 2.21.** Vérifier soigneusement le théorème précédent, puis essayer de formuler et de prouver une version pour les fonctions complexes  $f: \mathbb{C} \rightarrow \mathbb{C}$ , ou plus généralement pour  $f: \mathbb{R}^n \rightarrow \mathbb{R}^n$ .

*Exercice/M 2.22.* La méthode de Newton est assez universelle et sa convergence foudroyante en fait un outil omniprésent. En voici un exemple plus poussé, d'après Demailly, §IV, qui sert pour l'inversion des matrices :

Soit  $A$  une algèbre unitaire normée sur  $\mathbb{R}$ , par exemple les matrices carrées  $m \times m$ .

- Soit  $u \in A$  un élément inversible. Déterminer deux constantes  $\alpha, \beta \in \mathbb{R}$  de sorte que l'application  $\phi: A \rightarrow A$ ,  $\phi(x) = \alpha x + \beta xux$ , admette  $u^{-1}$  comme point fixe super-attractif. Montrer que  $|\phi'(x)| \leq 2|u| \cdot |x - u^{-1}|$ . En déduire que la suite  $x_n = \phi^n(x_0)$  converge vers  $u^{-1}$  pour tout  $x_0 \in B(u^{-1}, r)$  avec  $r < \frac{1}{2|u|}$ .
- On suppose désormais que  $A$  est complète. (C'est automatique si  $A$  est de dimension finie.) En supposant  $|v| < 1$  montrer que  $u = 1 - v$  est inversible et  $u^{-1} = \sum_{k=0}^{\infty} v^k$ . En quoi la méthode (a) est-elle plus efficace ? Déterminer un entier  $n \in \mathbb{N}$  de sorte que la méthode (a) converge pour  $x_0 = 1 + v + \dots + v^n$ .

Pour résumer : quels sont les avantages et inconvénients des deux méthodes (a) et (b) utilisées séparément ? Quel est l'intérêt de les combiner ?

**2.6. Critères pratiques.** La problématique de choisir de bonnes valeurs initiales pour la méthode de Newton reste d'actualité. On peut se poser une question plus facile : étant donné  $z_0$ , comment savoir si l'itération de Newton  $z_n = \phi^n(z_0)$  convergera ? Le théorème 2.20 part de la racine  $a$  et minore son rayon de Newton. Ceci n'est souvent pas pratique car on ignore les racines encore à trouver. Le critère suivant, par contre, part d'une valeur initiale  $z_0$  donnée :

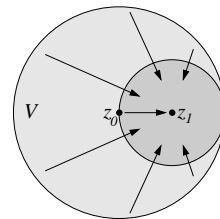
**Théorème 2.23.** *Considérons une fonction analytique  $f: \mathbb{C} \supset U \rightarrow \mathbb{C}$  avec son application de Newton  $\phi(z) = z - \frac{f(z)}{f'(z)}$ . Soit  $z_0 \in U$  une valeur initiale telle que  $f(z_0) \neq 0$  et  $f'(z_0) \neq 0$ , et soit  $r := \frac{|f(z_0)|}{|f'(z_0)|}$  le pas initial dans l'itération de Newton. Supposons que la boule  $B(z_0, 2r)$  soit incluse dans  $U$ , et que*

$$\frac{|f''(z)|}{|f'(z)|} \leq \frac{1}{8} \frac{|f'(z_0)|}{|f(z_0)|} \quad \text{pour tout } z \in B(z_0, 2r) =: V.$$

Alors  $\phi|_V$  est contractante de rapport  $\frac{1}{2}$  et vérifie  $\phi(V) \subset V$ . Par conséquent  $f|_V$  admet une unique racine  $a \in V$ , et la suite itérative  $z_n = \phi^n(z_0)$  converge vers  $a$ . La convergence est quadratique comme expliqué dans la proposition 2.4.

*Remarque.* — Par hypothèse on a  $|\phi(z_0) - z_0| = r$ . La première itération reste donc bien dans  $V = B(z_0, 2r)$ , mais il faut encore contrôler les suivantes. L'idée est de montrer que  $\phi|_V$  est contractante de rapport  $\frac{1}{2}$ , c'est-à-dire  $|\phi(x) - \phi(y)| \leq \frac{1}{2}|x - y|$  pour tout  $x, y \in V$ . Ceci entraîne  $\phi(V) \subset V$ , car pour tout  $z$  avec  $|z - z_0| \leq 2r$  on trouve  $|\phi(z) - \phi(z_0)| \leq r$  et donc

$$|\phi(z) - z_0| \leq |\phi(z) - \phi(z_0)| + |\phi(z_0) - z_0| \leq 2r.$$



**DÉMONSTRATION.** Après translation on peut supposer que  $z_0 = 0$ . Par hypothèse  $f'$  ne s'annule pas dans  $V = B(0, 2r)$ , donc  $\phi$  est définie sur  $V$ . Pour montrer que  $\phi|_V$  est contractante de rapport  $\frac{1}{2}$  on montrera que sa dérivée  $\phi'(z) = f(z)f''(z)/f'(z)^2$  vérifie  $|\phi'| \leq \frac{1}{2}$  sur  $V$ . Étudions le quotient  $u(z) = f(z)/f'(z)$  et sa dérivée  $u'(z) = 1 - u(z)f''(z)/f'(z)$ . On a  $|u(0)| = r$  et  $|u'(z)| \leq 1 + M|u(z)|$  avec  $M := \sup_V \frac{|f''|}{|f'|}$ . Pour tout  $R > r$  la fonction auxiliaire  $v: \mathbb{R} \rightarrow \mathbb{R}$  définie par  $v(t) = (R + M^{-1})\exp(Mt) - M^{-1}$  est croissante et vérifie  $v(0) > |u(0)|$  et  $v'(t) = 1 + Mv(t)$ . Ceci entraîne que  $|u(z)| < v(|z|)$  pour tout  $z \in B(0, 2r)$ . Sinon il existerait  $z$  tel que  $|u(z)| = v(|z|)$  avec  $|z|$  minimal, et on arriverait à la contradiction suivante :

$$\begin{aligned} |u(z)| - |u(0)| &\leq |u(z) - u(0)| = \left| \int_0^z u'(\xi) d\xi \right| \leq \int_0^{|z|} |u'(\xi)| d\xi \leq \int_0^{|z|} 1 + M|u(\xi)| d\xi \\ &< \int_0^{|z|} 1 + Mv(t) dt = \int_0^{|z|} v'(t) dt = v(|z|) - v(0) = |u(z)| - v(0) < |u(z)| - |u(0)|. \end{aligned}$$

On en déduit que  $|\phi'(z)| \leq M|u(z)| < Mv(2r)$  pour tout  $z \in B(0, 2r)$ . Puisque c'est valable pour tout  $R > r$  on obtient  $|\phi'(z)| \leq Mv(2r)$  dans le cas limite  $R = r$ . Il ne reste qu'à déterminer  $M$  de sorte que  $Mv(2r) = (Mr + 1)\exp(2Mr) - 1 \leq \frac{1}{2}$ . Le choix  $Mr = \frac{1}{8}$  convient, ce qui prouve le théorème.  $\square$

**Exercice 2.24.** Vérifier la preuve précédente, en particulier la majoration  $|u(z)| < v(|z|)$ . Montrer que l'équation  $(x + 1)\exp(2x) = 3/2$  admet une unique solution  $x_0$  et vérifier que  $x_0 \approx 0.1381 > \frac{1}{8} = 0.125$ .

Le théorème précédent est un premier pas vers un critère pratique, mais il nous demande encore d'étudier  $f$  dans une boule  $B(z_0, 2r)$  afin de majorer la proportion  $|f''(z)^{-1}f''(z)|$ . Le théorème suivant, développé par S. Smale, remplace cette étude par des informations locales en  $z_0$  seulement, à savoir les dérivées  $f^{(k)}(z_0)$  pour  $k \in \mathbb{N}$ . Ces informations suffisent à contrôler le comportement de  $f$  si  $f$  est un polynôme, ou plus généralement une fonction analytique :

**Théorème 2.25** (S. Smale, 1986). *Soit  $f: \mathbb{C} \rightarrow \mathbb{C}$  une fonction analytique. Pour que l'itération de la fonction  $\phi(z) = z - f'(z)^{-1}f(z)$  converge pour une valeur initiale  $z_0 \in \mathbb{C}$  il suffit que*

$$\sup_{k \geq 2} \sqrt[k-1]{\frac{|f^{(k)}(z_0)|}{k!|f'(z_0)|}} \leq \frac{1}{8} \frac{|f'(z_0)|}{|f(z_0)|}.$$

Dans ce cas  $f$  admet une unique racine  $a$  dans la boule  $B(z_0, 2r)$  où  $r := |f(z_0)/f'(z_0)|$  est le pas initial dans l'itération de Newton, et la suite  $z_n = \phi^n(z_0)$  vérifie  $|z_n - a| \leq (\frac{1}{2})^{2^n - 1} \cdot |z_0 - a|$ .  $\square$

A noter que le théorème de Smale s'appuie uniquement sur des estimations au point  $z_0$  donné, et que ce critère s'applique très facilement aux polynômes  $f$  : il suffit de vérifier la condition pour  $k = 2, \dots, n$ .

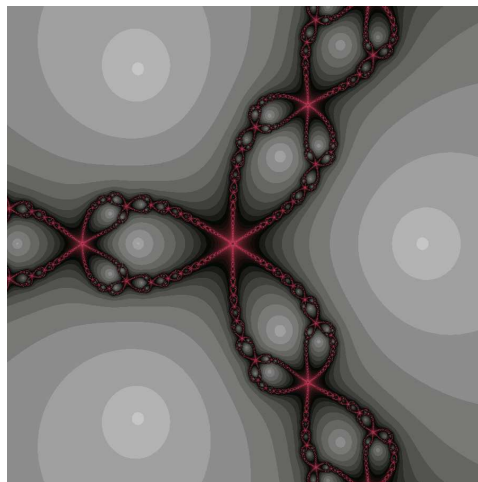
Pour une discussion détaillée voir le chapitre 8 intitulé "Newton's method" du livre de L. Blum, F. Cucker, M. Shub, S. Smale : *Complexity and real computation*, Springer, New York 1997.

### 3. Application aux polynômes complexes

En 1685 Wallis publia son livre nommé *Algèbre* dans lequel il décrit une méthode inventée par Newton pour la résolution d'équations. Une version modifiée fut publiée par Raphson en 1690 ; c'est la méthode qui est maintenant connue sous le nom de Newton ou de Newton-Raphson. Newton lui-même avait discuté sa méthode en 1669 et l'avait illustrée par l'exemple  $X^3 - 2X - 5 = 0$ . Continuons cette tradition vieille de plus de 300 ans qui veut que tout étudiant de méthodes numériques résolve cette vénérable équation :

**Exercice/P 3.1.** Appliquez la méthode de Newton au polynôme  $X^3 - 2X - 5$ . Expérimentez avec le programme `newton1669.cc` pour trouver les trois racines de ce polynôme. Comme de nos jours ces calculs ne coûtent pas cher, vous pouvez facilement faire varier la valeur initiale  $z_0$ . Voyez-vous une corrélation entre le choix de  $z_0$  et le résultat obtenu ? et le nombre d'itérations ? Tester plusieurs cas où  $z_0$  est « proche » d'une racine, puis des cas où  $z_0$  en est « loin ». Après plusieurs expériences réussies, regarder par exemple  $z_0 = -60, -61, -62, \dots, -70$  afin de tester délibérément les limites de la méthode.

Étant donné un polynôme  $P \in \mathbb{C}[X]$  on peut appliquer l'itération de Newton  $\phi(z) = z - P'(z)^{-1}P(z)$  à une valeur initiale  $z_0 \in \mathbb{C}$  choisie au hasard, en espérant que  $\phi^n(z_0)$  convergera vers une racine de  $P$ . De manière empirique, on constate que « la plupart » des choix initiaux mènent à une racine. Une fois on sait localiser une racine, la méthode de Newton permet de calculer de très bonnes approximations au bout de quelques itérations seulement. Sans connaissance préalable des racines, par contre, il est difficile de deviner de bonnes valeurs initiales. Pour une valeur initiale loin des racines il semble impossible de prédire vers quelle racine convergera l'itération. L'inconvénient pratique est surtout que la convergence peut être très lente. Dans le pire des cas, certains choix n'aboutiront jamais : pour  $X^2 + 1$  une valeur initiale  $z_0 \in \mathbb{R}$  ne permet pas de trouver une des deux racines. (Pourquoi ?)



On peut s'amuser avec des graphiques suivantes : Le polynôme  $P(z) = z^3 - 1$  possède trois racines complexes :  $1, j$  et  $j^2$ . On souhaite visualiser la dynamique de l'application  $\phi(z) = z - P'(z)^{-1}P(z)$ . Pour  $a \in \{1, j, j^2\}$  on veut représenter graphiquement l'ensemble  $A(a) = \{z_0 \in \mathbb{C} \mid \phi^n(z_0) \rightarrow a\}$ , dit « bassin d'attraction ». On peut écrire un programme représentant les ensembles  $A(a)$  par trois couleurs distinctes et utilisant une quatrième couleur pour l'ensemble des  $z_0 \in \mathbb{C}$  tels que la suite  $(z_n)$  « ne semble pas » converger. Il existe des logiciels spécialisés pour ces graphiques dites « fractales », comme `fractint`.

**Exemple 3.2.** À titre d'avertissement, considérons le polynôme  $p(z) = z^3 - 2z + 2$ . La méthode de Newton itère  $\phi(z) = z - \frac{z^3 - 2z + 2}{3z^2 - 2} = \frac{2z^3 - 2}{3z^2 - 2}$ , dont la dérivée est  $\phi'(z) = \frac{6z^4 - 12z^2 + 12z}{9z^4 - 12z^2 + 4}$ . Comme  $\phi(0) = 1$  et  $\phi(1) = 0$ , on a un cycle de longueur 2. Ce cycle est super-attractif car  $\phi'(0) = 0$  : tous les points dans un voisinage de  $\{0, 1\}$  sont attirés vers le cycle  $0 \rightarrow 1 \rightarrow 0 \rightarrow 1 \rightarrow \dots$ , et ne convergerons pas vers une racine de  $p$ . Ce phénomène montre que l'on ne peut pas espérer la convergence pour « presque tout » point initial.

**3.1. Le théorème de Gauss-d'Alembert.** Soit  $K$  un corps et  $P \in K[X]$  un polynôme de degré  $n$ . Pour  $a \in K$  la division euclidienne donne  $P(X) = (X - a)Q(X) + r$  avec  $Q \in K[X]$  et  $r \in K$ . Ceci veut dire que  $a$  est une racine de  $P$  si et seulement si le facteur linéaire  $(X - a)$  divise le polynôme  $P$ .

Soit  $a_1 \in K$  une racine de  $P$ . Par récurrence on trouve  $m_1 \geq 1$  tel que  $P(X) = (X - a_1)^{m_1}Q(X)$  et  $Q(a_1) \neq 0$ . On appelle  $m_1$  la *multiplicité* de la racine  $a_1$ . Si  $a_1, \dots, a_k$  sont des racines distinctes de  $P$  dans  $K$  alors  $P = (X - a_1)^{m_1} \dots (X - a_k)^{m_k}Q(X)$  avec des multiplicités  $m_1, \dots, m_k \geq 1$  et le polynôme restant  $Q \in K[X]$  ne s'annule pas en  $a_1, \dots, a_k$ . Puisque  $\deg(P) = m_1 + \dots + m_k + \deg(Q)$ , on conclut que  $m_1 + \dots + m_k \leq n$ , et ce procédé doit s'arrêter quand  $Q$  n'a plus aucune racine dans  $K$  :

**Proposition 3.3.** Sur un corps  $K$  tout polynôme  $P \in K[X]$  de degré  $n$  admet au plus  $n$  racines. □

Si  $P$  admet  $n$  racines  $a_1, a_2, \dots, a_n \in K$  (éventuellement avec répétitions) alors il se factorise comme

$$P = c \cdot (X - a_1)(X - a_2) \cdots (X - a_n).$$

Une telle factorisation en polynômes de degré 1 n'est pas toujours possible : le polynôme  $X^2 + 1 \in \mathbb{R}[X]$  par exemple n'admet aucune racine dans  $\mathbb{R}$ . Sur  $\mathbb{C}$  la situation est beaucoup meilleure :

**Théorème 3.4** (de Gauss-d'Alembert). *Pour tout polynôme  $P \in \mathbb{C}[X]$  unitaire de degré  $n$  il existe des nombres complexes  $a_1, a_2, \dots, a_n \in \mathbb{C}$  tels que  $P = (X - a_1)(X - a_2) \cdots (X - a_n)$ .*

On dit aussi que sur  $\mathbb{C}$  tout polynôme est *scindé*, ou qu'il se *factorise* en facteurs linéaires sur  $\mathbb{C}$ . Cette propriété est tellement importante en algèbre qu'on lui donne un nom : on dit que le corps  $\mathbb{C}$  est *algébriquement clos*. Ce superbe théorème est parfois appelé le « théorème fondamental de l'algèbre » (bien que le corps  $\mathbb{C}$  appartienne plutôt à l'analyse). Remarquons toutefois qu'il s'agit d'un pur résultat d'existence. Il ne nous indique nullement comment *trouver* les racines. Reste le problème pratique évident : étant donné  $P$ , comment « calculer » ses racines ? Au moins deux approches se présentent :

*L'approche algébrique.* — En degré 2 on peut calculer les racines de  $P = X^2 + aX + b$  par la formule  $\frac{1}{2}(-a \pm \sqrt{a^2 - 4b})$ . De formules similaires existent en degré 3 (formule de Cardan) et degré 4 (formule de Ferrari). Elles sont malheureusement assez complexes, et par conséquent rarement utilisées.

Jusqu'au XIXe siècle des généralisations aux degrés  $\geq 5$  furent cherchées en vain. C'est un des succès spectaculaires de la théorie de Galois d'en montrer l'impossibilité :

**Théorème 3.5** (N.H. Abel, 1824). *Il n'existe pas de formule construite à partir des opérations arithmétiques du corps  $\mathbb{C}$  et des racines nièmes qui résolve une équation polynomiale générale de degré  $\geq 5$ .*  $\square$

Sont résolubles par radicaux seulement certaines équations particulières, comme  $X^n - a = 0$  pour un exemple évident. L'équation  $X^5 - X + 1 = 0$  par contre n'est pas résoluble par radicaux. La théorie de Galois établit le critère général : une équation est résoluble par radicaux si et seulement si le groupe de Galois associé est résoluble.

*L'approche numérique.* — Étant donné un polynôme  $P \in \mathbb{C}[X]$  il ne reste souvent que le recours aux méthodes numériques pour localiser les racines. Évidemment on ne peut en général pas espérer de tomber sur les racines exactes, mais de bonnes valeurs approchées donnent déjà des informations précieuses. Le projet mettra en œuvre une telle démarche.

**3.2. Relation entre racines et coefficients.** Intuitivement, les racines  $a_1, \dots, a_n \in \mathbb{C}$  d'un polynôme  $p(z) = z^n + p_1 z^{n-1} + \dots + p_{n-1} z + p_n$  varient continûment en fonction des coefficients  $p_1, \dots, p_n \in \mathbb{C}$ . C'est vrai mais pas si facile à montrer. La proposition suivante donne une preuve dans le cas « générique » où toutes les racines sont simples ; la situation est plus lisse et le résultat est plus fort :

**Proposition 3.6.** *On considère l'application  $\Phi: \mathbb{C}^n \rightarrow \mathbb{C}^n$  qui associe à  $n$  racines  $(a_1, \dots, a_n) \in \mathbb{C}^n$  les coefficients  $(s_1, \dots, s_n) \in \mathbb{C}^n$  du polynôme*

$$p(z) = \prod_{k=1}^n (z - a_k) = \sum_{k=0}^n (-1)^k s_k z^{n-k}.$$

On a  $s_0 = 1$  et les fonctions  $s_1, \dots, s_n$  sont les fonctions symétriques élémentaires :

$$s_k = \sum_{i_1 < i_2 < \dots < i_k} a_{i_1} a_{i_2} \cdots a_{i_k}.$$

L'application  $\Phi$  est polynomiale, et  $\Phi' = \left( \frac{\partial s_i}{\partial a_j} \right)$  a pour déterminant  $\det \Phi' = \prod_{i < j} (a_i - a_j)$ . Par conséquent la dérivée  $\Phi'(a_1, \dots, a_n)$  est inversible si et seulement si  $a_1, \dots, a_n$  sont distincts deux à deux. Le théorème d'inversion locale implique que dans un voisinage d'un polynôme séparable  $p$  (avec discriminant non nul) les racines dépendent de manière  $C^\infty$  (même analytiquement) des coefficients.  $\square$

*Remarque 3.7.* Le carré  $(\det \Phi')^2 = \prod_{i < j} (a_i - a_j)^2$  est appelé le *discriminant* du polynôme  $p$ . C'est un polynôme symétrique dans les  $a_i$  et peut donc être exprimé comme un polynôme dans les coefficients  $s_j$ . Il existe des méthodes efficaces, similaires à l'algorithme d'Euclide, qui permettent de calculer le discriminant d'un polynôme donné.

*Exercice/M 3.8.* Si vous vous intéressez au calcul différentiel et/ou l'algèbre, vous êtes vivement invité à effectuer les calculs énoncés ci-dessus et de profiter des outils utilisés dans votre cours de calcul différentiel ou d'algèbre.

**3.3. Instabilité des racines mal conditionnées.** Même si les racines dépendent continûment des coefficients, on constate malheureusement un problème d'amplification d'erreur : des petites perturbations dans les coefficients peuvent entraîner des grandes perturbations dans les racines. Dans la pratique ceci se traduit à nouveau en un phénomène d'instabilité numérique.

**Exemple 3.9.** Les racines de  $P(X) = X^n - 2^{-bn}$  sont faciles à expliciter :  $a_k = 2^{-b} \exp(k \cdot 2\pi i/n)$  avec  $k = 1, \dots, n$ . Si l'on considère le polynôme voisin  $Q(X) = X^n$  on trouve la racine 0 de multiplicité  $n$ . Un changement d'ordre  $2^{-bn}$  dans les coefficients entraîne ici un changement d'ordre  $2^{-b}$  dans les racines. (Pensez par exemple à  $b = 50$  et  $n = 10$ .) Pour la pratique on retient la règle suivante :

*Afin de calculer les racines d'un polynôme de degré  $n$  à une précision de  $b$  bits, il faut en général connaître les coefficients de  $p$  à une précision de  $nb$  bits.*

**Exercice 3.10.** Dans des implémentations il est important de vérifier la qualité des racines approchées  $\hat{a}_1, \dots, \hat{a}_n$  en comparant  $\hat{P} = (X - \hat{a}_1) \cdots (X - \hat{a}_n)$  au polynôme  $P$  donné au départ. Nous sommes contents de notre factorisation si  $P$  et  $\hat{P}$  sont proches. Est-ce un critère valable ? Discuter cette vérification.

Précisons nos observations de manière qualitative : Soit  $P \in \mathbb{C}[X]$  un polynôme unitaire de degré  $n$  à racines distinctes  $a_1, \dots, a_n$ . En particulier  $P'(a_k) \neq 0$  en toute racine  $a_k$ . On regarde le polynôme perturbé  $P_t = P + tQ$ , où  $t \in \mathbb{R}$  et  $Q$  est un polynôme de degré  $< n$ .

Supposons que  $a$  est une racine de  $P$ , c'est-à-dire  $P(a) = 0$ , et que  $a_t = a + \delta(t)$  est la racine correspondante du polynôme perturbé  $P_t$ , donc  $0 = P_t(a_t) = P(a_t) + tQ(a_t)$ . La dérivation par rapport à  $t$  donne

$$0 = \left[ \frac{d}{dt} P_t(a_t) \right]_{t=0} = P'(a)\delta'(0) + Q(a) \quad \text{donc} \quad \delta'(0) = -\frac{Q(a)}{P'(a)}.$$

Pour  $Q = X^k$  on obtient ainsi  $\delta(0) = -a^k/P'(a)$  ; c'est le taux de changement de la racine  $a$  quand on perturbe le  $k$ ème coefficient du polynôme  $P$ .

- Pour une racine  $a$  avec  $|a| > 1$  le maximum est atteint pour  $k = n - 1$  ; ce sont donc les plus hauts coefficients qui influencent le plus les grandes racines.
- Pour une racine  $a$  avec  $|a| < 1$  le maximum est atteint pour  $k = 0$  ; ce sont donc les plus bas coefficients qui influencent le plus les petites racines.

*La sensibilité d'une racine  $a$  aux perturbations des coefficients est amplifiée par le facteur  $P'(a)^{-1}$ . Plus la dérivée  $P'(a)$  est proche de zéro, plus la racine  $a$  est instable.*

Le pire des cas arrive si  $m$  racines  $a_1, \dots, a_m$  sont très proches, car dans ce cas  $P'(a_k) \approx 0$ . Dans la limite on peut supposer qu'elles coïncident en une seule racine  $a$  de multiplicité  $m$ , et dans ce cas  $P'(a) = 0$ . L'instabilité d'une telle situation a déjà été illustrée dans l'exemple 3.9.

**Exercice 3.11.** Vérifiez ces constats de manière empirique sur des exemples de votre choix. Pour les calculs vous pouvez utiliser un logiciel de votre choix, ou bien votre propre implémentation issue du projet suivant. Est-ce que des petits changements dans les coefficients peuvent entraîner des grands changements des racines ? Qu'est-ce que cela veut dire pour notre vérification  $\text{dist}(\hat{P}, P) < \varepsilon$  ?

**Exercice/M 3.12.** Si le polynôme  $P$  risque d'avoir de racines multiples, on calcule d'abord  $Q = \text{pgcd}(P, P')$  : les racines de  $Q$  sont exactement les racines multiples de  $P$ , et  $P/Q$  n'admet que de racines simples. Cette astuce peut considérablement faciliter le calcul des racines. Détailler cette approche algorithmiquement pour  $P \in \mathbb{K}[X]$  avec  $\mathbb{K} = \mathbb{Q}$  ou  $\mathbb{K} = \mathbb{Q}[i]$ . Quels problèmes apparaissent pour  $\mathbb{K} = \mathbb{R}$  et  $\mathbb{K} = \mathbb{C}$  ?





*Vous avez trouvé par de long ennuis  
ce que Newton trouva sans sortir de chez lui.  
Voltaire (1694-1778)*

## PROJET XVII

# Factorisation de polynômes complexes

### 1. Approche probabiliste

Nous nous proposons de développer un programme pour factoriser des polynômes complexes. Soit  $P \in \mathbb{C}[X]$  un polynôme unitaire de degré  $n$ , donné sous la forme  $P(X) = X^n + p_1X^{n-1} + \dots + p_{n-1}X + p_n$  par ses coefficients  $p_1, \dots, p_n \in \mathbb{C}$ . Nous cherchons les racines  $a_1, a_2, \dots, a_n \in \mathbb{C}$  telles que

$$P = (X - a_1)(X - a_2) \cdots (X - a_n)$$

Le théorème de Gauss-d'Alembert nous assure que ces nombres existent : sur  $\mathbb{C}$  tout polynôme se factorise en facteurs linéaires. Il ne nous dit pas, par contre, comment les trouver.

Après nos expériences on peut espérer que pour « beaucoup » de valeurs initiales  $z_0$ , la suite de Newton  $\phi^n(z_0)$  convergera vers une racine  $a$ . Ceci donne un procédé probabiliste pour trouver une valeur approchée d'une des racines de  $P$ . Comment les trouver toutes ? On pourrait essayer différentes valeurs initiales  $z_0$  et espérer de tomber « par chance » sur les autres racines. . .

L'idée suivante se révèle plus efficace : on met la première racine trouvée  $a_1$  en facteur, par une division euclidienne  $P = (X - a_1)Q_1$ , puis on recommence avec  $Q_1$ . De manière itérée on calculera ainsi toutes les racines de  $P$ . Vous trouverez les rudiments d'une telle implémentation dans le fichier `newton.cc`. Les exercices suivants vous demandent de compléter les fonctions qui manquent.

**Remarque 1.1** (Propagation des erreurs d'arrondi). Afin d'approcher les racines de  $P$  une par une, la fonction `factoriser` met en facteur les racines déjà trouvées,  $P = (X - a_1) \cdots (X - a_k)Q_k$ . L'itération de Newton produira (au moins probablement) une racine  $a_{k+1}$  de  $Q_k$ ; sans erreurs d'arrondi, ce serait aussi une racine de  $P$ . Malheureusement, nous ne disposons que de racines approchées  $\hat{a}_1, \dots, \hat{a}_k$ , qui ne sont en général pas de racines exactes ! Déjà la première factorisation  $P = (X - \hat{a}_1)\hat{Q}_1 + \hat{r}_1$  laissera donc un petit reste  $\hat{r}_1 = P(\hat{a}_1) \neq 0$ . Les factorisations suivantes accumulent forcément de telles erreurs.

Par conséquent il faut s'attendre à un polynôme erroné  $\hat{Q}_k$ , et une racine  $\hat{a}_{k+1}$  de  $\hat{Q}_k$  peut s'éloigner considérablement de la racine correspondante  $a_{k+1}$  de  $P$ . Néanmoins elle peut nous servir pour localiser grossièrement la vraie racine  $a_{k+1}$  de  $P$ . Pour augmenter la précision de  $\hat{a}_{k+1}$  on réapplique donc l'itération de Newton par rapport à  $P$  (au lieu de  $\hat{Q}_k$ ). Cette deuxième itération convergera vite vers  $a_{k+1}$  pourvu que  $\hat{a}_{k+1}$  soit suffisamment proche. Cette approche en deux phases motive l'introduction des paramètres `localisation` et `affinage` : le premier spécifie le nombre d'itérations pour localiser une racine approchée de  $\hat{Q}_k$ , le deuxième spécifie le nombre d'itérations pour l'affinage par rapport à  $P$ .

### 2. Implémentation

**Exercice/P 2.1.** Lisez attentivement le code déjà implémenté dans `newton.cc` et survolez le fichier auxiliaire `polynome.cc`. Écrire une fonction `diff` qui implémente la dérivation  $P \mapsto P'$ .

**Exercice/P 2.2.** Pour mettre en facteur une racine approchée  $\hat{a}$ , implémenter une fonction `fact`, qui effectue la division euclidienne  $P = (X - \hat{a})Q + \hat{r}$ , puis renvoie  $Q$ .

**Exercice/P 2.3.** Afin de vérifier les racines approchées  $\hat{a}_1, \hat{a}_2, \dots, \hat{a}_n$ , on comparera  $P$  au produit  $\hat{P} = (X - \hat{a}_1)(X - \hat{a}_2) \cdots (X - \hat{a}_n)$ . Écrire une fonction `prod` qui implémente ce calcul.

### 3. Tests empiriques

**Exercice 3.1.** Pour tester votre programme, appliquez-le à quelques polynômes de votre choix, comme le polynôme  $X^3 - 2X - 5$  regardé ci-dessus. S'il vous manque d'idées, vous pouvez tester  $X^n - 1$  dont on connaît toutes les racines  $e^{2i\pi k/n}$ , ou bien  $P_n = 1 + 2X + 3X^2 + \dots + nX^{n-1} + X^n$ , pour  $n$  de plus en plus grand. Quelles difficultés rencontrez-vous ? Peut-on tout de même parvenir à une factorisation en augmentant les paramètres `localisation` et `affinage` ?

Expérimenter avec ces paramètres pour mieux comprendre la problématique des arrondis :

- (1) Sans affinage, jusqu'à quel degré les résultats sont-ils satisfaisants ?
- (2) Pour  $P_{10}$ ,  $P_{20}$ ,  $P_{30}$ , disons, quels paramètres donnent de résultats satisfaisants ?
- (3) Quand le degré croît, peut-on formuler une règle pour adapter les paramètres ?
- (4) Arrive-t-on à factoriser  $P_{50}$  ? puis  $P_{100}$  ? Quel est l'obstacle ?

**Exercice 3.2.** Les types primitifs `double` et `long double` du C++ sont très efficaces quant au temps d'exécution, mais ils imposent aussi de sévères restrictions (rappeler leur précision). Pour plus de précision on calculera avec des nombres flottants en précision arbitraire, par exemple notre classe `RReal` faite maison. Pour ceci il suffit de remplacer la première ligne par

```
#include "rreal.cc"      // nombres flottants faits maison
using namespace Numeric; // accès à l'espace de noms Numeric
typedef RReal Flo;      // synonyme utilisé dans la suite
```

Pour rassurer le compilateur les constantes littérales comme `2.0` sont à écrire comme `Flo(2.0)`, et de la même manière d'éventuelles conversions implicites ambiguës sont à rendre explicites. Avec cette précaution, votre programme devrait compiler avec la classe `RReal`. N'oubliez pas l'option `-lmpxx` : on utilise la bibliothèque GMP. Mettez au point et testez cette nouvelle version de votre programme.

**Exercice 3.3.** Expliquer pourquoi on doit calculer les premières racines avec une très grande précision si le degré  $n$  est grand. Supposons que l'on veut calculer toutes les racines à 20 décimales, soit 60 bits environs, ou bien une précision à demander à l'utilisateur. Comment adapter la précision durant les calculs intermédiaires afin d'assurer la précision souhaitée dans les résultats finaux ? Expliquez une règle pratique, implémentez-là, et testez-là sur des polynômes de plus en plus grands. Arrive-t-on à la précision souhaitée ? Comment la vérifier d'une manière convainquante ? Que dire du temps d'exécution ?

### 4. La danse des racines

On peut joindre deux polynômes unitaires  $P_0$  et  $P_1$  de degré  $n$  par le chemin  $P_t = (1-t)P_0 + tP_1$  paramétré par  $t \in [0, 1]$ . Si l'on connaît une racine  $a_0$  de  $P_0$  on peut la suivre : puisque  $P_t$  change continûment, il existe un chemin  $[0, 1] \rightarrow \mathbb{C}$ ,  $t \mapsto a_t$  tel que  $P_t(a_t) = 0$  pour tout  $t \in [0, 1]$ . Nous allons tacitement supposer que tout polynôme intermédiaire  $P_t$  est séparable : dans ce cas la proposition 3.6 nous assure que la fonction  $t \mapsto a_t$  est analytique en  $t$ .

Dans la pratique on pourra commencer par le polynôme  $P_0 = X^n - 1$  dont on connaît toutes les racines  $a_0^{(k)} = e^{2i\pi k/n}$ . On subdivise l'intervalle  $[0, 1]$  par  $0 = t_0 < t_1 < \dots < t_N = 1$ . Les racines initiales  $a_{t_0}^{(k)}$  sont connues. Ayant calculé les racines  $a_{t_j}^{(k)}$  du polynôme  $P_{t_j}$ , on peut appliquer la méthode de Newton pour le polynôme  $P_{t_{j+1}}$  : avec une subdivision suffisamment fine, la valeur initiale  $a_{t_j}^{(k)}$  convergera rapidement vers la racine  $a_{t_{j+1}}^{(k)}$  de  $P_{t_{j+1}}$ . Après  $N$  étapes on arrive finalement au polynôme  $P_1$  et toutes ses racines  $a_1^{(k)}$ .

**Exercice/P 4.1.** Si cette approche par « homotopie » vous intéresse, vous pouvez l'implémenter en étendant votre programme. Expérimenter avec la subdivision  $t_j = \frac{j}{N}$  et la précision avec laquelle vous calculez. On peut combiner cette approche avec le critère du théorème 2.25 pour assurer la convergence. Ceci permet d'adapter la subdivision localement comme nécessaire. (La seule difficulté est d'assurer  $\text{disc}(P_t)^2 > 0$ .)