II - APPROXIMATION DES NOMBRES REELS PAR DES SUITES : RECHERCHE DE SOLUTIONS APPROCHEES D'EQUATIONS NUMERIQUES.

Jean-Louis OVAERT (IREM de Marseille)

A - Analyse du problème et objectifs mathématiques

1) Représentation des nombres réels.

Un des moyens les plus puissants d'étude des nombres réels consiste, étant donné un nombre défini par un certain processus (solution d'une équation, somme d'une série, limite d'une suite, valeur d'une fonction, valeur d'une intégrale...), à *représenter* ce nombre par un autre processus mieux adapté au problème posé.

Le nombre d'Archimède π est à cet égard exemplaire. Nous nous bornons ici à une esquisse très brève, renvoyant pour plus de détail au numéro spécial du petit Archimède [23]. De sa définition à partir de la longueur du cercle ou de l'aire du disque, on peut passer à une représentation comme limite de longueurs ou d'aires de polygones réguliers inscrits ou exinscrits, ce qui permet d'en obtenir des valeurs approchées [2] [4]. L'utilisation des fonctions trigonométriques directes et réciproques permet d'obtenir d'autres représentations de π comme limite de suites ou somme de séries conduisant à des processus d'approximation plus efficaces [4] [31]. Des représentations de π utilisant des intégrales et des fractions continues permettent d'établir son irrationnalité et sa transcendance [14] [18]. Inversement, le nombre π permet d'exprimer des résultats très importants. Ainsi, la représentation de π par les intégrales de Wallis permet d'expliquer son intervention dans la formule de Stirling [4]. De même π permet d'évaluer des intégrales et des séries très remarquables. Ainsi

$$\sum_{n=0}^{+\infty} \frac{(-1)^n}{2n+1} = \frac{\pi}{4}, \quad \sum_{n=1}^{+\infty} \frac{1}{n^2} = \frac{\pi^2}{6}, \text{ et } \int_0^1 \frac{dt}{1+t^2} = \frac{\pi}{4}.$$

Ces relations sont peu adaptées au calcul de valeurs approchées de π. Elles sont néanmoins utiles en analyse numérique, pour tester *expérimentalement* la performance des procédés d'accélérations de convergence des séries et des méthodes de calcul de valeurs approchées des intégrales [4] [5] [6].

En conclusion, l'étude du nombre π met en jeu de nombreuses théories, offrant sur les problèmes posés des éclairages complémentaires. Il en est de même pour la plupart des grands problèmes mathématiques, et, en particulier, pour la résolution des équations numériques dont nous allons maintenant traiter.

2) Résolution des équations numériques.

a) Cette question comporte un aspect purement *algébrique* : étudier l'appartenance des solutions à certains domaines numériques donnés (corps des rationnels, extensions algébriques) et évaluer les solutions en fonction d'éléments convenablement choisis dans ces domaines (radicaux, résolvantes,...). A cet égard, les travaux de Lagrange, Gauss et Galois ont ouvert des voies très intéressantes [18] [42] [39] [38]. *L'aspect arithmétique* (résolution des équations diophantiennes) met en jeu une extraordinaire variété de secteurs mathématiques, et, dans bien des cas, a constitué le moteur même du développement de ces secteurs [14] [30].

Ces aspects algébriques et arithmétiques ne seront pas développés dans cet article, mais il ne faut pas les perdre de vue, même à un niveau très élémentaire.

b) Dans le cas où le domaine de nombres considéré est **R** ou **C**, on peut attaquer cette question par les méthodes de l'analyse, et on est conduit aux problèmes suivants :

Problème 1. Etudier l'existence et l'unicité des solutions ; localiser et séparer les différentes racines.

Ce problème est de type qualitatif ; il met en jeu des méthodes topologiques (théorème des valeurs intermédiaires, calcul d'indices, théorèmes de point fixe,...) [4] [16].

Problème 2. Lorsque l'équation comporte des paramètres, évaluer la *dépendance des solutions en fonction de ces paramètres* [4] [6]. Outre son intérêt purement mathématique, ce problème est essentiel pour l'analyse numérique (effet sur les solutions de *perturbations* des cœfficients, fortuites ou provoquées, évaluation de l'*influence d'un paramètre*). Il peut comporter des aspects qualitatifs (continuité des solutions) et quantitatifs (comportement asymptotique, majorations,...).

Problème 3. Etant donné une solution α de l'équation, *construire* des algorithmes d'approximation de α par une suite numérique (u_n), étudier la *convergence* de ces algorithmes et la *stabilité* de cette convergence, comparer leur *rapidité de convergence* et leur *performance*.

Au niveau élémentaire où nous nous plaçons, on pourra apprécier la rapidité de convergence en évaluant, pour tout entier p assez grand, le nombre de pas qu'il faut effectuer pour "gagner" la $p^{ième}$ décimale, c'est-à-dire passer de la précision 10^{-p+1} à la précision 10^{-p}. De même on pourra apprécier la performance en évaluant le temps total de calcul nécessaire pour obtenir la précision 10^{-p} sur un matériel de calcul donné. Nous préciserons ces points sur les exemples étudiés au B.

Les algorithmes d'approximation des racines d'une équation relèvent de principes très variés. Nous nous bornerons à étudier les suivants, selon le schéma proposé par les problèmes énoncés plus haut :
— Dichotomie et variantes (B1).
— Utilisation d'un système discret (B2).
— Utilisation d'une équation à point fixe (B3).

On trouvera d'autres types d'algorithmes dans [6] et [8].

B - Esquisse de quelques méthodes d'approximation et de leur utilisation pour l'enseignement de l'analyse à divers niveaux

I - DICHOTOMIE ET VARIANTES

1) Rappelons brièvement cette méthode :

On suppose donnée une fonction numérique f continue et strictement monotone sur un intervalle [a,b] telle que $f(a)f(b)<0$. Il existe alors un point α et un seul de]a,b[tel que $f(\alpha)=0$. Pour obtenir un algorithme d'approximation de α, on construit par dichotomie une suite $[a_n,b_n]$ d'intervalles emboîtés de la façon suivante : on pose $a_o=a, b_o=b$ et $c_o=\frac{a_o+b_o}{2}$. Deux cas peuvent se présenter : si $f(a_o)f(c_o)<0$, on prend $a_1=a$, $b_1=c_o$; sinon, on prend $a_1=c_o$, $b_1=b_o$. Supposant avoir construit a_n et b_n, on pose $c_n=\frac{a_n+b_n}{2}$.

Si $f(a_n)f(c_n)<0$, on prend $a_{n+1}=a_n$, $b_{n+1}=c_n$; sinon, on prend $a_{n+1}=c_n$, $b_{n+1}=b_n$. Dans ces conditions, la suite (a_n) est croissante, la suite (b_n) est décroissante, et, pour tout entier n, $a_n\leqslant\alpha\leqslant b_n$ et $b_n-a_n=\frac{b-a}{2^n}$.

La convergence est géométrique de raison 1/2.

2) Commentaires :

a) Sur des exemples simples (racines carrées, racines cubiques, équations algébriques) cette méthode peut être mise en œuvre bien avant la classe de première. Dans ces exemples, le nombre α est donné ; on cherche à approcher α et non à démontrer son existence.

b) On peut aussi couper à chaque pas l'intervalle en trois (trichotomie) ou en dix (décachotomie). La convergence est alors géométrique de raison 1/3 ou 1/10 suivant le cas. Il ne faut cependant pas en conclure que la décachotomie est plus performante que la dichotomie, car le nombre de "tris" à effectuer à chaque étape est en moyenne plus élevé. En fait, quelques expérimentations, sur $\sqrt{2}$, $\sqrt{5}$, $\sqrt[3]{2}$ par exemple, permettront de se convaincre que, sur une calculatrice non programmable, la dichotomie est nettement plus performante que la décachotomie ou la trichotomie, dès que l'on désire une précision de 10^{-2}, et a fortiori si l'on veut obtenir la précision 10^{-4} ou 10^{-6}. De toute façon, ces méthodes sont fastidieuses dès que l'on recherche une grande précision.

3) Objectifs

De telles activités peuvent répondre à plusieurs objectifs mathématiques, selon le niveau considéré :

a) — Familiarisation avec les suites et leur convergence.
— Exploration des ordres de grandeurs des suites géométriques.
— Construction et mise en œuvre d'un algorithme de tri.
— Pratique du calcul numérique et algébrique. Obtention de majorations et d'encadrements.
— Mise en évidence de la différence entre rapidité de convergence et performance.

b) — Familiarisation avec la méthode de dichotomie et mise en évidence sur quelques exemples de son *importance théorique* en analyse. A partir de la convergence des suites croissantes majorées, cette méthode permet de démontrer de façon simple quelques théorèmes fondamentaux sur les fonctions, dont l'énoncé figure au programme des Lycées :
— Existence des racines carrées, et des racines $n^{èmes}$.
Existence des fonctions réciproques des fonctions continues strictement monotones.
— Théorème des valeurs intermédiaires (la monotonie de f n'est pas utile pour prouver l'existence de α par dichotomie). C'est la méthode utilisée par Cauchy (cf.[35], note 3).
— Toute fonction continue sur [a,b] est bornée sur [a,b] (Raisonner par l'absurde, et effectuer une dichotomie). On peut en déduire, par une méthode due à Weierstrass, que, dans ces conditions, f atteint ses bornes supérieure M et inférieure m. (Raisonner par l'absurde et introduire les fonctions
$x\mapsto\frac{1}{M-f(x)}$ et $x\mapsto\frac{1}{f(x)-m}$).
— Principe de Lagrange : Si f est dérivable sur [a,b] et si $f'\geqslant 0$, alors f est croissante sur [a,b]. (Sinon il existe des points c et d de [a,b] tels que $c<d$ et $f(d)-f(c)<0$. On construit alors une suite dichotomique $[c_n,d_n]$ telle que pour tout entier n,

$$\frac{f(d_n)-f(c_n)}{d_n-c_n}\leqslant\frac{f(d)-f(c)}{d-c}.$$

Les suites (c_n) et (d_n) convergent vers un élément α de [a,b], et on montre que

$$\frac{f(d_n)-f(c_n)}{d_n-c_n} \to f'(\alpha).$$

L'inégalité $f'(\alpha) \leqslant \frac{f(d)-f(c)}{d-c}$ contredit l'hypothèse $f'(\alpha) \geqslant 0$).

c) Ces exemples montrent que les *activités numériques* peuvent fort bien mettre en jeu des méthodes très efficaces pour *l'approfondissement théorique* des concepts de l'analyse, qui en retour, permet de *contrôler les algorithmes utilisés* (existence et séparation des racines).

La méthode de dichotomie permet en effet de séparer les zéros d'une fonction de classe C^2 sur un intervalle compact [a,b]. Eliminons provisoirement le cas où f admet au moins un zéro multiple.

L'idée est d'effectuer une subdivision de [a,b] par dichotomie en intervalles $[c_k,c_{k+1}]$ de longueur $\frac{b-a}{2^n}$ et d'effectuer un test selon le signe de $f(c_k)\ f(c_{k+1})$. Une difficulté théorique non négligeable surgit : si $f(c_k)\ f(c_{k+1})<0$, on est sûr que f admet au moins un zéro sur $]c_k,c_{k+1}[$, mais l'unicité n'est pas assurée. De même, si $f(c_k)f(c_{k+1})>0$, il pourrait arriver qu'il y ait par exemple deux zéros, même si $\frac{b-a}{2^n}$ est très petit. A cette difficulté théorique s'ajoute une difficulté liée à la précision du calcul : si on calcule $f(c_k)f(c_{k+1})$ à la précision 10^{-p}, on ne pourra conclure que si $|f(c_k)f(c_{k+1})|$ est, par exemple, supérieur à 2.10^{-p}. On peut cependant surmonter la difficulté théorique, en observant que si $|f(c_k)|$ est grand, f(x) reste de signe constant sur $[c_k,c_{k+1}]$ si n est assez grand. Au contraire si $|f(c_k)|$ est petit, f'(x) reste de signe constant sur $[c_k,c_{k+1}]$ ce qui garantit la monotonie de f sur cet intervalle, auquel cas le test permet de conclure.

De façon précise, on est amené à déterminer des majorants

β et γ de $M_1(f)=\sup_{t\in[a,b]}|f'(t)|$ et $M_2(f)=\sup_{t\in[a,b]}|f''(t)|$.

Alors il existe un entier p tel que, pour tout entier $n \geqslant p$, et pour tout point x de [a,b], l'une au moins des conditions suivantes soit satisfaite :

a) $|f(x)|>\beta\frac{b-a}{2^{n-1}}$;

b) $|f'(x)|>\gamma\frac{b-a}{2^{n-1}}$.

On considère alors la subdivision de pas constant $\frac{b-a}{2^p}$.

— Si $f(c_{k-1})f(c_{k+1})>0$, f n'admet pas de zéro sur $[c_{k-1},c_{k+1}]$.

— Si $f(c_k)f(c_{k+1})\leqslant 0$, f admet un zéro et un seul sur $[c_{k-1},c_{k+1}]$.

(Il suffit de distinguer deux cas suivant que $|f(c_k)|$ satisfait ou non à la condition a), et on conclut à l'aide de l'inégalité des accroissements finis).

Enfin, le cas ambigu où $f(c_k)$ est très petit peut être étudié à l'aide d'une étude locale de f au voisinage de c_k.

Commentaire

On obtient ainsi un algorithme permettant de séparer les zéros de f. On prend d'abord n = 1. Lorsque $c_1=\frac{a+b}{2}$ satisfait à l'une des relations a) et b), le signe de f(a)f(b) détermine le nombre de zéros de f dans l'intervalle [a,b]. Lorsque c_1 ne satisfait à aucune de ces relations, on subdivise $[c_0,c_1]$ et $[c_1,c_2]$ en deux intervalles. On continue ce processus, et les résultats précédents montrent qu'on aboutit au bout d'un nombre fini de pas. En pratique, on a intérêt à visualiser les résultats en construisant le graphique de f point par point.

Bien entendu, cette méthode échoue en pratique lorsque les zéros de f sont très mal séparés. Néanmoins, elle permet alors de détecter un tel phénomène.

Pour un exposé plus détaillé et un choix d'exemples numériques, on pourra se reporter à[4] et [6].

4) Note historique

L'usage de la dichotomie pour séparer les racines d'une équation est fréquent au dix-huitième siècle, notamment chez L. Euler et J.L. Lagrange [36]. La première intervention théorique de la dichotomie apparaît dans le mémoire de Bolzano de 1817 sur le théorème des valeurs intermédiaires [34]. A.L. Cauchy utilise une variante de la dichotomie pour ce même théorème dans son cours d'analyse algébrique à l'école polytechnique de 1821, dont la note III est toute entière consacrée à la résolution numérique des équations. Dans ses cours d'analyse, à partir de 1860, K. Weierstrass systématise l'emploi de la dichotomie pour démontrer les théorèmes fondamentaux sur les suites de nombres réels et sur les fonctions continues ainsi que des résultats de base concernant les fonctions d'une variable complexe [29]. Dès lors, via le théorème de Bolzano-Weierstrass, la dichotomie figure parmi les procédés les plus puissants pour prouver des théorèmes d'existence, et son champ d'intervention est progressivement élargi dans deux directions importantes [12] :

— la propriété des segments emboîtés se généralise aux espaces métriques complets ;

— la propriété de subdivision en petits morceaux conduit au concept de précompacité.

Le procédé de Bolzano-Weierstrass est alors généralisé aux espaces métriques précompacts et complets, ce qui permet de l'appliquer à des problèmes très variés d'analyse fonctionnelle (équations différentielles et intégrales, théorie spectrale,...).

II. UTILISATION D'UN SYSTEME DYNAMIQUE DISCRET : METHODE DE BERNOULLI

1) Description de la méthode de Bernoulli :

a) Supposons donné un système dynamique discret dont le comportement est décrit par une suite $n \mapsto u(n)$ satisfaisant à la relation de récurrence suivante

(1) $$u(n+p) = \alpha_{p-1}u(n+p-1) + \alpha_{p-2}u(n+p-2) + \ldots + \alpha_1 u(n+1) + \alpha_0 u(n),$$

où $\alpha_0, \alpha_1, \ldots \alpha_p$ sont des nombres réels donnés, et aux conditions initiales

(2) $u(0) = \beta_0$, $u(1) = \beta_1, \ldots, u(p-1) = \beta_{p-1}$,

où $(\beta_0, \beta_1, \ldots, \beta_{p-1})$ sont des nombres réels donnés.

Il est immédiat que l'application

$$u \xrightarrow{\varphi} (u(0), u(1), \ldots, u(p-2))$$

est un isomorphisme de l'espace vectoriel E des solutions de (1) sur l'espace vectoriel $\mathbf{R}^p$. En particulier, dim E = p.

b) Parmi les solutions de (1), l'une d'elle joue un rôle privilégié. Il s'agit de la *solution fondamentale f* satisfaisant aux conditions initiales

(2′) $f(0) = 0$, $f(1) = 0$, ..., $f(p-2) = 0$, $f(p-1) = 1$.

Introduisons l'opérateur de translation T défini par la relation $(Tu)(n) = u(n+1)$. Il est immédiat que si $u \in E$, $Tu \in E$. En particulier, $f_1 = f$, $f_2 = Tf, \ldots, f_p = T^{p-1}f$ appartiennent à E. Il est alors facile de voir, en utilisant l'isomorphisme φ que $(f_1, f_2, \ldots, f_p)$ est une base de E.

Ainsi, la solution fondamentale f suffit pour engendrer, par translations successives, l'espace de toutes les solutions de (1). En outre la résolution explicite du problème de Cauchy (1) et (2) est alors très facile : on décompose la solution u dans la base $(f_1, \ldots, f_p)$. Le calcul des composantes de u se ramène à la résolution d'un système *triangulaire*.

c) L'équation (1) s'écrit encore $P(T)(u) = 0$, où $P = X^p - \alpha_{p-1}X^{p-1} \ldots - \alpha_1 X - \alpha_0$.

Ce polynôme s'appelle polynôme caractéristique de l'équation (1). Bornons-nous, pour simplifier, au cas où toutes les racines de P sont simples, et écrivons P sous la forme $P = \prod_{j=1}^{p} (X - \lambda_j)$, où, pour tout $j \in [1,p], \lambda j \in \mathbf{C}$.

L'étude du cas particulier $P = X - \lambda$, i.e. $u(n+1) = \lambda u(n)$, conduit à introduire la suite géométrique $e_\lambda : n \mapsto \lambda^n$.

On prouve alors que $(e_{\lambda_1}, e_{\lambda_2}, \ldots, e_{\lambda_p})$ est une base de l'espace vectoriel E.

d) Supposons maintenant que l'on veuille étudier le comportement asymptotique de la solution de (1), satisfaisant aux conditions initiales (2). On évalue f dans la base précédente. Pour tout entier n,

(3) $u(n) = a_1\lambda_1^n + a_2\lambda_2^n + \ldots + a_p\lambda_p^n$.

Plaçons nous dans le cas simple où il existe une racine réelle de P, soit λ_1, telle que, pour tout $j \in [2,p]$, $|\lambda j| < |\lambda_1|$. Alors, si $a_1 \neq o$,

(4) $u(n) \sim a_1\lambda_1^n$.

Le comportement symptotique de u est donc gouverné par la racine de plus grand module du polynôme caractéristique P.

e) Un exemple célèbre est celui de la suite de Fibonacci

(1) $u(n+2) = u(n+1) + u(n)$

(2) $u(0) = 0 \quad u(1) = 1$

étudiée en 1202 par Léonard de Pise (alias Fibonacci) à propos de la croissance d'une population de lapins.

Ici $P = X^2 - X - 1$, $\lambda = \frac{1+\sqrt{5}}{2}$, $\mu = \frac{1-\sqrt{5}}{2}$.

Donc $|\lambda| > 1$ et $|\mu| < 1$.

D'autre part $u(n) = \frac{1}{\sqrt{5}} [\lambda^n - \mu^n]$;

Donc $u(n) \sim \frac{1}{\sqrt{5}} \left(\frac{1+\sqrt{5}}{2}\right)^n$.

On peut vérifier cette évaluation expérimentalement, à la main ou sur calculatrice.

Remarque. Ici le coefficient de λ^n n'est pas nul. Il est intéressant de considérer la solution $u = e_\mu$ de (1), qui satisfait aux conditions initiales

$$e_\mu(o) = 1 \quad , \quad e_\mu(1) = \mu = \frac{1-\sqrt{5}}{2} .$$

Comme $|\mu| < 1$, $e_\mu(n) = \mu^n \to 0$. Cependant si on calcule les valeurs successives de u en utilisant (1) et (2), on aboutit au résultat paradoxal que $|u(n)|$ après une période de décroissance vers 0, se met à croître indéfiniment ! Ce paradoxe s'explique notamment par le fait que la calculatrice fait une erreur d'arrondi sur $u(1) = \frac{1-\sqrt{5}}{2}$; le calcul effectué n'est pas celui de e_μ mais d'une solution dont la composante suivante λ est *non nulle*, bien que très petite (de l'ordre de 10^{-8}, ou 10^{-10}, suivant la calculatrice utilisée). Au début du calcul, c'est le terme en μ^n qui est prépondérant, et $|u(n)|$ décroît vers 0, mais à partir d'un certain rang, le terme en λ^n prend le pas sur l'autre, si bien que $|u(n)| \to +\infty$.

f) Cet inconvénient n'apparaît jamais lorsque l'on considère la solution fondamentale f. En effet, aucune des composantes de f dans la base $(e_{\lambda_1}, \ldots, e_{\lambda_p})$ n'est nulle (sinon, les translatées de f appartiendraient à un sous espace vectoriel strict de E, ce qui contredit le fait que $(f_1, f_2, \ldots, f_p)$ est une base de E).

Considérons donc la solution fondamentale f de (1) et plaçons-nous dans le cas, décrit au d), où $|\lambda_j| < |\lambda_1|$.

Alors, d'après f), $a_1 \neq 0$. Donc $f(n) \sim a_1\lambda_1^n$, et, par suite, à partir d'un certain rang $f(n) \neq 0$, et

(5) $$\lim_{n \to +\infty} \frac{f(n+1)}{f(n)} = \lambda_1.$$

g) L'idée de D. Bernoulli est de *renverser toute la situation précédente*. Cette fois, on se donne un polynôme P dont toutes les racines sont simples, (mais inconnues). On suppose que, $\forall j \in [2,p], |\lambda_j| < |\lambda_1|$. Pour

trouver des valeurs approchées de λ_1, on associe à ce polynôme P le système dynamique défini par les conditions

$f(n+p) = \alpha_{p-1} f(n+p-1) + \ldots + \alpha_1 f(n+1) + \alpha_0 f(n)$

$f(0) = 0,\ f(1) = 0,\ \ldots,\ f(p-2) = 0,\ f(p-1) = 1.$

Alors $\lambda_1 = \lim_{n \to \infty} \dfrac{f(n+1)}{f(n)}$.

On peut facilement évaluer la rapidité de convergence de cette suite :

soit $k = \dfrac{\sup_{j \neq 1} |\lambda_j|}{|\lambda_1|}$; alors $\dfrac{f(n+1)}{f(n)} - \lambda_1$ est dominé par une suite géométrique de raison k. En particulier, la rapidité de convergence est d'autant plus grande que $|\lambda_1|$ est mieux séparé des autres valeurs $|\lambda_j|$.

h) *Quelques exemples numériques* permettant une analyse de la pertinence de la méthode précédente.

$\alpha)$ $P = X^3 - 2X^2 - X + 1$	Trois racines réelles λ, μ, ν ;
$\lambda \in]2,3[$	Appliquer directement l'algorithme de Bernoulli.
$\mu \in]1/2,1[$	Translation $X \mapsto X - 1$, et passage à l'équation aux inverses.
$\nu \in]-1,-1/2[$	Translation $X \mapsto X + 1$, et passage à l'équation aux inverses.
$\beta)$ $P = X^2 - X - 1$	(ce n'est pas une plaisanterie !)
$\gamma)$ $P = X^3 - X^2 - 2X + 1$	(analogue à α)
$\delta)$ $P = X^3 - 2X - 5$	Il y a une seule racine réelle λ, et $\lambda \in]2,3[$. Les racines complexes μ et $\bar{\mu}$ satisfont à $\lvert\mu\rvert^2 \lambda = 5$, donc $\lvert\mu\rvert < \lambda$, et la méthode de Bernoulli s'applique.
$\varepsilon)$ $P = X^3 - 7X + 7$	Trois racines réelles λ, μ, ν ; $\lambda \in]-4,-3[$ et $\mu, \nu, \in]1,2[$.

On trouvera d'autres exemples numériques dans le texte très remarquable de L. Euler [37] constituant le chapitre 17 de l'introduction à l'analyse des infiniments petits. Voir aussi [8], chapitre 7.

Remarque. Bien entendu, le rôle de ces exemples n'est pas de fournir une illustration numérique de la théorie esquissée ci-dessus. C'est à travers leur étude que l'on pourra aborder les différents facteurs mis en jeu, et mettre à l'épreuve son domaine de fonctionnement.

2) Commentaires.

Cette méthode permet de déterminer des valeurs approchées de toutes les racines réelles d'un polynôme P scindé sur **R** dont toutes les racines sont simples. On commence par séparer les racines de P. Pour calculer une valeur approchée de l'une de ces racines, notée λ, on effectue une translation de telle sorte que λ soit la racine de plus petit module. (Il suffit de placer l'origine au milieu d'un intervalle ne contenant que la racine λ.) On forme alors l'équation aux inverses, pour laquelle l'hypothèse de la méthode de Bernoulli est satisfaite, ce qui fournit un algorithme d'approximation de λ.

Remarque. Ayant déjà déterminé la plus grande racine λ_1, on pourrait essayer d'obtenir la racine suivante en introduisant la suite $f(n+1) - \lambda_1 f(n)$, et en prenant la limite du quotient de termes consécutifs. Cette méthode est d'une précision numérique désastreuse, car $f(n+1)$ et $\lambda_1 f(n)$ sont équivalents, et l'erreur commise sur λ_1 devient rapidement prépondérante lorsque n augmente. La méthode de Bernoulli peut s'étendre au cas où P admet des racines complexes, mais la technicité requise est nettement plus forte ; on pourra se reporter à [4], [6] et [8].

3) Objectifs.

a) La méthode de Bernoulli constitue un excellent terrain combinant *approfondissement expérimental* et *approfondissement théorique* :

- C'est bien entendu sur des exemples numériques que l'on peut aborder cette méthode, très simple à mettre en œuvre sur le plan expérimental.
- Les facteurs gouvernant la rapidité de convergence peuvent être mis facilement en évidence ; des transformations algébriques simples permettent d'agir sur ces facteurs.
- C'est encore sur des exemples très simples que l'on peut cerner le domaine de fonctionnement de la méthode (racines complexes, racines multiples ou mal séparées), de préciser les cas de stabilité ou d'instabilité et d'agir sur cette stabilité.
- Ce dernier point est à relier à la pertinence des moyens de calculs utilisés, notamment en ce qui concerne l'influence des erreurs d'arrondi.

Remarque 1. A un niveau d'approfondissement plus élevé, on peut comparer la performance de cette méthode à celle de variantes (calcul des sommes de Newton des racines, méthode de Graeffe ; voir [8] chap. 7 et [6] chapitre 7, § 3).

Remarque 2. Le texte de L. Euler cité plus haut met en valeur les aspects précédents d'une façon magistrale. Des extraits de ce texte peuvent être directement utilisés par les élèves.

b) La méthode de Bernoulli met en évidence l'importance du travail *intersectoriel* en mathématiques.

- D'une part la connaissance des racines du polynôme caractéristique (qui relève d'un problème *algébrique*) détermine le comportement asymptotique d'un système dynamique discret (problème *d'analyse*) lui-même pouvant être relié à des secteurs scientifiques très variés (croissance de populations, évolution de systèmes économiques et biologiques, problèmes combinatoires,...)
- Réciproquement, ce comportement asymptotique fournit une méthode de recherche de valeurs approchées des racines d'un polynôme. Il s'agit ici d'un exemple très intéressant de *simulation* d'une équation

algébrique par un *système dynamique discret* gouverné par cette équation, qui offre prise aux moyens de *calcul arithmétique*. Cette idée a une grande portée : on peut par exemple l'employer pour la recherche des *valeurs propres* et des *vecteurs propres* des matrices. (Voir par exemple [17], chap. 7, [8] tome II, chap. 11, et [6] chap. 8, § 7).

— Une méthode en tous points analogue permet de relier une équation algébrique au comportement d'un *système dynamique continu*, décrit pour des équations différentielles linéaires à coefficients constants, dont l'étude est reliée à celle de systèmes variés (mécaniques, électrique, biologiques,...) voir [11] et [15], ce qui permet des moyens de *calcul analogique*.

— Enfin, les concepts de *l'algèbre et de l'analyse linéaire* jouent un rôle central dans les démarches précédentes.

4) Note historique.

L'utilisation des équations algébriques pour l'étude de récurrences linéaires trouve son origine dans l'exemple célèbre de la suite de Fibonacci, qui figure dans le Liber Abacci de Léonard de Pise (1202) ; on remarquera que cette suite est associée à un système dynamique (croissance d'une population de lapins). Principalement en vue d'étudier le calcul des différences finies, l'emploi des méthodes algébriques s'impose progressivement à travers les travaux de Briggs, Wallis et Newton. Les travaux de Newton et des frères Jean et Jacques Bernoulli mettent en lumière l'importance du rôle des développements en série entière des fractions rationnelles. Ces travaux fonctionnent dans le sens : équations algébriques → équations récurrentes et différentielles. En 1728, Daniel Bernoulli [31] met en évidence l'intérêt de la démarche inverse. En 1748, Léonard Euler, dans le chapitre 17 de l'introduction à l'analyse des infiniments petits [37], effectue une étude systématique de la méthode ; l'emploi des développements en série entière, la convergence est envisagée sous un angle quantitatif, et ne fait l'objet d'aucune étude théorique. En 1759, Lagrange introduit le concept général d'équation caractéristique, et développe une théorie systématique des équations aux différences finies [43][44]. Enfin Laplace systématise l'emploi des séries génératrices d'une suite numérique [45]. Les extensions de la méthode de Bernoulli à la recherche des valeurs propres et des vecteurs propres des matrices ont été introduites à partir de 1930 et connaissent maintenant un développement important lié à l'apparition de moyens de calcul très performants. Pour cette extension, on pourra se reporter à [6], [8], [17].

III. UTILISATION D'UNE EQUATION A POINT FIXE

1) Introduction

On considère une équation numérique de la forme $F(x) = 0$, où F est de classe C^1 sur I, et admet un zéro a et un seul dans I. (On peut se ramener à ce cas, après séparation des racines de l'équation $F(x) = 0$).

Parmi les équations de ce type, figurent les *équations à point fixe*, du type $f(x) = x$, pour lesquelles on dispose, sous des hypothèses convenables, d'un algorithme de recherche de solutions approchées très efficace, consistant à utiliser la suite récurrente $u_n = f(u_n)$. (*Méthode des approximations successives*).

Dès lors la stratégie est claire : d'une part, étudier la méthode du point fixe et les facteurs qui gouvernent la performance de cette méthode ; d'autre part, construire des procédés permettant de ramener l'équation donnée $F(x) = 0$ à une équation à point fixe pour laquelle la méthode des approximations successives est performante.

2) Enoncé des problèmes

Nous allons préciser la démarche esquissée ci-dessus :

Considérons une fonction de classe C^1 sur un intervalle I de **R**, telle que $f(I) \subset I$. On se pose les quatre problèmes suivants :

Problème 1.

Existe-t-il un point a de I tel que $f(a) = a$? Ce point est-il unique ?

Problème 2.

Soit alors c un élément de I. La suite (u_n) définie par la relation de récurrence $u_{n+1} = f(u_n)$ et la condition initiale $u_0 = c$, converge-t-elle vers a ?

Problème 3.

Si oui, étudier la rapidité de convergence de (u_n) vers a, et la stabilité de cette convergence. Déterminer les facteurs qui gouvernent cette rapidité de convergence.

Les problèmes 1 et 2 sont d'ordre qualitatif ; le problème 3 est d'ordre quantitatif. Nous verrons que c'est la pente des sécantes, ou des tangentes, du graphe de f qui joue un rôle essentiel.
D'où le problème suivant.

Problème 4

Soit F une fonction de classe C^1 sur I admettant un zéro a et un seul sur I. Déterminer une fonction g de classe C^1 sur un intervalle J contenant a et contenu dans I satisfaisant aux deux conditions suivantes :

a) L'équation $F(x) = 0$ équivaut à l'équation $g(x) = x$ lorsque $x \in J$.

b) La valeur de $|g'(a)|$ est très petite, et, si possible, nulle.

La recherche de telles fonctions g procède de méthodes très diverses. Il s'agit alors de comparer la performance de ces différentes méthodes.

Dans ce qui suit, nous nous bornerons à étudier quelques points clefs, renvoyant pour plus de détail à la brochure *Analyse I* publiée par l'IREM de Marseille [5], partie 4, où l'on trouvera une bibliographie détaillée.

3) Etude de l'existence et unicité d'une solution et de sa stabilité. (Résolution des problèmes 1 et 2).

On dispose de plusieurs types d'énoncés. Le premier est de type *topologique* et s'appuie sur la notion de *connexité* :

Théorème 1. Soit f une application continue d'un intervalle compact $[\alpha,\beta]$ dans lui-même. Alors f admet au moins un point fixe.

(Il suffit d'appliquer le théorème des valeurs intermédiaires à la fonction $x \mapsto f(x)-x$).

On notera que ce théorème ne s'étend pas au cas d'un intervalle I non compact, même si I est fermé ; ainsi, lorsque $I=\mathbf{R}$, la fonction $f : x \mapsto x+1$ n'admet aucun point fixe.

D'autre part, ces points fixes ne s'obtiennent pas nécessairement par itération : c'est le cas lorsque

$$I = [-1, +1] \quad \text{et} \quad f(x) = -x.$$

On peut cependant décrire des algorithmes permettant d'approcher un point fixe, par exemple en utilisant une barycentration. (Voir Analyse I [5]).

L'énoncé le plus classique est de type *métrique* et se fonde sur la notion de *complétion*.

Théorème 2. Soient I un intervalle fermé de **R**, et f une application de I dans lui-même lipschitzienne dans un rapport $k < 1$. (On dit que f est k-contractante).

1. Il existe un élément a de I et un seul tel que $f(a) = a$.
2. Pour tout élément c de I, la suite définie par les relations $u_{n+1} = f(u_n)$ et $u_0 = c$ converge vers a.
3. Pour tout entier n,

$$(1) \qquad |u_n - a| \leqslant \frac{k^n}{1-k} |u_1-u_0|$$

Idée de la démonstration. Pour obtenir *l'existence du point fixe, et la convergence de* (u_n) *vers a,* on montre que la suite (u_n) est de Cauchy en utilisant la majoration

$$|u_q-u_p| \leqslant |u_{p+1}-u_p| + |u_{p+2}-u_{p+1}| + \ldots +$$
$$|u_q-u_{q-1}| \leqslant |u_1-u_0| (k^p+k^{p+1} \ldots\ldots + k^{q-1})$$

un passage à la limite dans cette relation fournit alors la relation (1).

Remarque 1. L'assertion 3 fournit une première indication sur la rapidité de convergence : elle est au moins d'ordre géométrique, et gouvernée par la valeur de k. Elle assure aussi la *stabilité* de l'algorithme utilisé.

Remarque 2. Lorsque f est continûment dérivable sur I, à dérivée bornée, le théorème des accroissements finis montre que f est lipschitzienne dans le rapport $k = \sup |f'(t)|$. *C'est donc la pente des tangentes qui gouverne la rapidité de convergence ;* plus précisément, comme $u_n \to a$, *c'est la valeur de* $f'(a)$ *qui gouverne la rapidité asymptotique de convergence.* Nous préciserons ce point au d.)

Remarque 3. Il suffit pas que, pour tout couple (x,y) d'éléments distincts de I, $|f(x) - f(y)| < |x-y|$. Ainsi, lorsque $I = [1,+\infty[$ et que $f(x) = x + \frac{1}{x}$, cette condition est réalisée, mais f n'a aucun point fixe. "Moralement", le point fixe est rejeté à l'infini. Pour empêcher l'apparition de ce phénomène, on peut supposer que I est *compact*.

Théorème 3. Soient $I = [\alpha,\beta]$ un intervalle compact de **R** et f une application de I dans lui-même telle que, pour tout couple (x,y) de points distincts de I,

$$|f(x) - f(y)| < |x-y| .$$

Alors

1. L'application f admet un point fixe a et un seul.
2. Pour tout élément c de I, la suite définie par les relations $u_{n+1} = f(u_n)$, $u_0 = c$ converge vers a.

Pour obtenir l'existence de a, on observe que la fonction continue $x \mapsto |f(x) - x|$ atteint sa borne inférieure sur l'intervalle compact $[\alpha,\beta]$ en au moins un point a. Alors $f(a) = a$. La démonstration de l'assertion 2 est plus délicate ; elle s'appuie sur la notion de valeur d'adhérence d'une suite ; on la trouvera esquissée dans Analyse 1 [5], et dans [4].

Contrairement au précédent, cet énoncé ne fournit plus aucune indication sur la rapidité de convergence, qui peut d'ailleurs être arbitrairement lente.

Notons finalement que l'existence d'une limite pour la suite (u_n) peut aussi être obtenue par *monotonie*. Plus précisément :

Théorème 4. Soient $I = [\alpha,\beta]$ un intervalle compact de **R**, f une application continue de I dans lui-même, c un point de I et (u_n) la suite définie par les relations $u_{n+1} = f(u_n)$ et $u_0 = c$. Soit enfin F l'ensemble des points fixes de f (qui est fermé et non vide d'après le théorème 1).

1. On suppose que f est croissante sur I.

— Si $u_1 \geqslant u_0$ la suite (u_n) est croissante et converge vers le plus petit élément a de $F \cap [u_0, +\infty[$.

— Si $u_1 \leqslant u_0$ la suite (u_n) est décroissante et converge vers le plus grand élément b de $]-\infty, u_0] \cap F$.

Dans les deux cas, on dit que l'itération est "en escalier" (voir figures 1 et 2).

2. On suppose que f est décroissante sur I ; alors f admet un point fixe a et un seul.

— Si $u_0 \leqslant a$ et $u_2 \geqslant u_0$, la suite (u_{2n}) est croissante et la suite (u_{2n+1}) est décroissante ; elles convergent toutes deux vers a.

— Si $u_0 \geqslant a$ et $u_2 \leqslant u_0$, la suite (u_{2n}) est décroissante et la suite (u_{2n+1}) est croissante ; elles convergent toutes deux vers a.

— Si $u_0 < a$ et $u_2 \leqslant u_0$, la suite (u_{2n}) est décroissante et la suite (u_{2n+1}) est croissante ; elles convergent vers deux points fixes b et c de f o f, tels que $b<a<c$. En particulier, la suite (u_n) diverge.

— Si $u_0 > a$ et $u_2 \geqslant u_0$, la suite (u_{2n}) est croissante et la suite (u_{2n+1}) est décroissante ; elles convergent vers deux points fixes b et c de f o f, tels que $c<a<b$. En particulier, la suite (u_n) diverge.

Dans les quatre cas on dit que l'itération est en spirale (cf. figures 3 et 4).

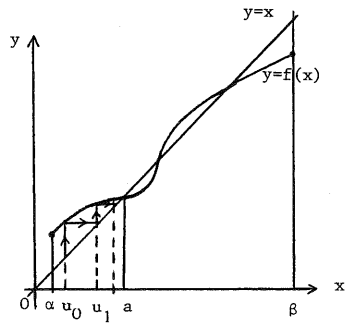


figure 1 : escalier croissant

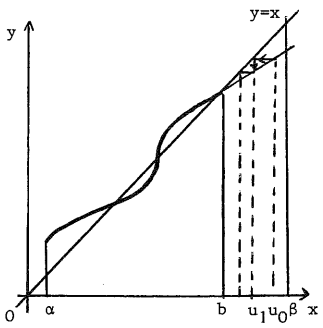


figure 2 : escalier décroissant

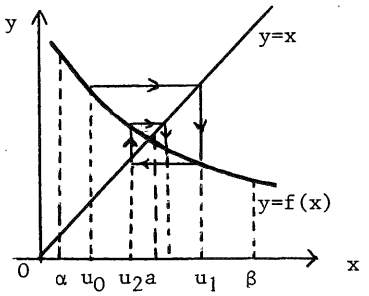


figure 3 : spirale convergente

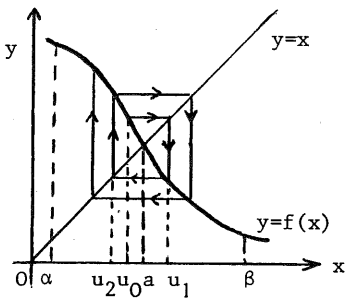


figure 4 : spirale divergente

La démonstration est immédiate.

Pour l'assertion 2, on observe que f o f est croissante.

Remarque 1. En dehors des cas couverts par les théorèmes 2 à 4, le comportement de la suite (u_n) peut être *très complexe,* comme le montre l'exemple, en apparence très simple, où $I = [0,1]$ et $f(x) = \alpha x(1-x)$, où $\alpha \in [0,4]$. L'étude topologique du comportement de (u_n) est alors liée de façon étroite à celle des points périodiques, i.e. des solutions de l'équation $f^n(x) = x$. Cet aspect fait à l'heure actuelle l'objet de nombreux travaux mathématiques (cf. notamment les travaux de Smale et de Jonker).

Remarque 2. Pour ce qui est de la stabilité de l'algorithme des approximations successives, on pourra se reporter à Analyse I [5]. Il nous suffira ici de savoir que cette stabilité est assurée si $|f'(a)|$ est nettement inférieur à 1.

4) Etude de la rapidité de convergence (Etude du problème 3)

Considérons à nouveau un intervalle fermé I de **R**, et f une application de I dans lui-même, continûment dérivable sur I. On suppose que f admet un point fixe a et un seul (nous avons étudié au 3) des conditions suffisantes pour qu'il en soit ainsi). Nous nous bornerons à étudier les cas

$$0 < |f'(a)| < 1 \quad , \quad f'(a) = 0,$$

qui assurent la stabilité de l'algorithme.

a) Cas où $0 < |f'(a)| < 1$.

Considérons un nombre réel positif k tel que

$$|f'(a)| < k < 1.$$

Comme f′ est continue, il existe un intervalle compact $J = [a-\alpha, a+\alpha]$, $\alpha > 0$, tel que pour tout point x de $J \cap I$, $|f'(x)| \leqslant k$.

Alors $J \cap I$ est un intervalle compact stable par f, et pour tout point c de $J \cap I$, la suite (u_n) définie par les relations $u_{n+1} = f(u_n)$ et $u_0 = c$ converge vers a et

$$(1) \qquad |u_n - a| \leqslant \frac{k^n}{1-k} |u_1 - u_0|.$$

$$(2) \qquad |u_n - a| \leqslant k^n \, |c - a|.$$

Ainsi, (u_n) *converge vers* a, *et* $|u_n - a|$ *est dominée par* k^n.

Remarque. L'inégalité (2) ne peut pas être fortement améliorée : prenons k′ tel que $0 < k' < |f'(a)|$. Alors,

$$(3) \qquad k'^n \, |u_0 - a| \leqslant |u_n - a|.$$

Les relations (2) et (3) suggèrent que $|u_n - a|$ se comporte comme ρ^n, où $\rho = |f'(a)|$ (mais elles ne suffisent pas à le prouver !).

Le résultat suivant, dont on trouvera la démonstration dans Analyse I [5], permet de conclure :

Théorème 1. Evaluation de la rapidité de convergence si $0 < |f'(a)| < I$.

Soit f une fonction de classe C^1 sur I satisfaisant aux deux conditions suivantes :

a) $0 < |f'(a)| < 1$

b) $f(x) = a + (x-a)\,f'(a) + 0(|x-a|^s)$, où $s > 1$.

Il existe alors un nombre réel $b > 0$ tel que

$$(4) \qquad |u_n - a| \sim b\,|f'(a)|^n.$$

Ce résultat fournit une estimation très précise de la rapidité de convergence, et montre en particulier que cette rapidité est gouvernée par $|f'(a)|$.

Remarque. Bien entendu, en pratique, on ignore la valeur exacte de a, et de $f'(a)$. C'est pourquoi on utilisera plutôt la relation (2), en prenant pour k le maximum de $|f'|$ sur un petit voisinage compact de a.

b) Cas où $f'(a) = 0$.

La première partie du raisonnement précédent reste valable, et montre que, pour tout nombre k tel que $0 < k < 1$, $|u_n - a|$ est dominée par k^n. Il faut donc s'attendre à une convergence très rapide. Pour majorer $|u_n - a|$ nous ferons l'hypothèse suivante satisfaite dans la quasi-totalité des cas rencontrés en pratique, et d'interprétation géométrique très simple :

Il existe un nombre réel r strictement supérieur à 1 et un nombre réel strictement positif λ tels que, au voisinage de a,

$$|f(x) - f(a)| \sim \lambda\,|x - a|^r.$$

On notera que cette condition est satisfaite lorsque f est de classe C^r sur I, où $r \geqslant 2$, et que

$$f'(a) = f''(a) = \ldots = f^{(r-1)}(a) = 0,\ f^{(r)}(a) \neq 0.$$

Soient alors μ et μ' des nombres réels tels que

$$\mu' < \lambda < \mu.$$

Il existe un nombre réel strictement positif η tel que, pour tout élément de x de $J = [a - \eta, a + \eta]$,

$$\mu'\,|x-a|^r \leqslant |f(x) - f(a)| \leqslant \mu\,|x-a|^r.$$

En outre si $\eta^{r-1} < \dfrac{1}{\mu}$, alors J est stable par f.

On suppose désormais que ces conditions sont réalisées et que c appartient à J. On pose

$$k = \mu^{1/(r-1)}\,|c-a| \text{ et } k' = \mu'^{1/(r-1)}\,|c-a|.$$

Il est immédiat que $0 < k' < k < 1$ et que, pour tout entier naturel n,

$$(5) \qquad k'^{(r^n)}\,\frac{|c-a|}{k'} \leqslant |u_n - a| \leqslant k^{(r^n)}\,\frac{|c-a|}{k}.$$

Comme $k < 1$, la rapidité de convergence est beaucoup plus grande que dans le cas où $|f'(a)| \neq 0$. Cette rapidité est essentiellement gouvernée par le nombre r, *ordre du contact* au point fixe a entre la courbe $y = f(x)$ et la tangente (horizontale) à cette courbe. Lorsqu'une suite (u_n) satisfait à une relation du type (5) on dit que la *convergence est d'ordre* r ; lorsque $k = 1/10$, à chaque pas on *multiplie par r le nombre de décimales exactes.*

Un calcul asymptotique permet de démontrer le résultat suivant, dont la démonstration est donnée dans Analyse I [5] et dans [4].

Lorsque c appartient à J, il existe un nombre réel δ appartenant à $]0,1[$ tel que

$$(6) \qquad |u_n - a| \sim \delta^{(r^n)}\,\frac{|c-a|}{\delta}.$$

En outre, $k' \leqslant \delta \leqslant k$.

Remarque 1. On peut aussi étudier la rapidité de convergence lorsque $|f'(a)| = 1$, auquel cas on obtient des convergences en $1/n^s$, c'est-à-dire lentes. L'intérêt didactique de ce cas est donc de fournir un procédé de construction d'approximations lentes. Nous renvoyons pour plus de détails à Analyse I [5]. Des exemples très simples de suites convergeant lentement vers 0 peuvent être étudiés ; par exemple

$$I = [0, +\infty[,\ f(x) = \frac{x}{1+x}\ ;\ I = [0, +\infty[,\ f(x) = x\,e^{-x}\ ;$$

$$I = [0, \pi/2],\ f(x) = \sin x\ ;\ I = [0, +\infty[,\ f(x) = 1 - e^{-x}.$$

5) Méthodes de transformations d'une équation en une équation à point fixe (Etude du problème 4).

Nous nous bornons ici à esquisser quatre méthodes, très classiques, renvoyant aux bons ouvrages [6] et [8] pour d'autres algorithmes.

a) Linéarisation de l'équation par le calcul différentiel (Méthode de Newton-Raphson)

Décrivons d'abord *l'idée* intuitive gouvernant cette méthode. Supposons que l'on connaisse déjà une valeur approchée c de l'unique nombre a tel que $F(a) = 0$. D'après la définition de la dérivée de F au point c, pour tout point x de I,

$$F(x) = F(c) + (x-c)\,F'(c) + o(x-c).$$

Comme $F(a) = 0$, et comme $a - c$ est petit, on peut vraisemblablement espérer que $F(c) + (a-c)\,F'(c)$ est petit par rapport à $(a-c)$.

Dans ces conditions on remplace l'équation *non linéaire* $F(x) = 0$ par l'équation *linéaire* approchée

$$(1) \qquad F(c) + (x-c)F'(c) = 0.$$

Supposons que F' ne s'annule pas sur I ; alors (1) admet une solution et une seule

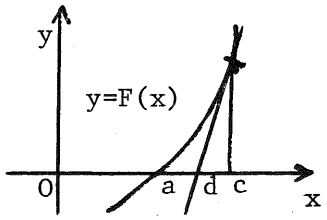
$$(2) \qquad d = c - \frac{F(c)}{F'(c)}.$$

On a ainsi construit une "meilleure" valeur approchée d de a. Il reste alors à itérer ce processus, en prenant d comme valeur initiale à la place de c. On est donc conduit à envisager la suite (u_n) définie par les relations.

$$(3) \quad u_{n+1} = u_n - \frac{F(u_n)}{F'(u_n)}. \quad \text{(Algorithme de Newton)}$$
$$u_0 = c$$

Interprétation géométrique de la méthode de Newton. On approche a par l'abscisse d de l'intersection avec Ox de la tangente au graphe de F au point c.

Il reste à préciser des conditions sous lesquelles (u_n) converge vers a.



Ici, on remplace l'équation $F(x) = 0$ par l'équation $g(x)=x$, où $g(x)=x-\frac{F(x)}{F'(x)}$. En appliquant les résultats sur les équations à point fixe, on peut étudier la rapidité de convergence de la méthode de Newton (voir Analyse I). Comme $g'(a) = 0$, la convergence est d'ordre deux, c'est-à-dire qu'on double le nombre de décimales exactes à chaque pas.

Remarque 1. L'expérimentation et l'analyse théorique de la rapidité de convergence mettent en lumière un point très important : la convergence ne devient très rapide que si l'on initialise avec une valeur c approchant déjà assez bien a (par exemple $|c-a| \leqslant \frac{1}{10}$). C'est pourquoi, en pratique, il peut y avoir avantage à dégrossir la résolution de $F(x) = 0$ par une autre méthode (dichotomies, Bernoulli,...) et à passer ensuite à l'algorithme de Newton. Le cas des racines carrées est à cet égard éclairant : si on initialise la recherche de $a = \sqrt{1234}$ avec $c = 1000$, la convergence est lente au début. Si on calcule d'abord la partie entière de a à savoir 35, et si l'on prend $c = 35$, la convergence est très rapide dès le départ. Cela tient, bien entendu, au fait que la condition $f'(a) = 0$ implique seulement que f' est petit *au voisinage de* a .

Remarque 2. Il n'est pas toujours aisé de vérifier que les conditions de validité de la méthode de Newton sont satisfaites ; mais, en pratique, cette vérification est inutile. Ayant obtenu une valeur approchée b de a, on teste le signe de $F(b+10^{-p})$ et de $F(b-10^{-p})$, si p fixe la précision demandée. Sous l'hypothèse où F est monotone, ces tests permettent de prouver en toute rigueur que $|b-a| \leqslant 10^{-p}$ (tout au moins, tant que les erreurs d'arrondi dans le calcul de F(x) sont négligeables devant 10^{-p}).

b) Ajustement linéaire. Considérons une équation de la forme $F(x) = 0$, admettant une solution a et une seule, et supposons que $F'(a) \neq 0$, ce qui signifie que a est un zéro simple de F. L'idée consiste à transformer l'équation $F(x) = 0$ en l'équation équivalente

$$F(x)-\lambda x = -\lambda x,$$

où λ est un nombre réel non nul à choisir convenablement. On est alors ramené à l'équation à point fixe

$$g(x) = x \ , \quad \text{où } g(x) = x - \frac{F(x)}{\lambda} \ .$$

Comme $g'(x) = 1 - \frac{F'(x)}{\lambda}$, on choisit $\lambda = F'(a)$ de telle sorte que $g'(a) = 0$. En pratique il est rare que l'on connaisse $g'(a)$. On prendra donc λ de telle sorte que $|g'(a)|$ soit suffisamment petit. La convergence est géométrique.

Remarque 1. Lorsque l'équation de départ est sous la forme $f(x) = x$, on se ramène à ce cas en posant

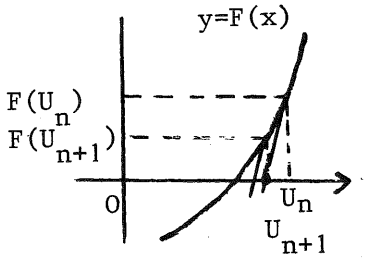
$$F(x) = f(x) - x.$$

Remarque 2. *Interprétation géométrique.*

La relation $u_{n+1} = g(u_n)$ équivaut à la relation

$$\lambda(u_{n+1} - u_n) + F(u_n) = 0.$$

Le point d'abscisse u_{n+1} s'obtient donc en coupant l'axe 0x par la droite passant par le point $(u_n, F(u_n))$ de pente λ.



Remarque 3. *Nouvelle approche de la méthode de Newton.*

En pratique, il est rare qu'on connaisse $F'(a)$, et il n'est pas possible d'effectuer le choix optimal $\lambda = F'(a)$. Ayant déjà obtenu un encadrement assez fin de a par une autre méthode (dichotomie, méthode de Bernoulli...) on pourra prendre $\lambda = F'(c)$, où c est voisin de a et poser $u_0 = c$. On peut alors penser à changer λ à chaque pas de façon à se rapprocher de la valeur optimale $\lambda = F'(a)$, ce qui conduit à l'algorithme suivant

$$u_0 = c \qquad \lambda_0 = F'(u_0)$$

$$u_1 = u_0 - \frac{F(u_0)}{\lambda_0} \qquad \lambda_1 = F'(u_1)$$

$$u_{n+1} = u_n - \frac{F(u_n)}{\lambda_n} \qquad \lambda_{n+1} = F'(u_{n+1})$$

Autrement dit,

$$(1) \qquad u_{n+1} = u_n - \frac{F(u_n)}{F'(u_n)}.$$

On retombe ainsi sur la méthode de Newton.

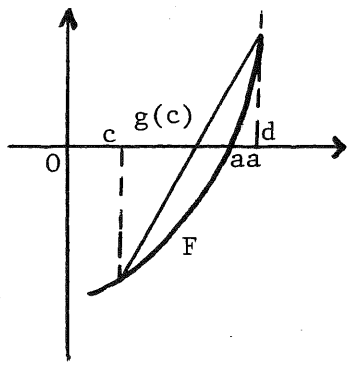
c) Interpolation linéaire

L'idée consiste à remplacer la fonction F par une fonction linéaire affine interpolant F sur un intervalle [c,d]. où c et d désignent des valeurs approchées de a telles que $c < a < d$. Dans ces conditions

$$g(x) = x - \frac{x-d}{F(x) - F(d)} F(x).$$

Interprétation géométrique de la méthode d'interpolation linéaire.

Supposons que l'on connaisse un encadrement [c,d] de la solution a. Alors g(c) n'est autre que l'abscisse du point d'intersection avec 0x de la sécante relative aux points c et d. Autrement dit, on effectue une *interpola-*



tion linéaire de F sur l'intervalle [c,d], et on remplace l'équation non linéaire F(x) = 0 par l'équation *linéaire* approchée.

$$x - \frac{c - d}{F(c) - F(d)} F(c) = 0.$$

Cette méthode porte des noms variés : interpolation linéaire, sécantes, fausse position, Lagrange, Descartes...

La convergence est géométrique ; l'étude théorique de cette convergence est esquissée dans Analyse I [5].

d) Passage aux fonctions réciproques

Considérons une équation f(x) = x et supposons que pour tout point x de I, $|f'(x)| > 1$. Alors, comme f' ne s'annule pas, f est strictement monotone sur I et admet donc une fonction réciproque g définie sur I' = f(I). En outre g est de classe C^1 sur I' et, pour tout point y de I'.

$$|g'(y)| = \frac{1}{|f'(x)|} \quad \text{où } y = f(x).$$

Donc $|g'(y)| < 1$, et la méthode des approximations successives s'applique à g.

Remarque. Bien entendu cette méthode est déconseillée si les fonctions réciproques sont compliquées à calculer, ou si les rapports de lipschitz sont trop voisins de 1.

e) Cas des racines multiples.

Les méthodes précédentes échouent si F'(a) = 0 (ou si F'(a) est trop voisin de 0). Il convient alors de transformer préalablement l'équation F(x) = 0 de telle sorte que la racine a devienne une racine simple.

En analyse numérique, cette difficulté survient dès que deux racines sont mal séparées. Pour l'étude de tels cas on pourra se reporter à Durand, Chapitre 2 [6], et Berezin et Zhydkov, Chapitre 7 [8].

f) Quelques exemples illustrant ces méthodes et permettant de comparer expérimentalement leur performance.

Exemple 1. Résoudre l'équation numérique tgx = x, où

$$x \in]\pi, \frac{3\pi}{2}[,$$

à la précision 10^{-8}.

a) Poser tgx = y (Méthode des fonctions réciproques)
b) Traiter ensuite les cas où $x \in]n\pi, n\pi + \pi/2[$, lorsque n = 2,3,4.
c) Utiliser la méthode de Newton pour l'équation tgx = x, puis pour l'équation y = Arctgy + nπ.

Exemple 2. Résoudre l'équation numérique $e^{5x} = x + 100$, où $x \geqslant 0$, à la précision 10^{-8}. Il est clair que $a \in [0,1]$.

a) Transformer cette équation en

$$y = \frac{1}{5} \text{Log}(y + 100).$$

b) Utiliser la méthode de Newton.
c) Comparer la performance de ces deux méthodes.

Exemple 3. Résoudre l'équation numérique shx = 100x, à la précision 10^{-8}.

a) Poser y = shx.
b) Poser $y = e^x$.
c) Utiliser un ajustement linéaire.
d) Comparer la performance de ces méthodes.

Exemple 4. Résoudre l'équation numérique $x = 30 \text{Log}x + 100$ à la précision 10^{-8}.

a) Déterminer d'abord la partie entière de la solution a.
b) Effectuer un ajustement linéaire.
c) Employer la méthode de Newton.
d) Comparer la performance des procédés b) et c).

Exemple 5. Résoudre l'équation numérique $x^3 + x - 1 = 0$ à la précision 10^{-8}.

a) Prouver que $a \in [2/3, 3/4]$.
b) Effectuer un ajustement linéaire en prenant $\lambda = F'(2/3)$.
c) Employer la méthode de Newton.
d) Comparer.

Exemple 6. Résoudre l'équation $x = \text{Log}x + 2$, $x \geqslant 1$, à la précision 10^{-8}.

a) Combien faudrait-il d'itérations par l'algorithme $u_{n+1} = \text{Log}\, u_n + 2$, $u_0 = 3$?
b) Effectuer un ajustement linéaire en prenant $\lambda = F'(3) = 2/3$, puis $\lambda = F'(u_1)$.
c) Employer la méthode de Newton.
d) Comparer la performance des méthodes b) et c).

Remarque. Dans les exemples précédents, on a choisi la précision 10^{-8} pour pouvoir apprécier plus clairement les performances des méthodes utilisées, ce que ne permettrait pas une précision moindre.

Pour le cas classique de calcul des racines carrées, des racines $n^{\text{èmes}}$, et des inverses, se reporter à la bibliographie sectorielle du thème 2 et, en particulier à Analyse 1.

6) Conclusion. L'étude précédente montre combien cette méthode du point fixe est au cœur des concepts fondamentaux de l'analyse, et ceci à tous les niveaux, puisque cette même méthode permet, ultérieurement, d'attaquer avec succès les équations matricielles, différentielles, intégrales, implicites, ...

Pour les notes historiques, nous renvoyons à la bibliographie sectorielle du thème 2. "Equations numériques".

BIBLIOGRAPHIE

On pourra aussi se reporter à la bibliographie sectorielle du thème 2 "Equations numériques".

OUVRAGES SCIENTIFIQUES

[1] DEMIDOVITCH B. et MARON I. - *Eléments de calcul numérique.* (Mir)

[2] ENGEL A. - *Mathématique élémentaire d'un point de vue algorithmique*, adapté par Daniel Reisz. (Cedic).

[3] HILDEBRAND F.B. - *Introduction to numerical analysis.* (Mc Graw-Hill).

[4] OVAERT J.L. et VERLEY J.L. - *Exercices de Mathématiques, Analyse 1.* (Cedic) (en préparation)

[5] IREM de MARSEILLE - *Brochure Analyse I.*

[6] BEREZIN I.S. et ZHYDKOV N.P. - *Computing methods.* (Pergamon Press).

[7] DIEUDONNE J. - *Calcul infinitésimal.* (Hermann).

[8] DURAND E. - *Solutions numériques des équations algébriques.* (Masson)

[9] HOUSEHOLDER A. - *Principles of numerical analysis.* (Mac Graw Hill).

MONOGRAPHIES ET CONTEXTE

[10] ADAMS J.F. - *Algebraic topology. A student's guide,* Cambridge University press, New-York and London, 1972.

[11] ARNOLD V. - *Equations différentielles ordinaires.* (Mir).

[12] DIEUDONNE J. - *Eléments d'Analyse (vol. 1).* (Gauthier-Villars).

[13] GREENBERG M.J. - *Lectures on algebraic topology,* (Benjamin, Reading, Mass.), 1967.

[14] HARDY G.H. et WRIGHT E.M. - *An introduction to the theory of numbers.* (Clarendon Press). Oxford.

[15] HIRSCH M.W. et SMALE S. - *Differential equations, dynamical systems and linear algebra.* (Academic Press).

[16] KUROSH, *Cours d'algèbre supérieure,* (Mir), Moscou, 1973.

[17] STEWART G.W. - *Introduction to matrix computation.* (Academic Press).

[18] STEWART I. - *Galois theory,* Chapman and Hall, 1973.

[19] VARGA R.F. - *Matrix iterative analysis.* (Englewood cliffs).

ENCYCLOPEDIES, REVUES, HISTOIRE

[20] *Encyclopedic Dictionnary of Mathematics* (EDM) MIT Press. Cambridge (Mass) and London. 1977.

[21] *Encyclopédie des Sciences Mathématiques pures et appliquées.* Paris Leipzig. 1904-1914.

[22] *Encyclopaedia Universalis.* Paris, 1968.

[23] *Petit Archimède : numéro spécial sur* Π.

[24] ABDELJAOUAD M. - *Vers une épistémologie des décimaux* (in Fragments d'histoire des mathématiques - Brochure APMEP).

[25] BOYER C. - *A History of mathematics.* (J. Wiley and sons).

[26] DIEUDONNE J. et alii : *Abrégé d'histoire des mathématiques* (Hermann)

[27] DHOMBRES J. - *Nombre, mesure et continu.* (Cedic).

[28] DUGAC P. - *Cours d'histoire des mathématiques.* (Paris VII).

[29] DUGAC P. - *Eléments d'analyse de K. Weierstrass.* (Archive for history). (Volume 10).

[30] ELLISON W. et F. - *Théorie des nombres* in Abrégé d'histoire des mathématiques (Hermann).

[31] GOLDSTINE H. - *A history of numerical analysis.* (Springer).

[32] KLINE M. - *Mathematical thought from ancient to modern times.* New-York. (Oxford University Press).

[33] OVAERT J.L. - Article *Calcul numérique* in Encyclopaedia Universalis.

TEXTES HISTORIQUES

[34] BOLZANO B. - *(Une preuve analytique...), (1817),* Revue d'histoire des Sciences, tome 17.

[35] CAUCHY A.L. - *Cours d'Analyse à l'Ecole Polytechnique. Analyse algébrique. (1821).* (Gauthier-Villars et Diffusion IREM).

[36] EULER L. - *Eléments d'algèbre.* (Diffusion IREM).

[37] EULER L. - *Introduction à l'analyse des infiniments petits. (1748).* (Diffusion IREM).

[38] GALOIS E. - *Sur les conditions de résolubilité des équations par radicaux.* Oeuvres complètes, (Gauthier-Villars).

[39] GAUSS K.F. - *Recherches arithmétiques (1801),* (Blanchard).

[40] LAGRANGE J.L. - *Théorie des fonctions analytiques, (1797)* (Diffusion IREM).

[41] LAGRANGE J.L. - *Leçons sur le calcul des fonctions. (1808).*

[42] LAGRANGE J.L. - *Réflexions sur la résolution algébrique des équations (1771)* Oeuvres complètes, (Gauthiers-Villars).

[43] LAGRANGE J.L. - *Sur l'intégration d'une équation... (1759)* Oeuvres, vol. 1, pp. 23-36.

[44] LAGRANGE J.L. - *Recherches sur les suites récurrentes... (1775)* Oeuvres, vol. 4 pp. 151-251 et *(1793)* vol. 5 pp. 627-641.

[45] LAPLACE P.S. - *Théorie analytique des probabilités,* Oeuvres, tome 7, Gauthier-Villars, Paris.

[46] NEWTON I. - *Méthodes des fluxions et des suites infinies. (1736)* (Blanchard).

[47] SERRET J.A. - *Cours d'algèbre supérieure.* 5e édition. Paris 1885. (Gauthier-Villars).