# Résolution numérique d'équations non linéaires

On étudie dans ce chapitre des méthodes de résolution de systèmes d'équations non linéaires dans $\mathbb{R}^n$.

Exemple: résolution numérique de l'équation de Poisson – Boltzmann

Ce modèle décrit le potentiel électrique $u(r)$ autour d'un cylindre chargé en surface et plongé dans une solution ionique.
Le modèle adimensionné s'écrit:

$$\begin{cases} u'' + \dfrac{1}{r}\dfrac{du}{dr} = K^2 \, sh(u), & r \in \,]1, 1+\ell[ \\ u'(1) = -\mu, \quad u(1+\ell) = 0 & (\ell \gg K^{-1}) \end{cases}$$

On discrétise le problème par un schéma différences finies d'ordre 2:

$$\begin{cases} \dfrac{u_{i+1} - 2u_i + u_{i-1}}{h^2} + \dfrac{1}{r_i}\dfrac{u_{i+1} - u_{i-1}}{2h} = K^2 sh(u_i), & 1 \leq i \leq n-1 \\ \dfrac{u_1 - u_0}{h} = -\mu\left(1 - \dfrac{h}{2}\right) + \dfrac{h}{2} K^2 sh\, u_0, & u_n = 0 \end{cases}$$

avec $r_i = 1 + ih \quad (0 \leq i \leq n)$, $h = \dfrac{\ell}{n}$.

Cela donne un système d'équations non linéaires pour $x = (u_0, u_1, \dots, u_{n-1})^T \in \mathbb{R}^n$, qui peut s'écrire sous la forme:

$$A x = G(x)$$

où $A \in M_n(\mathbb{R})$ est inversible (et tridiagonale) et $G = (G_0, \dots, G_{n-1})^T$ défini par $G_i = h^2(sh\, u_i - u_i)$ $\forall i \geq 1$, $G_0 = h^2(sh\, u_0 - u_0) + \mu h(h-2)$

Le problème à résoudre revient donc à chercher un point fixe de l'application $\phi = A^{-1}G$:

$$x = \phi(x).$$

Pour calculer un point fixe d'une application $\phi : \mathbb{R}^n \to \mathbb{R}^n$ continue, la "méthode des approximations successives" considère une suite $x_{k+1} = \phi(x_k)$ ($x_0 \in \mathbb{R}^n$) qui en cas de convergence tend vers un point fixe de $\phi$ : si $x_k \xrightarrow[k\to+\infty]{} x$ alors $x = \phi(x)$.

Dans l'exemple de l'équation de Poisson-Boltzmann, cette méthode itérative s'écrit $A x_{k+1} = G(x_k)$ (avec $A = LU$).

Nous allons étudier la convergence de cette méthode plus en détail.

## 1) Méthode des approximations successives :

**Théorème** (point fixe contractant dans $(\mathbb{R}^n, \| \ \|)$).

Soit $F$ une partie fermée non vide de $\mathbb{R}^n$ et $\phi : F \to F$ une application contractante : $\exists K < 1 \ / \ \|\phi(x) - \phi(y)\| \leq K \|x-y\| \ \forall x, y \in F$

Alors l'équation :

$$x = \phi(x), \quad x \in F$$

admet une solution unique. De plus, $\forall x_0 \in F$ la suite définie par $x_{k+1} = \phi(x_k)$ converge vers $x$, avec : $\forall k \geq 1$ :

$$\|x - x_k\| \leq \frac{K}{1-K} \|x_k - x_{k-1}\| \leq \frac{K^k}{1-K} \|x_1 - x_0\|$$

preuve des bornes d'erreur :

$$\begin{cases} x = \phi(x) \\ x_k = \phi(x_k) + x_k - x_{k+1} \end{cases}$$ donne en soustrayant : $$x - x_k = \phi(x) - \phi(x_k) + x_{k+1} - x_k$$

Donc $\|x - x_k\| \leq \|\phi(x) - \phi(x_k)\| + \|\phi(x_k) - \phi(x_{k-1})\| \leq K\left(\|x - x_k\| + \|x_k - x_{k-1}\|\right)$

d'où $$\|x - x_k\| \leq \frac{K}{1-K} \|x_k - x_{k-1}\| = \frac{K}{1-K} \|\phi(x_{k-1}) - \phi(x_{k-2})\|$$

$$\leq \frac{K^2}{1-K} \|x_{k-1} - x_{k-2}\| \leq \dots \leq \frac{K^k}{1-K} \|x_1 - x_0\|$$

On peut notamment utiliser le critère d'arrêt

$\| x_k - x_{k-1} \| \leq \left(\frac{1}{K} - 1\right) \varepsilon$ qui garantit $\| x - x_k \| \leq \varepsilon$ ($=$ tolérance d'erreur)

La convergence de la méthode des approximations successives est linéaire : $\frac{\| x_{k+1} - x \|}{\| x_k - x \|} \leq K < 1$ $\quad ( x_{k+1} - x = \phi(x_k) - \phi(x) )$

et lente lorsque $K$ est proche de $1$.

<u>application :</u> schémas de Runge-Kutta implicites à $s$ étages :

$$\begin{cases} y_{k+1} = y_k + h \sum_{i=1}^{s} b_i k_i & y_k \in \mathbb{R}^n \\ k_i = f(t_k + c_i h,\ y_k + h \sum_{j=1}^{s} a_{ij} k_j), & \forall i = 1, \dots, s \end{cases} \qquad (S)$$

pour une équation différentielle $y' = f(t, y)$ dont le second membre $f : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$ est $L$-lipschitzien $/ y$. On suppose que $h < \frac{1}{L \||A|\|_\infty}$,

où $A = (a_{ij})_{1 \leq i, j \leq s}$ et $\||A|\|_\infty = \max_i \sum_j |a_{ij}|$.

Alors le système $(S)$ admet une solution $y_{k+1}$ unique.

<u>preuve</u> Notons $x = (k_1, k_2, \dots, k_s) \in (\mathbb{R}^n)^s$, $\phi(x) = (\phi_1(x), \dots, \phi_s(x))$

avec $\phi_i(x) = f(t_k + c_i h,\ y_k + h \sum_{j=1}^{s} a_{ij} k_j) \in \mathbb{R}^n$.

Alors $\forall x, x' \in (\mathbb{R}^n)^s$, $\forall i = 1 \dots s$ :

$$\| \phi_i(x) - \phi_i(x') \| \leq L h \left\| \sum_{j=1}^{s} a_{ij} (k_j - k_j') \right\| \leq L h \sum_{j=1}^{s} |a_{ij}| \, \| x - x' \|_\infty$$

où $\forall x = (k_1, \dots, k_s) \in (\mathbb{R}^n)^s$, $\| x \|_\infty := \max_{1 \leq i \leq s} \| k_i \|$

En prenant le max sur $i = 1 \dots s$ : $\| \phi(x) - \phi(x') \|_\infty \leq L h \||A|\|_\infty \| x - x' \|_\infty$

Donc pour $h < (L \||A|\|_\infty)^{-1}$, $\phi$ est contractante sur $((\mathbb{R}^n)^s, \| \ \|_\infty)$, donc admet un unique point fixe $x = (k_1, \dots, k_s)$, qui détermine $y_{k+1}$ à partir de la $1^{\text{ère}}$ équation de $(S)$.

Pour montrer qu'une application $\phi \in C^1$ est lipschitzienne / contractante, on utilise souvent l'inégalité des accroissements finis :

$$\| \phi(x) - \phi(y) \| \leq \sup_{[x,y]} ||| D\phi ||| \; \| x-y \|$$

où $[x,y] = \{ tx + (1-t)y, \; t \in [0,1] \}$ et $||| \; |||$ est la norme matricielle subordonnée à $\| \; \|$.

En particulier, si $\phi \in C^1(\mathbb{R}^n, \mathbb{R}^n)$ vérifie $\sup_F ||| D\phi ||| \leq K$ sur une partie $F$ convexe ($\forall x,y \in F$, $[x,y] \subset F$) alors

$\| \phi(x) - \phi(y) \| \leq K \| x-y \| \quad \forall \, x,y \in F$.

Nous allons utiliser cela pour obtenir un résultat de convergence local pour la méthode des approximations successives.

<u>Proposition</u>

Soit $\phi \in C^1(\mathbb{R}^n, \mathbb{R}^n)$ admettant un point fixe $a \in \mathbb{R}^n$ tel que $\rho(D\phi(a)) < 1$. On considère une norme $\| \; \|$ sur $\mathbb{R}^n$ qui induit $||| D\phi(a) ||| < 1$. Alors il existe $\eta_0 > 0$ tel que $\forall \eta \in [0, \eta_0[$, $\phi$ est une contraction sur la boule fermée $B_f(a, \eta) = \{ x \in \mathbb{R}^n, \; \| x-a \| \leq \eta \}$.

<u>preuve</u> $\forall M \in M_n(\mathbb{C})$, $\forall \varepsilon > 0$ il existe une norme matricielle subordonnée / $||| M ||| \leq \rho(M) + \varepsilon$.

Puisque $\rho(D\phi(a)) < 1$, il existe donc une norme subordonnée telle que $||| D\phi(a) ||| < 1$ ; on note $\| \; \|$ la norme associée sur $\mathbb{R}^n$.

Par continuité de $x \mapsto ||| D\phi(x) |||$ on a donc $||| D\phi ||| < 1$ sur $B_f(a, \eta)$ lorsque $\eta$ est choisi assez petit.

Soit $\eta_0 = \sup \{ \eta \geq 0, \sup_{B_f(a,\eta)} |||D\phi||| < 1 \}$.

$B_f(a,\eta)$ étant convexe, l'inégalité des accroissements finis implique que $\forall \eta \in [0, \eta_0[$, $\forall x, y \in B_f(a,\eta)$,

$\|\phi(x) - \phi(y)\| \leq K_\eta \|x-y\|$, $K_\eta = \sup_{B_f(a,\eta)} |||D\phi||| < 1$.

De plus, $\forall x \in B_f(a,\eta)$, $\|\phi(x) - a\| = \|\phi(x) - \phi(a)\| \leq K_\eta \|x-a\| < \eta$

donc $\phi(x) \in B(a,\eta) \subset B_f(a,\eta)$. Donc $\phi$ est une contraction sur $B_f(a,\eta)$ $\forall \eta < \eta_0$. $\square$

On en déduit le résultat suivant :

<u>Théorème</u> (convergence locale)

Soit $\phi \in C^1(\mathbb{R}^n, \mathbb{R}^n)$ admettant un point fixe $a \in \mathbb{R}^n$ tel que $\rho(D\phi(a)) < 1$. Alors il existe un voisinage $\Omega$ de $a$ tel que $\phi$ admet $a$ pour unique point fixe dans $\Omega$, et $\forall x_0 \in \Omega$ la suite $(x_k)_{k \geq 0}$ définie par $x_{k+1} = \phi(x_k)$ converge vers $a$.

<u>preuve</u>

On reprend les notations de la proposition précédente et on pose $\Omega = B(a, \eta_0) = \{ x \in \mathbb{R}^n, \|x-a\| < \eta_0 \}$.

Si $x \in \Omega$ et $\phi(x) = x$, alors $x = a$ car $\phi$ admet $a$ comme unique point fixe dans $B_f(a, \|x-a\|)$ où $\phi$ est contractante.

Si $x_0 \in \Omega$, $\phi$ étant une contraction sur $B_f(a, \|x_0 - a\|)$, alors $x_k \xrightarrow[k \to +\infty]{} a$ d'après le théorème du point fixe contractant $\square$

<u>Remarque</u> On retrouve la condition de convergence $\rho < 1$ des méthodes itératives linéaires, mais ici la convergence est locale (pour $x_0 \in \Omega$). De plus, il peut arriver que $\rho(D\phi(a)) = 1$ et la méthode converge.

Il est classique de ramener la résolution d'une équation non linéaire à la recherche d'un point fixe (comme nous l'avons vu pour l'équation de Poisson-Boltzmann).

Etant donné $f : \mathbb{R}^n \to \mathbb{R}^n$ et $A \in M_n(\mathbb{R})$ inversible,

$$f(x) = 0 \iff Ax = Ax - f(x)$$
$$\iff x = x - A^{-1} f(x) := \phi(x).$$

La méthode des approximations successives s'écrit : $x_{k+1} = \phi(x_k)$, ou

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

Supposons $f \in C^1(\mathbb{R}^n, \mathbb{R}^n)$, $f(a) = 0$ et $Df(a)$ inversible.

Supposons connue une approximation $\tilde{a}$ de $a$ (obtenue par exemple à partir d'une équation linéarisée, de l'itéré précédent si on intègre une éq. différentielle, d'une approximation analytique...)

Si on choisit $A = Df(\tilde{a})$, alors au point fixe $a$ de $\phi$ :

$$D\phi(a) = I - Df(\tilde{a})^{-1} Df(a) \qquad (Df(\tilde{a}) \text{ inversible si } \tilde{a} \simeq a)$$

Puisque $\lim_{\tilde{a} \to a} D\phi(a) = 0$, on a $\rho(D\phi(a)) < 1$ pour $\tilde{a}$ suffisament proche de $a$. D'après le théorème de convergence locale, la méthode des approximations successives est alors convergente sur un voisinage de $a$.

<u>exemple</u> calcul de $\sqrt{c}$ $(c > 0)$ en résolvant $f(x) = x^2 - c = 0$

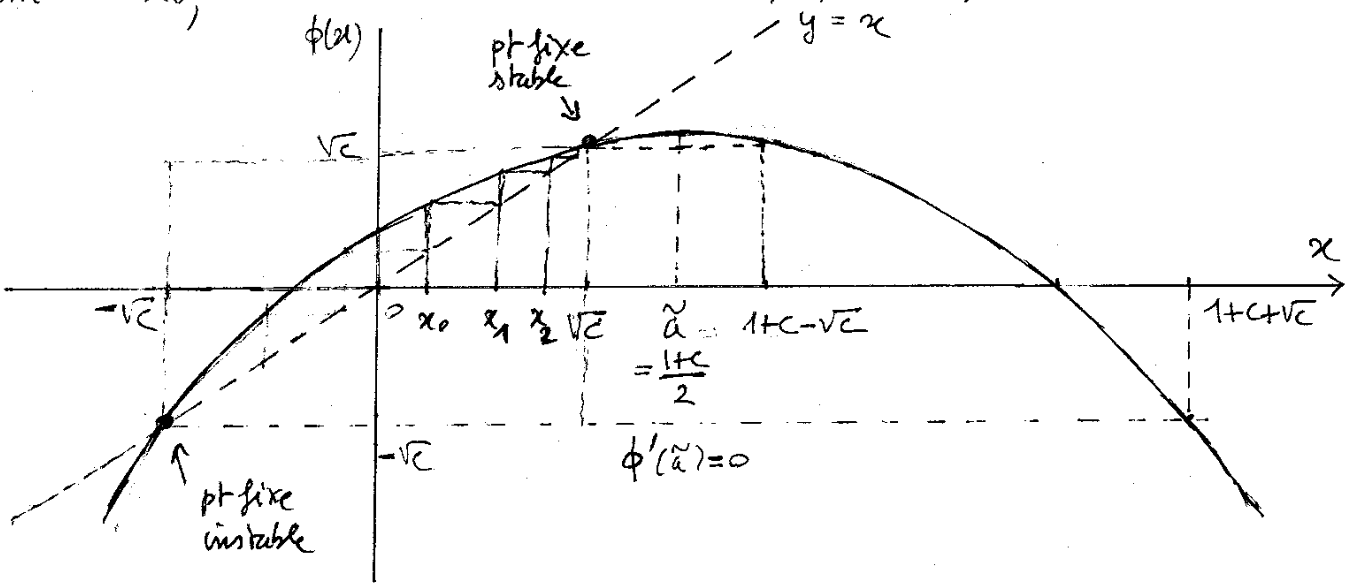
Approximation (précise lorsque $c \simeq 1$) : $\sqrt{c} = \sqrt{1 + (c-1)} \simeq \frac{1+c}{2} = \tilde{a}$

$$\phi(x) = x - \frac{f(x)}{f'(\tilde{a})} = x - \frac{x^2 - c}{1+c}$$

La méthode des approximations successives s'écrit :

$$x_{k+1} = \phi(x_k) \iff x_{k+1} = x_k - \frac{x_k^2 - c}{1+c}$$

Pour étudier le comportement de $(x_k)_{k\geq 0}$ suivant la condition initiale $x_0$, il est utile de tracer le graphe de $\phi$ (une parabole) :




$\phi$ admet deux points fixes : $x = \pm\sqrt{c}$

$\phi'(\sqrt{c}) = 1 - \frac{2\sqrt{c}}{1+c} = \frac{(1-\sqrt{c})^2}{1+c} \in [0,1[$ donc par le Thm de cv locale,

$x_k \xrightarrow[k\to+\infty]{} \sqrt{c}$ si $x_0$ est assez proche de $\sqrt{c}$ (point fixe "stable").

Plus précisément, $x_k \to \sqrt{c}$ lorsque $x_0 \in ]-\sqrt{c}, 1+c+\sqrt{c}[$, en particulier pour $x_0 = \tilde{a} = \frac{1+c}{2}$. En effet :

- si $x_0 \in ]-\sqrt{c}, \sqrt{c}] = I$, alors $x_k \in I$ $\forall k \geq 0$ ($\phi(I) = I$), $(x_k)_{k\geq 0}$ est croissante ($\phi(x) \geq x$ sur $I$) donc converge vers $\sqrt{c}$.
- si $x_0 \in [1+c-\sqrt{c}, 1+c+\sqrt{c}[$ alors $x_1 \in I$, $x_k \xrightarrow[k\to+\infty]{} \sqrt{c}$
- si $x_0 \in ]\sqrt{c}, 1+c-\sqrt{c}]$ alors $x_k \xrightarrow[k\to+\infty]{} \sqrt{c}$ en décroissant
- si $x_0 < -\sqrt{c}$ ou $x_0 > 1+c+\sqrt{c}$ alors $x_1 < -\sqrt{c}$, $(x_k)$ est décroissante ($\phi(]-\infty,-\sqrt{c}[) = ]-\infty,-\sqrt{c}[$, $\phi(x) < x$ sur cet intervalle) et $x_k \to -\infty$ ($x_k$ non minoré en l'absence de point fixe $< -\sqrt{c}$)

## 2) Méthode de Newton

Soit $f \in C^2(\mathbb{R}^n, \mathbb{R}^n)$. On suppose qu'il existe $a \in \mathbb{R}^n$ / $f(a)=0$ et $Df(a)$ inversible.

La méthode de Newton est une méthode itérative convergeant localement vers $a$ plus rapidement qu'avec la méthode des approximations successives décrite p.6. (Convergence quadratique au lieu de la convergence linéaire)

principe : on part d'une approximation $x_0$ de $a$ puis on calcule $x_1, x_2, x_3, \ldots$ de la manière suivante :

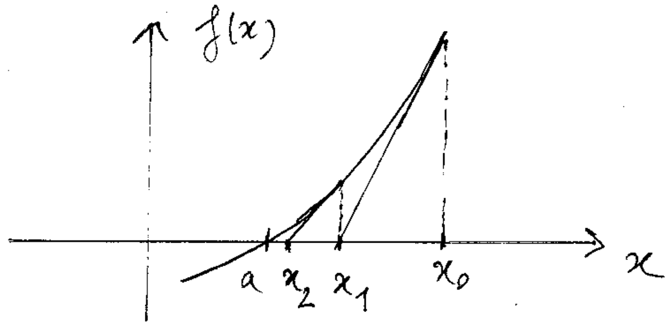
lorsque $x_k \simeq x$, $f(x) = f(x_k + x - x_k) \simeq f(x_k) + Df(x_k)(x-x_k)$

$x_k$ étant connu, on détermine l'approximation $x_{k+1}$ de $a$ (solution de $f(x)=0$) en résolvant : $f(x_k) + Df(x_k)(x-x_k) = 0$ (approximation affine de $f$ au voisinage de $x_k$).

On obtient alors :

$$\left\| \begin{array}{l} Df(x_k)\,(x_{k+1}-x_k) = -f(x_k) \qquad (N) \\ x_{k+1} = x_k - Df(x_k)^{-1} f(x_k) := \phi(x_k) \end{array} \right.$$

Illustration en dimension 1 :




algorithme : $x_k$ et $f(x_k)$ étant connus :

- calculer $Df(x_k)$
- résoudre $Df(x_k)\, d_k = -f(x_k)$ (ne pas calculer $Df(x_k)^{-1}$)
- $x_{k+1} = x_k + d_k$
- itérer si $\|f(x_{k+1})\|$ ou $\|d_k\|$ dépassent une tolérance d'erreur.

Nous allons montrer la convergence locale de la méthode de Newton en utilisant les résultats de la partie 1)

<u>Proposition</u> : Si $f$ est $C^2$ au voisinage de $a$ où $f(a)=0$, $Df(a)$ inversible, alors $\phi(x) = x - Df(x)^{-1} f(x)$ est $C^1$ au voisinage de $a$ avec $\phi(a) = a$, $D\phi(a) = 0$, $\phi(a+h) = a + O(\|h\|^2)$ lorsque $h \to 0$ dans $\mathbb{R}^n$.

<u>preuve</u> $x \mapsto Df(x)$ est $C^1$ au voisinage de $a$ avec $Df(a)$ inversible, et l'application $GL_n(\mathbb{R}) \to GL_n(\mathbb{R})$, $A \mapsto A^{-1}$ est $C^\infty$, donc $x \mapsto Df(x)^{-1}$ est $C^1$ au voisinage de $a$, et par suite $\phi$ est $C^1$ au voisinage de $a$.

Quand $h \to 0$ dans $\mathbb{R}^n$, $Df(a+h)^{-1} = Df(a)^{-1} + O(\|h\|)$ et $f(a+h) = Df(a)\, h + O(\|h\|^2)$, donc

$$\begin{aligned}
\phi(a+h) &= a+h - Df(a+h)^{-1} f(a+h) \\
&= a+h - \left[ Df(a)^{-1} + O(\|h\|) \right] \left( Df(a)\, h + O(\|h\|^2) \right) \\
&= a+h - h + O(\|h\|^2) \\
&= a + O(\|h\|^2)
\end{aligned}$$

On a donc $\phi(a+h) - \phi(a) = O(\|h\|^2)$, d'où $D\phi(a) = 0$.

□

On en déduit le résultat suivant ($\| \ \|$ désigne une norme sur $\mathbb{R}^n$)

<u>Théorème</u> : convergence locale de la méthode de Newton (N).

Si $f$ est $C^2$ au voisinage de $a$ où $f(a)=0$, $Df(a)$ inversible, alors il existe un voisinage $V$ de $a$ et $c \geq 0$ tels que $f$ admet $x=a$ pour unique zéro dans $V$, et $\forall x_0 \in V$ :

i) $\lim_{k \to +\infty} x_k = a$

ii) $\|x_{k+1} - a\| \leq c \|x_k - a\|^2 \quad \forall k \geq 0$ ("convergence quadratique") preuve : $f(x)=0 \iff \phi(x)=x$ au voisinage de $x=a$ ($Df(a)$ inversible)

$\phi(a)=a$, $D\phi(a)=0$ donc $\rho(D\phi(a))=0<1$, donc (application du théorème donné p-5) $\phi$ admet $a$ pour unique point fixe dans un voisinage de $a$, et $x_k \xrightarrow[k\to+\infty]{} a$ pour $x_0$ assez proche de $a$.

Il reste maintenant à montrer ii).

Puisque $\phi(a+h)-a = O(\|h\|^2)$ :

$\exists \eta_1>0, c\geq 0 \;/\; (\|h\|<\eta_1) \Rightarrow (\|\phi(a+h)-a\| \leq c\|h\|^2)$

d'où pour $h=x_k-a$ : $(\phi(a+h)=x_{k+1})$

$$(\|x_k-a\|<\eta_1) \Rightarrow (\|x_{k+1}-a\| \leq c\|x_k-a\|^2)$$

D'après la proposition p4, puisque $\rho(D\phi(a))=0<1$,

$\exists B_g(a,\eta) \subset B=\{x, \|x-a\|<\eta_1\}$ tel que $\phi(B_g(a,\eta)) \subset B_g(a,\eta) \subset B$.

donc si $x_0 \in B_g(a,\eta)$ alors $\|x_k-a\|<\eta_1 \;\forall k\geq 0$,

d'où $\|x_{k+1}-a\| \leq c\|x_k-a\|^2 \;\forall k\geq 0$.

On a donc montré le théorème avec $V=B_g(a,\eta)$ (voisinage de $a$ sur lequel $\phi$ est contractante). □

La convergence quadratique ii) donne pour $\tilde{e}_k = c(x_k-a)$ :

$$\|\tilde{e}_k\| \leq c^2\|x_{k-1}-a\|^2 = \|\tilde{e}_{k-1}\|^2$$
$$\leq \|\tilde{e}_{k-2}\|^4 \leq \dots \leq \|\tilde{e}_0\|^{2^k}$$

d'où $\|x_k-a\| \leq \frac{1}{c}(c\|x_0-a\|)^{2^k}$.

Donc $\|x_k-a\| \xrightarrow[k\to+\infty]{} 0$ à vitesse doublement exponentielle si $\|x_0-a\|<\frac{1}{c}$.

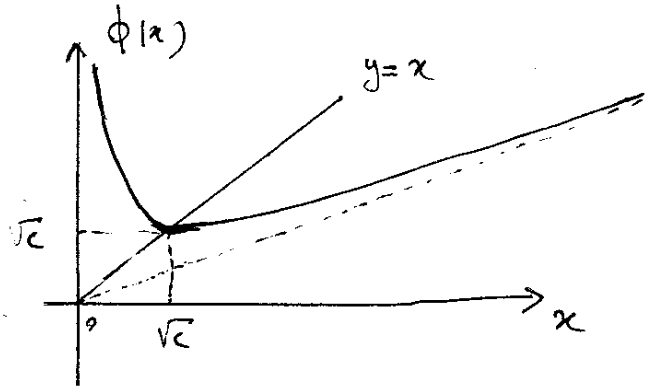
(donc bien plus rapidement que la vitesse exponentielle qui caractérise la convergence linéaire).

## Calcul de $\sqrt{c}$ par la méthode de Newton, $c$ constante $>0$ :

On veut annuler $f(x) = x^2 - c \Rightarrow f'(x) = 2x$ et la méthode de Newton s'écrit :

$$x_{k+1} = x_k - \frac{1}{2x_k}\left(x_k^2 - c\right) = \frac{1}{2}\left(x_k + \frac{c}{x_k}\right) := \phi(x_k)$$

On a $\phi(\sqrt{c}) = \sqrt{c}$ et $\phi'(\sqrt{c}) = 0$.




On étudie la convergence globale de la méthode pour toute condition initiale $x_0 > 0$. On commence par le cas $x_0 \geq \sqrt{c}$.

$\phi([\sqrt{c}, +\infty[) = [\sqrt{c}, +\infty[$ donc si $x_0 \in [\sqrt{c}, +\infty[$ alors $x_k \in [\sqrt{c}, +\infty[\ \forall k \geq 0$.

Par ailleurs $\phi(x) \leq x$ sur cet intervalle, donc $x_{k+1} = \phi(x_k) \leq x_k$ dans cet intervalle, ie la suite $(x_k)_{k \geq 0}$ est décroissante si $x_0 \in [\sqrt{c}, +\infty[$.

Cette suite étant décroissante et minorée, elle converge vers $\sqrt{c}$ qui est le seul point fixe de $\phi$ dans $[\sqrt{c}, +\infty[$.

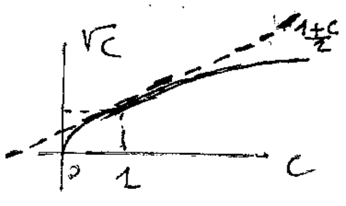
Par ailleurs, si $x_0 \in ]0, \sqrt{c}[$ alors $x_1 = \phi(x_0) \in ]\sqrt{c}, +\infty[$ et on est ramené au cas précédent.

Donc $\forall x_0 \in ]0, +\infty[$, la suite $(x_k)_{k \geq 0}$ converge vers $\sqrt{c}$.

$x = (0,1\dots\dots)_2 \cdot 2^p = c \times 2^p$

En utilisant la représentation des réels en base 2, on peut se ramener au cas $c \in [\frac{1}{2}, 1[$ : $\sqrt{c \times 2^{2m}} = \sqrt{c}\, 2^m$, $\sqrt{c \times 2^{2m-1}} = \sqrt{c}\sqrt{1/2}\, 2^m$.

De plus $\sqrt{c} = \sqrt{1 + c - 1} = 1 + \frac{c-1}{2} + O(c-1)^2 \simeq \frac{1+c}{2}$ si $c \simeq 1$




Par exemple $\sqrt{0,5} \simeq 0,707$ et $\frac{1+0,5}{2} = 0,75$

On peut donc choisir $x_0 = \frac{1+c}{2}$ lorsque $c \in [\frac{1}{2}, 1[$.

Majoration de l'erreur :

$$x_{k+1} = \frac{1}{2}\left(x_k + \frac{c}{x_k}\right).$$

L'erreur $e_k = x_k - \sqrt{c}$ vérifie :

$$e_{k+1} = \frac{1}{2}\left(x_k + \frac{c}{x_k}\right) - \sqrt{c} = \frac{1}{2}\left(e_k + \frac{c}{e_k + \sqrt{c}} - \sqrt{c}\right)$$

$$= \frac{1}{2}\left(e_k - \frac{\sqrt{c}\, e_k}{e_k + \sqrt{c}}\right) = \frac{e_k}{2}\left(1 - \frac{1}{\frac{e_k}{\sqrt{c}} + 1}\right)$$

$$\boxed{e_{k+1} = \frac{e_k^2}{2\sqrt{c}} \times \frac{1}{1 + \frac{e_k}{\sqrt{c}}}}$$

On pose $\varepsilon_k = \frac{e_k}{2\sqrt{c}}$. Alors : $\varepsilon_{k+1} = \varepsilon_k^2 \times \frac{1}{1 + 2\varepsilon_k}$

Pour $\boxed{x_0 = \frac{1+c}{2} \geqslant \sqrt{c}}$, on a $e_0 \geqslant 0$ et $\varepsilon_0 \geqslant 0$.

Donc $\varepsilon_k \geqslant 0 \ \forall k \geqslant 0$ par récurrence, et $\varepsilon_{k+1} \leq \varepsilon_k^2$,

d'où $\varepsilon_k \leq \varepsilon_0^{2^k}$ par récurrence.

Pour avoir $\frac{e_k}{\sqrt{c}} = 2\varepsilon_k \leq \eta$, il suffit d'avoir $2\varepsilon_0^{2^k} \leq \eta$

ce qui équivaut à $2^k \geqslant \frac{|\ln(\eta/2)|}{|\ln \varepsilon_0|}$.

On a fixé ici $\varepsilon_0 = \frac{1}{2\sqrt{c}} e_0 = \frac{1}{2\sqrt{c}}\left(\frac{1+c}{2} - \sqrt{c}\right) = \frac{1}{2}\left(\frac{1}{2\sqrt{c}} + \frac{\sqrt{c}}{2} - 1\right)$

Lorsque $c \in [\frac{1}{2}, 1[$, cette fonction est maximale en $c = 1/2$ d'où

$$\varepsilon_0 \leq \underbrace{\frac{1}{2}\left(\frac{1}{\sqrt{2}} + \frac{1}{2\sqrt{2}} - 1\right)}_{(\simeq 0.03)} < 2^{-5} \quad (\simeq 0.031)$$

donc $|\ln \varepsilon_0| > |\ln 2^{-5}| = 5 \ln 2$, d'où $\frac{1}{|\ln \varepsilon_0|} < \frac{1}{5 \ln 2}$

$\boxed{\text{Pour avoir } \dfrac{e_k}{\sqrt{c}} < \eta, \text{ il suffit donc d'avoir } 2^k \geqslant \dfrac{|\ln(\eta/2)|}{5\ln 2}.}$

Fixons $\eta = 2^{-53} \simeq 10^{-16}$ correspondant à la double précision machine

On obtient la condition $2^k \geqslant \dfrac{54 \times \ln 2}{5 \times \ln 2} = \dfrac{54}{5} \simeq 10.8$

Donc en <u>$k = 4$ itérations</u>, on a $\dfrac{x_k - \sqrt{c}}{\sqrt{c}} < 2^{-53} \simeq 10^{-16}$,

ie $\sqrt{c}$ est calculé à la précision machine.

<u>exemple</u> : $\sqrt{2} = \sqrt{\frac{1}{2} \times 4} = 2 \times \sqrt{\frac{1}{2}}$

| $k$ | approx de $\sqrt{0.5}$ | approx de $\sqrt{2}$ | nb décimales exactes $\sqrt{2}$ |
|---|---|---|---|
| 0 | $3/4$ | $3/2$ | 0 |
| 1 | $\frac{1}{2}\left(\frac{3}{4}+\frac{2}{3}\right)$ | $\frac{17}{12} = 1.41666\ldots$ | 2 |
| 2 | $\frac{1}{2}\left(\frac{17}{24}+\frac{12}{17}\right)$ | (*) $\frac{577}{408} = 1.414215\ldots$ | 5 |
| 3 | $\frac{1}{2} \times \left(\frac{577}{408} \times \frac{1}{2} + \frac{408}{577}\right)$ | $\frac{665857}{470832} = 1.414213562374\ldots$ | 11 |
| 4 | | $1{,}41421356237309{,}49\ldots$ en calcul double précision (**) | 14 en calcul double précision et 23 en arithmétique exacte. |

(*) approximation déjà connue vers 1700 av. JC (tablette babylonienne)

(**) en arithmétique exacte :

$1{,}4142135623730950488016896\ldots$

$\to$ erreur relative $\simeq 0.6 \times 10^{-24}$ $\quad \left(2\varepsilon_0^{2^4} \simeq 10^{-24}\right)$

Le théorème de Kantorovich donne des conditions suffisantes sur les conditions initiales $x_0$ pour que la méthode de Newton converge, ainsi qu'une borne d'erreur explicite.

Théorème de Kantorovich :

Soit $D$ un ouvert de $\mathbb{R}^n$ et $f : D \to \mathbb{R}^n$ une application différentiable sur $D_0 \subset D$ avec $D_0$ fermé et convexe.

Soit $x_0 \in D_0$. On suppose $Df(x_0)$ inversible, $Df(x_0)^{-1} Df$ lipschitzienne sur $D_0$, de constante $K$ pour la norme $|||\ |||$.

On suppose $\| Df(x_0)^{-1} f(x_0) \| \leq \eta$ avec $h := K\eta \leq \frac{1}{2}$.

Soient $r_0 = \dfrac{1-\sqrt{1-2h}}{h}\,\eta$, $r_1 = \dfrac{1+\sqrt{1-2h}}{h}\,\eta$

Alors, si $B(x_0, r_0) \subset D_0$, la suite des itérés de Newton $x_{k+1} = x_k - Df(x_k)^{-1} f(x_k)$ est bien définie, reste dans $B(x_0, r_0)$, et converge vers $a \in B(x_0, r_0)$ avec $f(a) = 0$.

Si $h < \frac{1}{2}$, $a$ est l'unique zéro de $f$ dans $B(x_0, r_1) \cap D_0$, et si $h = \frac{1}{2}$, $a$ est unique dans $\overline{B}(x_0, r_1) \cap D_0$.

De plus, on a la borne d'erreur :

$$\| a - x_k \| \leq \frac{1}{2^k} \left(1 - \sqrt{1-2h}\right)^{2^k} \frac{\eta}{h}.$$

(convergence quadratique pour $h < \frac{1}{2}$).

Par exemple, pour $f(x) = x^2 - c$, le théorème de Kantorovich donne la convergence de la méthode de Newton lorsque $x_0 \geq \sqrt{\frac{c}{2}}$

( condition $h = \dfrac{|x_0^2 - c|}{2x_0^2} \leq \frac{1}{2}$ )

Quelques remarques concernant la méthode de Newton :

- avantage : convergence locale très rapide (quadratique).
- ce n'est généralement plus vrai loin de $x = a$ : p.ex. pour $f(x) = x^2 - c$, la méthode de Newton $x_{k+1} = \frac{1}{2}\left(x_k + \frac{c}{x_k}\right)$ donne lorsque $x_k \gg \sqrt{c}$ : $x_{k+1} - \sqrt{c} = \frac{x_k - \sqrt{c}}{2}\left(1 - \frac{\sqrt{c}}{x_k}\right) \simeq \frac{x_k - \sqrt{c}}{2}$

  L'erreur est donc approximativement divisée par 2 à chaque itération tant que $x_k$ reste éloigné de $\sqrt{c}$.

  Par ailleurs, la convergence n'est pas garantie en général lorsque $x_0$ est trop éloigné de $x = a$.
- à chaque itération, il faut résoudre le système linéaire $Df(x_k)\, d_k = -f(x_k)$ pour calculer $x_{k+1} = x_k + d_k$.

  Le coût peut être élevé ( $\underset{n\to+\infty}{\sim} \frac{2}{3} n^3$ opérations par la méthode de Gauss ).
- Lorsqu'on ne dispose pas d'expression analytique simple pour $Df(x_k)$ (par exemple si $f(x)$ est le résultat d'un algorithme numérique, sans qu'on dispose d'une expression explicite pour $f$) on peut approcher $Df(x_k)$ par différences finies :

$$\frac{\partial f_i}{\partial x_j}(x) \simeq \frac{f_i(x + \delta e_j) - f_i(x)}{\delta}$$

avec $e_j$ $j^{\text{ème}}$ vecteur de la base canonique et $\delta \simeq 0$.

Dans les calculs en virgule flottante, si $\varepsilon$ désigne la précision machine (le plus petit flottant $> 1$ est $1 + \varepsilon$), il est classique de choisir $\delta$ pour minimiser $\max\left(\frac{\varepsilon}{\delta}, \delta\right)$ (erreurs d'arrondi et de troncature) qui conduit à $\delta = \sqrt{\varepsilon}$. En format double précision, $\varepsilon = 2^{-52} \simeq 2.2 \times 10^{-16}$ cela donne $\delta \sim 10^{-8}$.

- Lorsque $f(x)$ est coûteux à évaluer, le calcul de $Df(x_k)$ par différences finies est coûteux car il nécessite de calculer $f(x_k + \delta e_j)$ pour $j = 1, 2, \cdots, n$.

- Pour ces différentes raisons, on utilise souvent des méthodes dites "quasi-Newton" dans lesquelles on remplace $Df(x_k)^{-1}$ par une approximation moins coûteuse à actualiser à chaque itération.

  Un exemple est donné par la méthode de Broyden étudiée en TD, qui généralise la méthode de la sécante dans $\mathbb{R}^n$.

Résolution de l'équation de Poisson-Boltzmann par la méthode de Newton :

$$\begin{cases} U_{i+1} - 2U_i + U_{i-1} + \dfrac{h}{2r_i}(U_{i+1} - U_{i-1}) - K^2 h^2 \,\mathrm{sh}\, U_i = 0, & 1 \le i \le n-1 \quad (U_n = 0) \\ 2(U_1 - U_0) - K^2 h^2 \,\mathrm{sh}\, U_0 + \mu h (2-h) = 0 & \end{cases}$$

s'écrit :

$$f(U) = MU - K^2 h^2 \,\mathrm{sh}\, U + \mu h (2-h) e_1 = 0 \quad \text{avec} \quad U = \begin{pmatrix} u_0 \\ u_1 \\ \vdots \\ u_{n-1} \end{pmatrix} \in \mathbb{R}^n, \quad M \in M_n(\mathbb{R})$$

tridiagonale.

La méthode de Newton s'écrit :

$$Df(U^{(k)})\,(U^{(k+1)} - U^{(k)}) = -f(U^{(k)})$$

avec $Df(U^{(k)}) = M - K^2 h^2 \,\mathrm{diag}\,(\mathrm{ch}(U^{(k)}))$ → tridiagonale et à diagonale strictement dominante.

Dans cet exemple, l'évaluation de $Df(U^{(k)})$ est donc peu coûteuse, ainsi que la résolution du système de matrice $Df(U^{(k)})$ (coût $O(n)$ avec Gauss sans permutations).

L'approximation initiale $U^{(0)}$ peut être déterminée en résolvant le système linéaire obtenu avec l'approximation $\mathrm{sh}\, U_i \simeq U_i$ (équation de Debye-Hückel).