232

MÉTHODES D'APPROXIMATION DES SOLUTIONS D'UNE ÉQUATION $F(X)=0$. EXEMPLES.

Cadre : On se donne une fonction $f : D \subset \mathbb{R}^n \to \mathbb{R}^n$ dont on souhaite approcher un zéro.

Déf 1 : Soient (X,d) un espace métrique, $c \in X$ et $(x_p)_{p\geq 0}$ une suite d'éléments de X qui converge vers c.
- S'il existe $\tau < 1$ et $p_0 \in \mathbb{N}$ tels que $\forall p \geq p_0$, $d(x_{p+1}, c) \leq \tau\, d(x_p, c)$, on dit que la convergence est linéaire.
- S'il existe $k > 0$ tel que $\forall p \geq 0$, $d(x_{p+1}, c) \leq k\, d(x_p, c)^m$ pour un certain $m > 1$ indépendant de p, on dit que la convergence est d'ordre m. Pour $m = 2$, on dit aussi qu'elle est quadratique.

I - INTRODUCTION AUX MÉTHODES ITÉRATIVES

1 - Méthode de dichotomie

Thm 2 (Méthode de dichotomie) : Soient $a < b \in \mathbb{R}$ et $f : [a,b] \to \mathbb{R}$ une fonction continue telle que $f(a) f(b) \leq 0$.
- Posons $a_0 = a$ et $b_0 = b$.
- Supposons $a_0, \dots, a_p, b_0, \dots, b_p$ construits pour un certain $p \geq 0$.

  Si $f(a_p) f\left(\frac{a_p + b_p}{2}\right) \leq 0$, on pose $a_{p+1} = a_p$ et $b_{p+1} = \frac{a_p + b_p}{2}$.

  Sinon, on pose $a_{p+1} = \frac{a_p + b_p}{2}$ et $b_{p+1} = b_p$.

Alors les suites $(a_p)_{p\geq 0}$ et $(b_p)_{p\geq 0}$ ainsi définies sont adjacentes. De plus, leur limite commune, notée c, est un zéro de f.

Rq 3 : La méthode de dichotomie ne se généralise pas en dimensions supérieures.

2 - Théorème du point fixe de Picard

Rq 4 : La recherche d'un zéro d'une fonction peut se ramener à celle d'un point fixe d'une fonction auxiliaire.
Par exemple, pour tout $\lambda \neq 0$, les zéros de f sont exactement les points fixes de $\varphi : x \mapsto x + \lambda f(x)$.

Thm 5 (Théorème du point fixe de Picard) : Soient (X,d) un espace métrique complet et $\varphi : X \to X$ une fonction contractante.
Alors φ possède un unique point fixe $c \in X$. De plus, pour tout $x_0 \in X$, la suite $(x_p)_{p\geq 0}$ définie par $x_{p+1} = \varphi(x_p)$ converge vers c à vitesse linéaire.

Ex 6 : La fonction $f : (x,y) \mapsto (4x - \sin(x+y), 3y - 2\arctan(x-y))$ possède un unique zéro $c \in \mathbb{R}^2$. De plus, pour tout $z_0 \in \mathbb{R}^2$, la suite $(z_p)_{p\geq 0}$ définie par $z_{p+1} = \varphi(z_p)$, où $\varphi : (x,y) \mapsto \left(\frac{1}{4}\sin(x+y), \frac{2}{3}\arctan(x-y)\right)$, converge vers c à vitesse linéaire.

3 - Différents types de points fixes [Dem, p 55]

Déf 7 : Soient I un intervalle ouvert de $\mathbb{R}$, $\varphi : I \to \mathbb{R}$ une fonction et $c \in I$ un point fixe de φ. On suppose que φ est dérivable en c.
- Si $|\varphi'(c)| < 1$, on dit que c est un point fixe attractif de φ. Lorsque $\varphi'(c) = 0$, on dit que c est superattractif.
- Si $|\varphi'(c)| > 1$, on dit que c est un point fixe répulsif de φ.

Thm 8 : Avec les notations précédentes, on a :
- Si c est un point fixe attractif de φ, alors il existe $\delta > 0$ tel que $[c-\delta, c+\delta]$ soit contenu dans I et stable par φ et tel que pour tout $x_0 \in [c-\delta, c+\delta]$, la suite $(x_p)_{p\geq 0}$ définie par $x_{p+1} = \varphi(x_p)$ converge vers c à vitesse linéaire.
Si de plus, φ est m fois dérivable en c pour un certain $m \geq 2$ et $\varphi^{(\ell)}(c) = 0$ pour $\ell \in [\![1, m-1]\!]$, alors la convergence est d'ordre m.
En outre, si $\varphi^{(m)}(c) \neq 0$, alors $x_{p+1} - c \sim \frac{\varphi^{(m)}(c)}{m!}(x_p - c)^m$.
- Si c est un point fixe répulsif de φ, alors il existe $\delta > 0$ tel que $[c-\delta, c+\delta]$ soit contenu dans I et tel que pour tout $x_0 \in [c-\delta, c+\delta] \setminus \{c\}$, il existe p_0 tel que $x_{p_0} \notin [c-\delta, c+\delta]$ où $(x_p)_{p\geq 0}$ est définie par $x_{p+1} = \varphi(x_p)$.

Ex 9 : Les zéros de $f : x \mapsto x^3 - 4x + 1$ sont exactement les points fixes de $\varphi : x \mapsto \frac{1}{4}(x^3 + 1)$. L'étude de f montre que f possède exactement trois zéros : $-2{,}5 < c_1 < -2$; $0 < c_2 < 0{,}5$ et $1{,}5 < c_3 < 2$.
c_2 est un point fixe attractif de φ. c_1 et c_3 sont des points fixes répulsifs de φ et donc attractifs de $\varphi^{-1} : x \mapsto \sqrt[3]{4x - 1}$ où $\sqrt[3]{\cdot}$ est la bijection réciproque de $x \mapsto x^3$ sur $\mathbb{R}$.

Rq 10 : Si c est un point fixe de φ tel que $|\varphi'(c)| = 1$, on ne peut rien dire sur la convergence des itérées de φ sur un voisinage de c. Il faut regarder le premier nombre dérivé non nul de φ en c (s'il existe).

Ex 11 :
- Pour tout $x_0 > 0$, la suite $(x_p)_{p\geq 0}$ définie par $x_{p+1} = \sin(x_p)$ converge vers 0.
- Pour tout $x_0 > 0$, la suite $(x_p)_{p\geq 0}$ définie par $x_{p+1} = \mathrm{sh}(x_p)$ tend vers $+\infty$.

Rq 12 : Les notions de points fixes attractif et superattractif se généralisent en dimensions supérieures.

Déf 13 : Soient Ω un ouvert de $\mathbb{R}^n$, $\varphi : \Omega \to \mathbb{R}^n$ une fonction et $c \in \Omega$ un point fixe de φ. On suppose que φ est différentiable en c.
Si $\rho(d\varphi(c)) < 1$ (où ρ désigne le rayon spectral), on dit que c est un point fixe attractif de φ.
Lorsque $d\varphi(c) = 0$, on dit que c est superattractif.

Thm 14 : Avec les mêmes notations, si c est un point fixe attractif de φ, alors il existe un voisinage V de c contenu dans Ω, stable par φ et tel que pour tout $x_0 \in V$, la suite $(x_p)_{p\geq 0}$ définie par $x_{p+1} = \varphi(x_p)$ converge vers c et soit telle que pour toute norme $\|.\|$ sur $\mathbb{R}^n$, il existe $k>0$ et $\tau<1$ tels que $\forall p \geq 0$, $\|x_p - c\| \leq k\tau^p$.
Si de plus, φ est m fois différentiable en c pour un certain $m\geq 2$ et $d^\ell\varphi(c) = 0$ pour $\ell \in [\![1, m-1]\!]$, alors la convergence est d'ordre m (ceci ne dépend pas du choix de la norme sur $\mathbb{R}^n$).

II - MÉTHODES DE TYPE NEWTON

1 - Méthode de Newton en dimension 1 [Dem]

Thm 15 (Méthode de Newton) : Soient I un intervalle ouvert de $\mathbb{R}$ et $f : I \to \mathbb{R}$ une fonction de classe C^1 dont la dérivée ne s'annule pas et qui possède un zéro $c \in I$.
Alors c est l'unique point fixe de $\varphi : x \mapsto x - \frac{f(x)}{f'(x)}$. De plus, φ est dérivable en c et c est superattractif. Il existe donc $\delta > 0$ tel que pour tout $x_0 \in [c-\delta, c+\delta]$, la suite $(x_p)_{p\geq 0}$ définie par $x_{p+1} = \varphi(x_p)$ converge vers c à vitesse linéaire.
De plus, si f est deux fois dérivable, alors la convergence est quadratique.
En outre, si $f' \times f'' \leq 0$ (resp. $f'\times f'' \geq 0$), alors $I \cap]-\infty, c]$ (resp. $I \cap [c, +\infty[$) est stable par φ et pour tout $x_0 \in I \cap]-\infty, c]$ (resp. $I \cap [c, +\infty[$), la suite $(x_p)_{p\geq 0}$ définie comme ci-dessus est croissante (resp. décroissante) et converge vers c.

Appli 16 (Méthode de Héron) : Soit $y > 0$.
Alors pour tout $x_0 \in [\sqrt{y}, +\infty[$, la suite $(x_p)_{p\geq 0}$ définie par $x_{p+1} = \frac{1}{2}\left(x_p + \frac{y}{x_p}\right)$ converge vers $\sqrt{y}$ à vitesse quadratique.

Rq 17 : Un inconvénient de la méthode de Newton est qu'il n'y a pas convergence des itérées de φ pour toute condition initiale.
La condition initiale doit souvent être "assez proche" du zéro que l'on souhaite approcher.

Ex 18 : Pour tout x_0 de valeur absolue "suffisamment grande", la suite $(x_p)_{p\geq 0}$ définie par $x_{p+1} = \varphi(x_p)$, où φ est la fonction associée à arctan donnée par la méthode de Newton, vérifie : $|x_p| \xrightarrow[p\to\infty]{} +\infty$.

Thm 19 (Méthode de Newton pour les polynômes) :
Soit $P = \lambda \prod_{i=1}^{r} (X - \alpha_i)^{m_i} \in \mathbb{R}[X]$ un polynôme non constant et scindé sur $\mathbb{R}$ de racines $\alpha_1 < \dots < \alpha_r$ de multiplicités $m_1, \dots, m_r \geq 1$.
Alors $]\alpha_r, +\infty[$ est stable par $\varphi : x \mapsto x - \frac{P(x)}{P'(x)}$ et pour tout $x_0 \in]\alpha_r, +\infty[$, la suite $(x_p)_{p\geq 0}$ définie par $x_{p+1} = \varphi(x_p)$ est strictement décroissante et converge vers c.
Si de plus, $m_r \geq 2$, alors $x_p - \alpha_r \sim \beta\left(1 - \frac{1}{m_r}\right)^p$ pour un certain $\beta > 0$.

Rq 20 : Les racines d'un polynôme sont exactement les valeurs propres de sa matrice compagnon associée.

2 - Méthode de la sécante [Dem]

Rq 21 : Un inconvénient de la méthode de Newton est qu'elle nécessite le calcul du nombre dérivé de la fonction en plusieurs points. La méthode de la sécante évite ceci en remplaçant le nombre dérivé par un taux d'accroissement.

DVPT

Thm 22 (Méthode de la sécante) : Soient I un intervalle ouvert de $\mathbb{R}$ et $f : I \to \mathbb{R}$ une fonction de classe C^2 dont la dérivée ne s'annule pas et qui possède un zéro $c \in I$.
Alors il existe $\delta > 0$ et $k > 0$ tels que pour tous x_0, x_1 distincts dans $[c-\delta, c+\delta]$, la suite (x_p) définie par $x_{p+1} = x_p - \frac{f(x_p)}{\tau_p}$, où $\tau_p = \frac{f(x_p) - f(x_{p-1})}{x_p - x_{p-1}}$, converge vers c et vérifie :
$\forall p$, $|x_p - c| \leq \frac{1}{k}\left(k \max\{|x_0 - c|, |x_1 - c|\}\right)^{F_p}$ où $(F_p)_{p\geq 0}$ est la suite de Fibonacci de premiers termes $F_0 = F_1 = 1$.

3 - Méthode de Newton-Raphson [Dem]

Thm 23 (Méthode de Newton-Raphson) : Soient Ω un ouvert de $\mathbb{R}^n$ et $f : \Omega \to \mathbb{R}^n$ une fonction de classe C^1 dont la différentielle est inversible en tout point et qui possède un zéro $c \in \Omega$.
Alors $\varphi : x \mapsto x - (df(x))^{-1}(f(x))$ est différentiable en c et c est un point fixe superattractif de φ. Il existe donc un voisinage V de c tel que pour tout $x_0 \in V$, la suite $(x_p)_{p\geq 0}$ définie par $x_{p+1} = \varphi(x_p)$ converge vers c.
Si de plus, f est deux fois différentiable, alors la convergence est quadratique.

Rq 24 : La méthode de Newton-Raphson est une généralisation de la méthode de Newton en dimension 1.

Ex 25 : Pour la condition initiale $x_0 = (-2 ; 0{,}2)$, la méthode de Newton-Raphson donne une approximation de l'unique zéro de $f : (x,y) \mapsto (x^2 + xy - 2y^2 - 4, xe^x + ye^y)$ avec 9 décimales exactes en 3 itérations.

III _ MÉTHODES DE RÉSOLUTION APPROCHÉE DE SYSTÈMES LINÉAIRES

But : On se donne $A \in GL_n(\mathbb{R})$ et $b \in \mathbb{R}^n$. On souhaite approcher l'unique solution $c = A^{-1}b$ du système linéaire $Ax = b$.

Rq 26 : Il existe des méthodes directes de résolution de systèmes linéaires mais celles-ci sont lentes en "grande dimension".

1. Méthodes itératives linéaires [QSS]

Déf 27 : On appelle décomposition régulière de A tout couple $(M,N) \in GL_n(\mathbb{R}) \times M_n(\mathbb{R})$ tel que $A = M - N$.

Rq 28 : Généralement, la matrice M est choisie de manière à être facilement inversible.

Thm 29 : Soit (M,N) une décomposition régulière de A.
On a équivalence entre :
(i) Pour tout $x_0 \in \mathbb{R}^n$, la suite $(x_p)_{p\geq 0}$ définie par $x_{p+1} = M^{-1}Nx_p + M^{-1}b$ converge vers c.
(ii) $\rho(M^{-1}N) < 1$.
En particulier, l'hypothèse (i) ne dépend pas de b.
De plus, si (i) ou (ii) est vérifiée, alors pour toute norme $\|.\|$ sur $\mathbb{R}^n$ et pour tout $x_0 \in \mathbb{R}^n$, il existe $k > 0$ et $\tau < 1$ tels que la suite $(x_p)_{p\geq 0}$ définie comme ci-dessus vérifie : $\forall p \geq 0, \|x_p - c\| \leq k\tau^p$.

Déf 30 : Avec les notations précédentes, si (i) ou (ii) est vérifiée, on dit que la décomposition régulière (M,N) de A converge.

Thm 31 : Supposons que $A \in S_n^{++}(\mathbb{R})$. Soit (M,N) une décomposition régulière de A telle que ${}^tM + N \in S_n^{++}(\mathbb{R})$. Alors (M,N) converge.

Déf 32 : Supposons que les coefficients diagonaux de A sont tous non nuls. Soient D diagonale, E triangulaire inférieure stricte et F triangulaire supérieure stricte telles que $A = D - E - F$.
- On appelle méthode de Jacobi la décomposition régulière $(D, D-A)$ de A.
- Pour $\omega \neq 0$, on appelle méthode de relaxation de paramètre ω la décomposition régulière $\left(\frac{1}{\omega}D - E, \frac{1-\omega}{\omega}D + F\right)$ de A.

Pour $\omega = 1$, on parle aussi de méthode de Gauss-Seidel.

Prop 33 :
- Si les coefficients diagonaux de A sont tous non nuls et $\omega \notin]0,2[$, alors la méthode de relaxation ne converge pas.
- Si $A \in S_n^{++}(\mathbb{R})$ et $\omega \in]0,2[$, alors la méthode de relaxation converge.
- Si A est à diagonale strictement dominante, alors les méthodes de Jacobi et Gauss-Seidel convergent.

Ex 34 : Si $A = \begin{pmatrix} 2 & -1 & & 0 \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ 0 & & -1 & 2 \end{pmatrix}$, alors les méthodes de Jacobi et de relaxation pour $\omega \in]0,2[$ convergent.

2 - Méthode du gradient à pas optimal

Notation : Notons $\langle .,. \rangle$ le produit scalaire usuel sur $\mathbb{R}^n$ et $\|.\|$ sa norme associée.

Prop 35 : Supposons que $A \in S_n^{++}(\mathbb{R})$. On considère la fonction $F : x \mapsto \frac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle$. Alors pour tout $x \in \mathbb{R}^n$, $\nabla F(x) = Ax - b$ et c est l'unique élément de $\mathbb{R}^n$ qui minimise F.

DVPT

Thm 36 (Méthode du gradient à pas optimal) :
Supposons que $A \in S_n^{++}(\mathbb{R})$. Soit $x_0 \in \mathbb{R}^n$.
- Supposons $x_0, \dots, x_p$ construits pour un certain $p \geq 0$.

Si $\nabla F(x_p) = 0$ (où F est définie comme précédemment), on arrête la construction de la suite.

Sinon, on pose $x_{p+1} = x_p - t_p \nabla F(x_p)$ où $t_p = \dfrac{\|\nabla F(x_p)\|^2}{\langle A\nabla F(x_p), \nabla F(x_p)\rangle}$ est l'unique élément de $\mathbb{R}_+^*$ qui minimise $t \mapsto F(x_p - t\nabla F(x_p))$.

Alors la suite (x_p) ainsi définie converge vers c et vérifie :
$\forall p, \|x_p - c\| \leq \sqrt{\mathrm{cond}(A)}\, \|x_0 - c\| \left(\dfrac{\mathrm{cond}(A) - 1}{\mathrm{cond}(A)+1}\right)^p$
où $\mathrm{cond}(A)$ est le conditionnement de A.

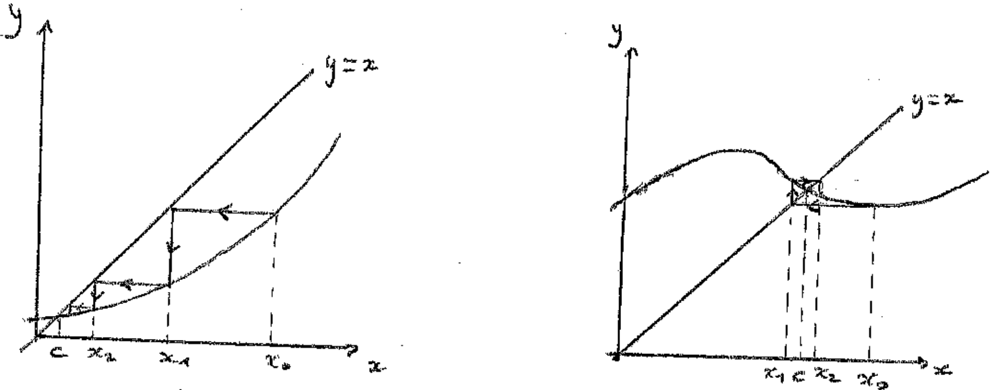
Rq 37 : Pour une matrice A "mal conditionnée", la méthode du gradient à pas optimal s'avère "peu efficace".

Rq 38 : Un autre algorithme, appelé méthode du gradient conjugué, permet d'approcher "plus efficacement" c.
Pour tout $x_0 \in \mathbb{R}^n$, la suite (x_p) donnée par cet algorithme converge vers c en au plus n itérations et vérifie :
$\forall p, \|x_p - c\| \leq 2\sqrt{\mathrm{cond}(A)}\, \|x_0 - c\| \left(\dfrac{\sqrt{\mathrm{cond}(A)} - 1}{\sqrt{\mathrm{cond}(A)}+1}\right)^p$ (ADMIS)

[Dem] Demailly
[QSS] Quarteroni, Sacco, Saleri

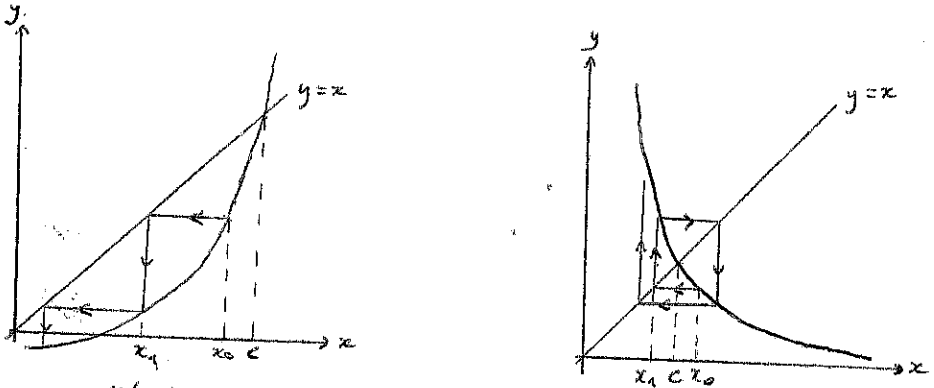
ANNEXE

Figure 1 : Point fixe attractif



$0 \leq \varphi'(c) < 1$ $\qquad$ $-1 < \varphi'(c) < 0$

Figure 2 : Point fixe répulsif



$\varphi'(c) > 1$ $\qquad$ $\varphi'(c) < -1$

Figure 3 : Méthode de Newton

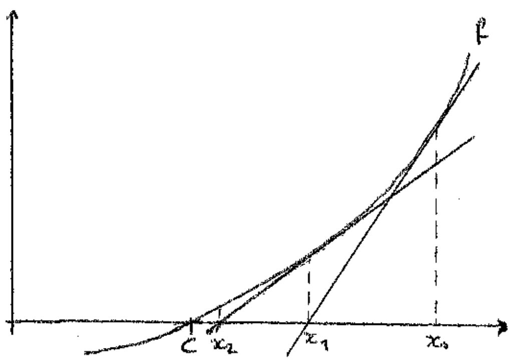


Figure 4 : Méthode de la sécante

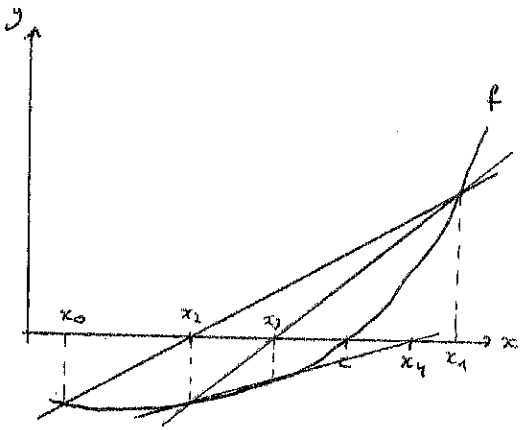
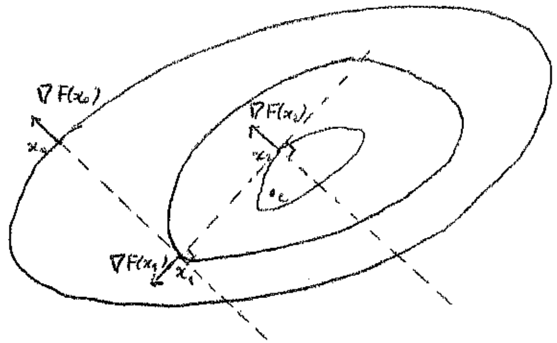


Figure 5 : Méthode du gradient à pas optimal



DÉVELOPPEMENT : MÉTHODE DE LA SÉCANTE

Théorème : Soient I un intervalle ouvert de $\mathbb{R}$ et $f : I \to \mathbb{R}$ de classe $\mathcal{C}^2$, dont la dérivée ne s'annule pas, et qui possède un zéro $c \in I$. Alors il existe $\delta > 0$ et $K > 0$ tels que pour tous x_0, x_1 distincts dans $[c-\delta, c+\delta]$, la suite (x_p) définie par
$$x_{p+1} = x_p - \frac{f(x_p)}{\tau_p} \quad \text{où} \quad \tau_p = \frac{f(x_p) - f(x_{p-1})}{x_p - x_{p-1}}$$
converge vers c et vérifie
$$\forall p \geqslant 0, \ \|x_p - c\| \leqslant \frac{1}{K}\left(K \max\{|x_0 - c|; |x_1 - c|\}\right)^{F_p}$$
où $(F_p)_{p \geqslant 0}$ est la suite de Fibonacci de premiers termes $F_0 = F_1 = 1$.

Démonstration :

Comme I est ouvert et contient c, il existe $r > 0$ tel que le segment $J = [c-r, c+r]$ soit inclus dans I.

On note, pour $i \in \{0, 1, 2\}$, $M_i = \max\limits_J |f^{(i)}|$ et $m_i = \min\limits_J |f^{(i)}|$.

Étape 1 On considère $\tau : \left| \begin{array}{l} I \times I \longrightarrow \mathbb{R} \\ (x,y) \longmapsto \begin{cases} \dfrac{f(y) - f(x)}{y - x} & \text{si } y \neq x \\ f'(x) & \text{si } y = x \end{cases} \end{array} \right.$

On suppose $x, y \in I$ et $x \neq y$.

La fonction $\theta : t \in [0,1] \mapsto f(x + t(y-x))$ est de classe $\mathcal{C}^1$ avec $\theta(0) = f(x)$ et $\theta(1) = f(y)$.

On a donc
$$\theta(1) - \theta(0) = \int_0^1 \theta'(t)\, dt$$
soit
$$f(y) - f(x) = \int_0^1 (y-x)\, f'(x + t(y-x))\, dt$$
d'où $\tau(x,y) = \int_0^1 f'(x + t(y-x))\, dt$

(et cette égalité est encore valable si $y = x$, car $\int_0^1 f'(x)\,dt = f'(x) = \tau(x,x)$).

Selon le théorème de dérivation sous le signe intégrale, τ possède des dérivées partielles par rapport à x et y en tout point de $I \times I$, et on a :

$$\forall (x,y) \in I \times I, \qquad \frac{\partial \tau}{\partial x}(x,y) = \int_0^1 (1-t) f''(x+t(y-x))\,dt$$

$$\frac{\partial \tau}{\partial y}(x,y) = \int_0^1 t\, f''(x+t(y-x))\,dt$$

De plus, le théorème de continuité sous le signe intégrale assure que $\frac{\partial \tau}{\partial x}$ et $\frac{\partial \tau}{\partial y}$ sont continues sur $I \times I$, et donc τ est de classe $\mathcal{C}^1$ sur $I \times I$.

Pour $(x,y) \in J \times J$, on a en particulier :

$$\left|\frac{\partial \tau}{\partial x}(x,y)\right| \leq M_2 \int_0^1 (1-t)\,dt = \frac{M_2}{2}$$

$$\left|\frac{\partial \tau}{\partial y}(x,y)\right| \leq M_2 \int_0^1 t\,dt = \frac{M_2}{2}$$

$$|\tau(x,y)| = \left|\int_0^1 f'(x+t(y-x))\,dt\right| \geq m_1 \quad \text{(car } f' \text{ est de signe constant)}$$

$$|\tau(x,y) - f'(x)| = |\tau(x,y) - \tau(x,x)| \leq \underbrace{\max_{\substack{t \in [x,y] \\ \text{ou } [y,x]}} \left|\frac{\partial \tau}{\partial y}(x,t)\right|}_{\leq M_2/2 \text{ car } x \in J \text{ et } [x,y] \text{ (ou } [y,x]) \subset J} \cdot |y-x| \leq \frac{M_2}{2}|y-x|$$

Étape 2 On pose, pour tout $(x,y) \in I \times I$,

$$\psi(x,y) = x - \frac{f(x)}{\tau(x,y)}$$

Le but est de majorer $|\psi(x,y) - c|$, pour $(x,y) \in J \times J$, en fonction de $|x-c|$ et $|y-c|$.

Soit $(x,y) \in J \times J$. $\psi(x,y) - c = \psi(x,y) - \psi(c,c) = \alpha(1) - \alpha(0)$

où $\alpha : t \in [0,1] \mapsto \psi(c+t(x-c), c+t(y-c))$ est de classe $\mathcal{C}^1$ sur $[0,1]$. Alors

$$\psi(x,y) - c = \alpha(1) - \alpha(0) = \int_0^1 \alpha'(t)\,dt$$

$$= \int_0^1 \left((x-c)\frac{\partial \psi}{\partial x}(c+t(x-c), c+t(y-c)) + (y-c)\frac{\partial \psi}{\partial y}(c+t(x-c), c+t(y-c))\right)dt$$

Par ailleurs, pour tout $(a,b) \in I \times I$,

$$\frac{\partial \psi}{\partial x}(a,b) = \frac{\tau(a,b) - f'(a)}{\tau(a,b)} + \frac{f(a)}{\tau(a,b)^2} \cdot \frac{\partial \tau}{\partial x}(a,b)$$

$$\frac{\partial \psi}{\partial y}(a,b) = \frac{f(a)}{\tau(a,b)^2} \cdot \frac{\partial \tau}{\partial y}(a,b)$$

Prenant $a = c + t(x-c)$ et $b = c + t(y-c)$, on obtient les inégalités suivantes à l'aide de l'étape 1 :

$$\left|\frac{\partial \psi}{\partial x}(c+t(x-c), c+t(y-c))\right| \leq \frac{M_2}{2m_1} \underbrace{|y-x|}_{\leq |y-c|+|x-c|} t + \frac{M_2}{2m_1^2} |f(c+t(x-c))|$$

or $|f(c+t(x-c))| = |f(c+t(x-c)) - \underbrace{f(c)}_{=0}|$

$\leq M_1 |x-c| t$ (inégalité des accroissements finis)

d'où $\left|\frac{\partial \psi}{\partial x}(c+t(x-c), c+t(y-c))\right| \leq \frac{M_2}{2m_1}(|y-c|+|x-c|)t + \frac{M_1 M_2}{2m_1^2}|x-c| t$

De même $\left|\frac{\partial \psi}{\partial y}(c+t(x-c), c+t(y-c))\right| \leq \frac{M_1 M_2}{2m_1^2} |x-c| t$

On en déduit que, pour tout $(x,y) \in J \times J$,

$$|\psi(x,y) - c| \leq \int_0^1 \left(|x-c| \left(\frac{M_2}{2m_1}(|y-c|+|x-c|) + \frac{M_1M_2}{2m_1^2}|x-c| \right) + |y-c| \cdot \frac{M_1 M_2}{2m_1^2} |x-c| \right) t \, dt$$

$$\leq \frac{1}{2} |x-c| \underbrace{\left(|x-c| + |y-c| \right)}_{\leq 2\max\{|x-c|, |y-c|\}} \underbrace{\left(\frac{M_2}{2m_1} + \frac{M_1 M_2}{2m_1^2} \right)}_{:= K}$$

donc

$$|\psi(x,y) - c| \leq K \cdot |x-c| \cdot \max\{|x-c|\,;\, |y-c|\}$$

<u>Étape 3</u> On pose $\delta = \min\{r\,;\, 1/K\}$

On suppose que $x_0, x_1 \in [c-\delta, c+\delta]$ et que $x_0 \neq x_1$.

On a : $\psi([c-\delta, c+\delta]^2) \subseteq [c-\delta, c+\delta] \subseteq J \subset I$

↳ car $|\psi(x,y) - c| \leq K \underbrace{|x-c|}_{\leq \delta \leq \frac{1}{K}} \underbrace{\max\{|x-c|; |y-c|\}}_{\leq \delta}$

Par suite, tous les termes de la suite (x_p) définie dans l'énoncé du théorème sont dans $[c-\delta, c+\delta]$

On peut alors écrire : $\forall p \geqslant 1,\ x_{p+1} = \psi(x_p, x_{p-1})$

Puisque $[c-\delta, c+\delta] \subseteq J$,

$$|x_{p+1} - c| \leqslant K\,|x_p - c|\, \underbrace{\max\{|x_p - c| ; |x_{p-1} - c|\}}_{\leqslant\, \delta\, \leqslant\, \frac{1}{K}} \leqslant |x_p - c|$$

donc la suite $(|x_p - c|)_{p \geqslant 1}$ est décroissante.

Il vient donc, pour tout $p \geqslant 2$, $|x_{p+1} - c| \leqslant K\,|x_p - c| \cdot |x_{p-1} - c|$.

<u>Étape 4</u> Montrons par récurrence que, pour tout $p \geqslant 0$,

$$|x_p - c| \leqslant \frac{1}{K}\left(K \max\{|x_0 - c| ; |x_1 - c|\}\right)^{F_p}$$

I/ Pour $p = 0$ et $p = 1$, le résultat est immédiat.

Pour $p = 2$: $|x_2 - c| \leqslant K\, \underbrace{|x_1 - c|}_{\leqslant \frac{1}{K} \cdot K \max\{|x_0 - c| ; |x_1 - c|\}} \cdot \max\{|x_0 - c| ; |x_1 - c|\}$

(c'est vrai pour $p = 1$)

$$\leqslant \frac{1}{K}\left(K \max\{|x_0 - c| ; |x_1 - c|\}\right)^2$$

H/ Soit $p \geqslant 2$ tel que, pour tout $q \in [\![0 ; p]\!]$,

$$|x_q - c| \leqslant \frac{1}{K}\left(K \max\{|x_0 - c| ; |x_1 - c|\}\right)^{F_q}$$

$$\begin{aligned}
|x_{p+1} - c| &\leqslant K\,|x_p - c| \cdot |x_{p-1} - c| \\
&\leqslant K \cdot \frac{1}{K}\left(K \max\{|x_0 - c| ; |x_1 - c|\}\right)^{F_p} \cdot \frac{1}{K}\left(K \max\{|x_0 - c| ; |x_1 - c|\}\right)^{F_{p-1}} \\
&\leqslant \frac{1}{K}\left(K \max\{|x_0 - c| ; |x_1 - c|\}\right)^{F_{p+1}}
\end{aligned}$$

car $F_{p+1} = F_p + F_{p-1}$.

ce qui conclut la récurrence.

DÉVELOPPEMENT : MÉTHODE DU GRADIENT À PAS OPTIMAL

Soient $A \in S_n^{++}(\mathbb{R})$ et $b \in \mathbb{R}^n$. Soient $\lambda_1 \leqslant \ldots \leqslant \lambda_n$ les valeurs propres de A. On note $c = A^{-1}b$ la solution de $Ax = b$.

D'après le théorème spectral, il existe $P \in O_n(\mathbb{R})$ telle que
$$A = {}^tPDP \text{ avec } D = \mathrm{diag}(\lambda_1, \ldots, \lambda_n).$$

Proposition: On considère la fonction
$$F : x \in \mathbb{R}^n \longmapsto \frac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle$$

1. Pour tout $x \in \mathbb{R}^n$, $\nabla F(x) = Ax - b$.
2. c est l'unique élément de $\mathbb{R}^n$ qui minimise F.

Démonstration:

1. Soient $x, h \in \mathbb{R}^n$.

$$\begin{aligned} F(x+h) &= \frac{1}{2}\langle A(x+h), x+h\rangle - \langle b, x+h\rangle \\ &= \underbrace{\frac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle}_{} + \underbrace{\frac{1}{2}\langle Ax, h\rangle + \frac{1}{2}\langle Ah, x\rangle - \langle b, h\rangle}_{} + \frac{1}{2}\langle Ah, h\rangle \\ &= \quad F(x) \quad + \quad \langle Ax - b, h\rangle \quad + o(\|h\|) \end{aligned}$$

car A est symétrique

et $|\langle Ah, h\rangle| \underset{\text{théorème spectral}}{\leqslant} \lambda_n \|h\|^2 \xrightarrow[\|h\| \to 0]{} 0$

Donc $\nabla F(x) = Ax - b$.

2. Soit $x \in \mathbb{R}^n$.

$$F(x) = \frac{1}{2}\underbrace{\langle Ax, x\rangle}_{\substack{\geqslant \lambda_1 \|x\|^2 \\ \text{(th. spectral)}}} - \underbrace{\langle b, x\rangle}_{\substack{\leqslant \|b\| \cdot \|x\| \\ \text{(Cauchy-Schwarz)}}} \geqslant \frac{1}{2}\lambda_1 \|x\|^2 - \|b\| \cdot \|x\| \xrightarrow[\|x\| \to +\infty]{} +\infty$$

Ainsi il existe $R > 0$ tel que pour tout $x \in \mathbb{R}^n$ vérifiant $\|x\| > R$, on a $F(x) \geqslant F(0)$. Sur $\overline{B(0,R)}$ qui est compacte, F est continue, donc elle y admet un minimum et atteint celui-ci. De plus ce minimum est global, car si $\|x\| > R$, on a $F(x) \geqslant F(0) \geqslant \min\limits_{\overline{B(0,R)}} F$.

Si $d \in \mathbb{R}^n$ minimise F, on a $\quad \nabla F(d) = Ad - b = 0$

d'où $\quad d = A^{-1}b = c$.

Donc c est l'unique minimiseur de F sur $\mathbb{R}^n$.

<u>Lemme</u> : (Inégalité de Kantorovitch)

Pour tout $x \in \mathbb{R}^n$,

$$\langle Ax, x\rangle \langle A^{-1}x, x\rangle \leqslant \frac{1}{4}\left(\sqrt{\frac{\lambda_1}{\lambda_n}} + \sqrt{\frac{\lambda_n}{\lambda_1}}\right)^2 \|x\|^4$$

<u>Démonstration</u> :

Comme $A \in S_n^{++}(\mathbb{R})$, alors A^{-1} est aussi dans $S_n^{++}(\mathbb{R})$.

Soient $x \in \mathbb{R}^n$ et $t > 0$. $\frac{1}{t}\langle Ax, x\rangle \geqslant 0$ et $t\langle A^{-1}x, x\rangle \geqslant 0$.

D'après l'inégalité arithmético-géométrique :

$$\sqrt{\langle Ax, x\rangle\langle A^{-1}x, x\rangle} \leqslant \frac{1}{2}\left(\frac{1}{t}\langle Ax, x\rangle + t\langle A^{-1}x, x\rangle\right) = \frac{1}{2}\left\langle\left(\frac{1}{t}A + tA^{-1}\right)x, x\right\rangle$$

Or $\frac{1}{t}A + tA^{-1} = {}^tP \operatorname{diag}\left(\frac{\lambda_i}{t} + \frac{t}{\lambda_i}\right)P$, et on en déduit que

$$\sqrt{\langle Ax, x\rangle\langle A^{-1}x, x\rangle} \leqslant \frac{1}{2} \underbrace{\max_{i \in [\![1,n]\!]}\left\{\frac{\lambda_i}{t} + \frac{t}{\lambda_i}\right\}}_{\leqslant \max\limits_{\lambda \in [\lambda_1, \lambda_n]}\{\theta_t(\lambda)\}} \|x\|^2$$

où $\theta_t : \lambda > 0 \longmapsto \frac{\lambda}{t} + \frac{t}{\lambda}$ est convexe.

Par conséquent : $\quad \forall \lambda \in [\lambda_1, \lambda_n],\ \theta_t(\lambda) \leqslant \max\{\theta_t(\lambda_1), \theta_t(\lambda_n)\}$

Ceci est vrai pour tout $t > 0$, donc en particulier pour $t = \sqrt{\lambda_1\lambda_n}$.

Or $\theta_t(\lambda_1) = \sqrt{\frac{\lambda_1}{\lambda_n}} + \sqrt{\frac{\lambda_n}{\lambda_1}} = \theta_t(\lambda_n)$

donc
$$\sqrt{\langle Ax, x\rangle\langle A^{-1}x, x\rangle} \leqslant \frac{1}{2}\left(\sqrt{\frac{\lambda_1}{\lambda_n}} + \sqrt{\frac{\lambda_n}{\lambda_1}}\right)\|x\|^2$$

et on en déduit le résultat en mettant au carré.

Théorème : Soit $x_0 \in \mathbb{R}^n$. Supposons $x_0, x_1, \dots, x_p$ construits pour un certain $p \geqslant 0$.

Si $\nabla F(x_p) = 0$ (où F est définie comme précédemment), on arrête la construction de la suite (x_p).

Sinon, on pose $x_{p+1} = x_p - t_p \nabla F(x_p)$

où $t_p = \dfrac{\|\nabla F(x_p)\|^2}{\langle A\nabla F(x_p), \nabla F(x_p)\rangle}$ est l'unique élément de $\mathbb{R}^n$

qui minimise $t \mapsto F(x_p - t\nabla F(x_p))$.

Alors la suite (x_p) ainsi définie converge vers c et vérifie :

$$\forall p, \quad \|x_p - c\| \leqslant \sqrt{\operatorname{cond}(A)} \left(\frac{\operatorname{cond}(A)-1}{\operatorname{cond}(A)+1}\right)^p \|x_0 - c\|$$

où $\operatorname{cond}(A) = \dfrac{\lambda_n}{\lambda_1}$ est le conditionnement de A.

Démonstration :

* Soit $x \in \mathbb{R}^n$ tel que $\nabla F(x) \neq 0$.

  On note $\tilde{F} : t > 0 \mapsto F(x - t\nabla F(x))$.

  Soit $t > 0$.

$$\tilde{F}(t) = F(x) + \langle \nabla F(x), -t\nabla F(x)\rangle + \frac{1}{2}\langle A t\nabla F(x), t\nabla F(x)\rangle$$
$$= F(x) - t\|\nabla F(x)\|^2 + \frac{t^2}{2}\langle A\nabla F(x), \nabla F(x)\rangle$$

  donc $t_x = \dfrac{\|\nabla F(x)\|^2}{\langle A\nabla F(x), \nabla F(x)\rangle}$ est l'unique élément de $\mathbb{R}_+^*$ qui minimise $\tilde{F}$.

* Soit $p \geqslant 0$ tel que $\nabla F(x_p) \neq 0$.

  Montrons que $\langle A(x_{p+1}-c), x_{p+1}-c\rangle \leqslant \left(\dfrac{\operatorname{cond} A - 1}{\operatorname{cond} A + 1}\right)^2 \langle A(x_p - c), x_p - c\rangle$.

$$\langle A(x_{p+1}-c), x_{p+1}-c\rangle = \langle A(x_p - c) - t_p A\nabla F(x_p), (x_p - c) - t_p\nabla F(x_p)\rangle$$
$$= \langle A(x_p-c), x_p - c\rangle - t_p\langle A(x_p-c), \nabla F(x_p)\rangle$$
$$- t_p\langle A\nabla F(x_p), x_p - c\rangle + t_p^2\langle A\nabla F(x_p), \nabla F(x_p)\rangle$$

Puisque A est symétrique et $A(x_p - c) = Ax_p - Ac = Ax_p - b = \nabla F(x_p)$, on a :

$$\langle A(x_{p+1}-c), x_{p+1}-c\rangle = \langle A(x_p-c), x_p-c\rangle - 2t_p \|\nabla F(x_p)\|^2 + t_p^2 \langle A\nabla F(x_p), \nabla F(x_p)\rangle$$

En remplaçant t_p par son expression, il vient

$$\langle A(x_{p+1}-c), x_{p+1}-c\rangle = \langle \underbrace{A(x_p-c)}_{=\nabla F(x_p)}, \underbrace{x_p-c}_{=A^{-1}\nabla F(x_p)}\rangle - \frac{\|\nabla F(x_p)\|^4}{\langle A\nabla F(x_p), \nabla F(x_p)\rangle}$$

Comme $\nabla F(x_p) \neq 0$:

$$\langle A(x_{p+1}-c), x_{p+1}-c\rangle = \langle A(x_p-c), x_p-c\rangle \underbrace{\left(1 - \underbrace{\frac{\|\nabla F(x_p)\|^4}{\langle A\nabla F(x_p), \nabla F(x_p)\rangle \langle A^{-1}\nabla F(x_p), \nabla F(x_p)\rangle}}_{\geqslant \frac{1}{\frac{1}{4}\left(\sqrt{\frac{\lambda_1}{\lambda_n}}+\sqrt{\frac{\lambda_n}{\lambda_1}}\right)^2} \text{ (lemme)}}\right)}_{\leqslant \left(\frac{\mathrm{cond}(A)-1}{\mathrm{cond}(A)+1}\right)^2}$$

ce qui prouve l'inégalité voulue.

Par récurrence immédiate, on a pour tout p :

$$\langle A(x_p-c), x_p-c\rangle \leqslant \left(\frac{\mathrm{cond}(A)-1}{\mathrm{cond}(A)+1}\right)^{2p} \langle A(x_0-c), x_0-c\rangle$$

Enfin

$$\|x_p - c\|^2 \leqslant \frac{1}{\lambda_1} \langle A(x_p-c), x_p-c\rangle \quad \text{(th. spectral et } \lambda_1 \text{ est la plus petite valeur propre de } A)$$

$$\leqslant \frac{1}{\lambda_1}\left(\frac{\mathrm{cond}(A)-1}{\mathrm{cond}(A)+1}\right)^{2p} \underbrace{\langle A(x_0-c), x_0-c\rangle}_{\leqslant \lambda_n \|x_0-c\|^2} \quad \text{(th. spectral et } \lambda_n \text{ est la plus grande valeur propre de } A)$$

$$\leqslant \mathrm{cond}(A)\left(\frac{\mathrm{cond}(A)-1}{\mathrm{cond}(A)+1}\right)^{2p} \|x_0-c\|^2$$

L'inégalité énoncée dans le théorème s'en déduit en prenant la racine carrée.

Comme $\frac{\mathrm{cond}(A)-1}{\mathrm{cond}(A)+1} \in [0,1[$, la suite (x_p) converge vers c.