TD 8
Résolution numérique d'équations non-linéaires

Correction

Thibaut METIVET

09/04/2021

Exercice 1 : Méthode de la sécante

On s'intéresse dans cet exercice à une méthode de résolution numérique de l'équation générale

$$f(x) = 0 \tag{1}$$

où f est une fonction abitraire supposée C^2 au voisinage des solutions de eq. (1). On suppose également que les solutions de eq. (1) ne sont pas des points singuliers de f, i.e.

$$\forall a \in \mathbb{R}, f(a) = 0 \Rightarrow f'(a) \neq 0. \tag{2}$$

Cette méthode, appelée *méthode de la sécante*, est une version "discrète" de la méthode de Newton ne faisant pas intervenir à chaque itération $k+1$ directement la dérivée $f'(x_k)$ de f en x_k, mais son approximation $T(x_k, x_{k-1})$ avec

$$T(x, y) = \frac{f(x) - f(y)}{x - y}. \tag{3}$$

définie pour $x \neq y$. La méthode s'écrit ainsi

$$x_{k+1} = x_k - \frac{f(x_k)}{T(x_k, x_{k-1})} \tag{4}$$

et définit une suite $(x_k)_{k\geq 1}$ étant données deux conditions initiales x_0 et x_1. On observe que cette suite est d'ordre 2 ; il est donc nécessaire d'augmenter artificiellement la dimension de l'espace d'analyse à $\mathbb{R}^2$ pour la considérer comme une suite d'ordre 1 en $z_k \in \mathbb{R}^2 \equiv (x_k, y_k) \equiv (x_k, x_{k-1})$.

Q1

On a alors

$$\begin{cases} x_{k+1} = x_k - \dfrac{f(x_k)}{T(x_k, x_{k-1})} = x_k - \dfrac{f(x_k)}{T(x_k, y_k)} \\ y_{k+1} = x_k \end{cases} \tag{5}$$

i.e. en définissant $\phi : \mathbb{R}^2 \to \mathbb{R}^2$

$$\phi(x,y) = \left(x - \frac{f(x)}{T(x,y)}, x\right) \equiv (\varphi(x,y), x) \tag{6}$$

la méthode s'écrit $z_{k+1} = \phi(z_k)$.

Q2

La façon dont on a défini ϕ à la Q1 n'est possible que pour $x \neq y$, ce qui s'avère limitant pour l'analyse de la méthode proche d'une solution a quelconque de eq. (1), qui devrait vérifier $\phi(a,a) = (a,a)$.

C'est ici qu'intervient l'hypothèse de régularité de f et la non-singularité des racines de f, qui nous permet de prolonger par continuité T et ϕ en (a,a).

En effet, f étant C^2 sur un voisinage de a, on sait qu'il existe un ouvert $U_a \subseteq \mathbb{R}$ contenant a tel que f soit C^2 sur U_a. On a donc $\forall x \in U_a$, $\forall (h_1, h_2), h_1 \neq h_2$

$$\begin{aligned} T(x+h_1, x+h_2) &= \frac{f(x+h_1) - f(x+h_2)}{h_1 - h_2} \\ &= \frac{f(x+h_1) - f(x)}{h_1 - h_2} + \frac{f(x+h_2) - f(x)}{h_2 - h_1} \\ &= \frac{1}{h_1 - h_2}\left(h_1 \frac{f(x+h_1) - f(x)}{h_1} - h_2 \frac{f(x+h_2) - f(x)}{h_2}\right) \\ &\xrightarrow[(h_1,h_2)\to(0,0)]{} f'(x) \end{aligned} \tag{7}$$

ce qui suggère le prolongement de T en (x,x), $\forall x \in U_a$

$$T(x,x) \equiv f'(x). \tag{8}$$

On a alors bien $T(x+h_1, x+h_2) \xrightarrow[(h_1,h_2)\to(0,0)|h_1\neq h_2]{} T(x,x)$, et dans le cas $h_1 = h_2 = h$,

$$\begin{aligned} T(x+h, x+h) &= f'(x+h) \\ &\xrightarrow[h\to 0]{} f'(x) \quad \text{car } f' \text{ continue} \end{aligned} \tag{9}$$

ce qui montre que le prolongement de T est continu sur U_a.

De plus, f étant C^2 sur U_a, le prolongement de T est même C^1 (on admet cette propriété ici ; elle se démontrerait de la même façon que la continuité de T, mais en étudiant les dérivées partielles).

L'hypothèse $T(a,a) = f'(a) \neq 0$ nous permet également de prolonger ϕ par continuité sur un voisinage de (a,a). Plus précisément, on a $T(a,a) = f'(a) \neq 0$ avec T continue. Il existe donc un voisinage ouvert V_a de a tel que $\forall x \in V_a$, $T(x,x) = f'(x) \neq 0$.

On peut donc définir le prolongement de ϕ sur $(U_a \cap V_a)^2$:

$$\forall x \in U_a \cap V_a,\ \phi(x,x) = \left(x - \frac{f(x)}{f'(x)}, x\right) \tag{10}$$

ce qui nous donnera aussi naturellement un prolongement C^1 comme addition et division de fonctions C^1.

Q3

On peut donc maintenant analyser les propriétés de la méthode en utilisant les propriétés de la matrice jacobienne $D\phi$ en (a, a). On a de façon générale

$$D\phi = \begin{pmatrix} \frac{\partial \varphi}{\partial x} & \frac{\partial \varphi}{\partial y} \\ 1 & 0 \end{pmatrix} \tag{11}$$

et

$$\begin{aligned} \frac{\partial \varphi}{\partial x} &= 1 - \frac{\partial}{\partial x}\left(\frac{f(x)}{T(x,y)}\right) \\ &= 1 - \frac{f'(x)T(x,y) - f(x)\frac{\partial T}{\partial x}(x,y)}{T^2(x,y)} \end{aligned}$$

Evalué en une racine a (i.e. $f(a) = 0$), on obtient alors

$$\frac{\partial \varphi}{\partial x}(a,a) = 1 - \frac{f'(a)}{T(a,a)} = 0 \tag{12}$$

par définition de $T(a, a) = f'(a)$. On a de plus $\forall y, \varphi(a, y) = a$, d'où

$$\frac{\partial \varphi}{\partial y}(a,a) = 0 \tag{13}$$

et finalement

$$D\phi(a,a) = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} \tag{14}$$

Le rayon spectral de $D\phi(a, a)$ est donc $\rho(D\phi(a, a)) = 0 < 1$, ce qui implique que a est asymptotiquement stable d'après un théorème du cours, i.e. pour une condition initiale suffisamment proche de a (i.e. en considérant la norme 1, $\|z_0 - (a, a)\|_1 = |x_0 - a| + |x_1 - a|$ assez petit), la suite (x_k) converge vers a.

Bonus

Etudions maintenant la vitesse de convergence de la méthode de la sécante. On considère donc une solution a de eq. (1) ($f(a) = 0$) telle que $f'(a) \neq 0$, et on introduit la suite des erreurs (e_k) définie par

$$e_k \equiv x_k - a. \tag{15}$$

D'après eq. (4), on a alors

$$\begin{aligned} e_{k+1} &\equiv x_{k+1} - a \\ &= x_k - \frac{f(x_k)(x_k - x_{k-1})}{f(x_k) - f(x_{k-1})} - a \\ &= e_k - \frac{f(x_k)(e_k - e_{k-1})}{f(x_k) - f(x_{k-1})} \\ &= \frac{e_{k-1}f(x_k) - e_k f(x_{k-1})}{f(x_k) - f(x_{k-1})} \end{aligned} \tag{16}$$

où on observe que la suite des erreur est également une suite récurrente d'ordre 2. En supposant que les conditions initiales x_0 et x_1 sont choisies suffisamment proches de a, on sait que la

méthode est convergente, et que pour k assez grand, x_k est donc dans le voisinage de a où f est C^2. On peut alors écrire

$$\begin{aligned} f(x_k) &= f(a+e_k) \\ &= \underbrace{f(a)}_{=0} + e_k f'(a) + \frac{e_k^2}{2} f''(a) + o(e_k^2) \\ &= e_k \left(f'(a) + \frac{e_k}{2} f''(a) \right) + o(e_k^2) \end{aligned} \tag{17}$$

ce qui donne dans l'eq. (16)

$$\begin{aligned} e_{k+1} &= \frac{e_{k-1}e_k \left(f'(a) + \frac{e_k}{2} f''(a) \right) + e_{k-1} o(e_k^2) - e_k e_{k-1} \left(f'(a) + \frac{e_{k-1}}{2} f''(a) \right) e_k o(e_{k-1}^2)}{e_k \left(f'(a) + \frac{e_k}{2} f''(a) \right) + o(e_k^2) - e_{k-1} \left(f'(a) + \frac{e_{k-1}}{2} f''(a) \right) + o(e_{k-1}^2)} \\ &= \frac{e_k e_{k-1} f''(a)(e_k - e_{k-1}) + o(e_{k-1}^3)}{2f'(a)(e_k - e_{k-1}) + o(e_{k-1}^2)} \\ &= \frac{e_k e_{k-1} f''(a) + o(e_{k-1}^3)}{2f'(a) + o(e_{k-1}^2)} \end{aligned} \tag{18}$$

où on a utilisé le fait que $e_k < e_{k-1}$ car la suite converge vers a. On a donc asymptotiquement

$$e_{k+1} = e_k e_{k-1} \frac{f''(a)}{f'(a)} \tag{19}$$

ou encore en prenant le log de la valeur absolue

$$\ln |e_{k+1}| = \ln |e_k| + \ln |e_{k-1}| + \ln \left| \frac{f''(a)}{f'(a)} \right|. \tag{20}$$

On observe alors que la suite $(\ln |e_k|)$ est une suite de Fibonacci, qui possède donc le comportement asymptotique

$$\frac{\ln |e_{k+1}|}{\ln |e_k|} \xrightarrow[k\to\infty]{} \Phi \tag{21}$$

où Φ est le nombre d'or

$$\Phi = \frac{1+\sqrt{5}}{2} \approx 1.618. \tag{22}$$

On a donc asymptotiquement

$$|e_{k+1}| \sim |e_k|^{\Phi} \tag{23}$$

ce qui montre que la méthode de la sécante converge "à la vitesse ϕ", i.e. plus rapidement que le point fixe (qui converge linéairement), mais moins vite que la méthode de Newton qui converge quadratiquement.

Exercice 2

Dans cet exercice, on étudie le comportement de la méthode de Newton appliquée au calcul de l'inverse $\frac{1}{a}$ d'un réel $a > 0$, i.e. à la résolution de l'équation

$$f(x) \equiv \frac{1}{x} - a = 0. \tag{24}$$

Q1

La méthode de Newton s'écrit de façon générale

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

Dans le cas d'une fonction réelle, on a donc

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

Appliquée à $f(x) \equiv \frac{1}{x} - a$ ($f'(x) = -\frac{1}{x^2}$), on obtient donc

$$\begin{aligned} x_{k+1} &= x_k + x_k^2 \left(\frac{1}{x_k} - a \right) \\ &= x_k(2 - a x_k) \equiv \phi(x_k) \end{aligned} \tag{25}$$

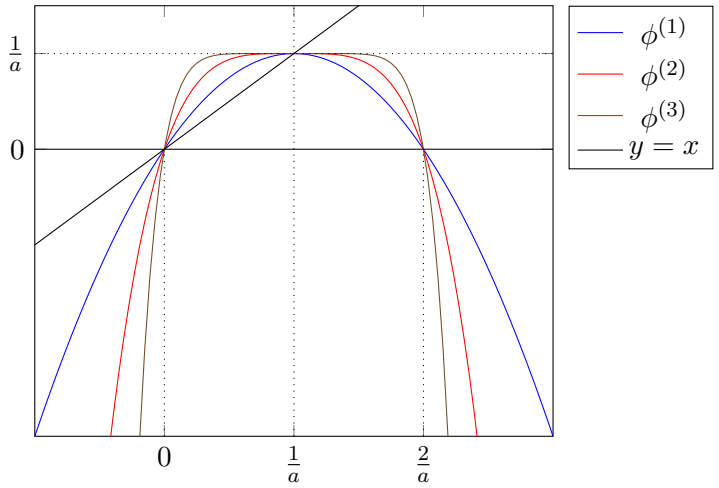
où on a défini la fonction d'itération $\phi(x) \equiv x(2 - ax)$.

Q2

L'analyse de la convergence de la méthode de Newton revient donc l'étude de la convergence de la suite des itérés de ϕ (i.e. la suite $(\phi^{(k)})$ où $\phi^{(k)} \equiv \underbrace{\phi \circ \cdot \circ \phi}_{k \text{ fois}}$), qui dépend naturellement de la condition initiale x_0.

L'étude de ce type de suite de fonction itérée n'admet que peu de résultats généraux, et dépend la plupart du temps de la fonction particulière considérée.

Commençons donc par tracer les premières itérations de la suite.



Comme on peut le voir sur le graphe, la fonction ϕ admet deux points fixes :

$$\begin{aligned} \phi(x) = x &\iff x - ax^2 = 0 \\ &\iff x = 0 \text{ ou } x = \frac{1}{a}. \end{aligned} \tag{26}$$

On sait donc que si la suite converge, alors elle converge vers l'un de ces points fixes. ϕ admet également deux racines :

$$\begin{aligned} \phi(x) = 0 &\iff x(2-ax) = 0 \\ &\iff x = 0 \text{ ou } x = \frac{2}{a} \end{aligned} \tag{27}$$

et on a

$$\phi'(x) = 2(1-ax). \tag{28}$$

Le graphe suggère que l'intervalle $\left[0; \frac{1}{a}\right]$ est stable par ϕ et que l'intervalle $\left[\frac{1}{a}; \frac{2}{a}\right]$ a pour image $\left[0; \frac{1}{a}\right]$, ce qui nous amène à considérer les quatre cas $x_0 \in \left]0; \frac{1}{a}\right]$, $x_0 \in \left[\frac{1}{a}; \frac{2}{a}\right[$, $x_0 \in]-\infty; 0[$ et $x \in \left]\frac{2}{a}; +\infty\right[$, ainsi que les points particuliers $x_0 = 0$ et $x_0 = \frac{2}{a}$.

- $x_0 \in \left]0; \frac{1}{a}\right]$

  On a $\phi'(x) \geq 0$, $\forall x \in \left]0; \frac{1}{a}\right]$, et $\phi(0) = 0$, $\phi\left(\frac{1}{a}\right) = \frac{1}{a}$. ϕ est donc croissante sur $\left]0; \frac{1}{a}\right]$ et

  $$\phi\left(\left]0; \frac{1}{a}\right]\right) = \left]0; \frac{1}{a}\right]. \tag{29}$$

  De plus, $\phi(x) - x = x(1-ax) \geq 0$, $\forall x \in \left]0; \frac{1}{a}\right]$. La suite (x_k) est donc croissante et bornée, ce qui implique qu'elle converge vers un point fixe dans $\left[x_0; \frac{1}{a}\right]$, i.e. *la suite converge vers* $\frac{1}{a}$.

- $x_0 \in \left[\frac{1}{a}; \frac{2}{a}\right[$

  Dans ce cas, on a $\phi'(x) \leq 0$, $\forall x \in \left[\frac{1}{a}; \frac{2}{a}\right[$, et $\phi\left(\frac{2}{a}\right) = 0$, d'où

  $$\phi\left(\left[\frac{1}{a}; \frac{2}{a}\right[\right) = \left]0; \frac{1}{a}\right]. \tag{30}$$

  On a donc $x_1 \equiv \phi(x_0) \in \left[\frac{1}{a}; \frac{2}{a}\right[$ et on se ramène au premier cas, i.e. *la suite converge vers* $\frac{1}{a}$.

- $x_0 \in]-\infty; 0[$

  On a alors ϕ négative, et donc

  $$\phi\left(]-\infty; 0[\right) =]-\infty; 0[. \tag{31}$$

  De plus, on a $\phi(x) - x = x(1-ax) < 0$, $\forall x \in]-\infty; 0[$. La suite est donc strictement décroissante, et sans point fixe dans l'intervalle stable $]-\infty; 0[$. La *suite est donc divergente* (et tend vers $-\infty$).

- $x \in \left]\frac{2}{a}; +\infty\right[$

  On a dans ce cas ϕ strictement négative, d'où $x_1 \equiv \phi(x_0) \in]-\infty; 0[$ De façon similaire au cas précédent, *la suite est donc divergente* (vers $-\infty$).

- $x_0 = 0$

  On a alors $x_k = 0$, $\forall k$, et la *suite est convergente* (constante égale à 0).

- $x_0 = \frac{2}{a}$

  On a $x_1 \equiv \phi\left(\frac{2}{a}\right) = 0$, et on retrouve le cas précédent ($x_k = 0$, $\forall k \geq 1$), i.e. *la suite converge vers* 0.

En résumé, on observe donc que la méthode de Newton ne converge vers la solution voulue ($\frac{1}{a}$) que si la condition initiale appartient à $]0; \frac{2}{a}[$, ce qui illustre une limitation classique de cette méthode, fortement dépendante de son initialisation.

Exercice 3 : Méthode de Broyden

Comme vu dans le cours, la *méthode de Newton* est une méthode très efficace de résolution des équations non-linéaires de la forme $f(x) = 0$, avec une convergence quadratique dès lors que la condition initiale est choisie "assez près" de la solution. Cependant, cette méthode requiert le calcul à chaque itération de l'inverse de la jacobienne de la fonction (ou a minima la résolution d'un système linéaire $Df(x_k)\Delta x_k = -f(x_k)$), ce qui peut s'avérer coûteux (construction de la jacobienne et inversion du système), voire impossible si l'expression analytique de la jacobienne n'est pas disponible. On a vu dans dans l'exercice 1 qu'il était possible d'utiliser une approximation "différences finies" de la jacobienne dans le cas d'une fonction d'une seule variable : la méthode dite *de la sécante*.

Comme nous allons le voir, la *méthode de Broyden* est la généralisation de la méthode de la sécante au cas des fonctions de plusieurs variables, qui consiste donc à approcher à chaque itération la jacobienne par une matrice B_k construite aussi itérativement de manière à satisfaire l'équation de la sécante $B_k(x_{k+1} - x_k) = f(x_{k+1}) - f(x_k)$.

Etant donnée une fonction $f : \mathbb{R}^n \to \mathbb{R}^n$ de classe C^2 au voisinage d'une racine $a \in \mathbb{R}^n$ non-singulière, la méthode de Broyden s'écrit

$$x_{k+1} = x_k - B_k^{-1} f(x_k) \tag{32}$$

avec $B_k \in M_n(\mathbb{R})$ définie itérativement par le choix d'un B_0 et

$$B_{k+1} = B_k + \frac{f(x_{k+1}) s_k^T}{s_k^T s_k} \tag{33}$$

où s_k est l'incrément $s_k \equiv x_{k+1} - x_k$.

Q1

En utlisant l'eq. (33), on obtient

$$\begin{aligned} B_{k+1} s_k &= B_k s_k + \frac{f(x_{k+1}) s_k^T}{s_k^T s_k} s_k \\ &= B_k s_k + f(x_{k+1}) \end{aligned} \tag{34}$$

Or, d'après l'eq. (32), on a $x_{k+1} = x_k - B_k^{-1} f(x_k)$, i.e. $s_k = -B_k^{-1} f(x_k)$, d'où

$$B_k s_k = -f(x_k)$$

et finalement

$$B_{k+1} s_k = f(x_{k+1}) - f(x_k) \tag{35}$$

qui est l'équation de la sécante en dimension arbitraire.

Q2

On introduit dans cette question le produit scalaire de Frobenius (produit scalaire "naturel" entre deux matrices), défini pour deux matrices A et B de mêmes dimensions par

$$A : B \equiv \langle A, B \rangle_F \equiv \mathrm{Tr}(A^T B). \tag{36}$$

On va montrer en utilisant ce produit scalaire que B_{k+1} est le projeté orthogonal de B_k sur l'espace affine

$$\Delta_k \equiv \{B \in M_n(\mathbb{R}), Bs_k = f(x_{k+1}) - f(x_k)\} \tag{37}$$

i.e. $B_{k+1} = \mathrm{P}^{\perp}_{\Delta_k} B_k$.

On observe tout d'abord que $B_{k+1} \in \Delta_k$, d'après l'eq. (35). Il nous faut donc montrer que $B_{k+1} - B_k$ est orthogonal à tous les vecteurs de Δ_k, i.e. $\forall B \in \Delta_k$, $\langle B_{k+1} - B_k, B - B_{k+1} \rangle_F = 0$.

Soit donc $B \in \Delta_k$. On a alors

$$\begin{aligned}
\langle B_{k+1} - B_k, B - B_{k+1} \rangle_F &= \mathrm{Tr}\left((B_{k+1} - B_k)^T (B - B_{k+1}) \right) \\
&= \mathrm{Tr}\left((B_{k+1} - B_k)(B - B_{k+1})^T \right) \\
&= \mathrm{Tr}\left(\frac{f(x_{k+1}) s_k^T}{s_k^T s_k} (B - B_{k+1})^T \right) \quad \text{d'après (33)} \\
&= \mathrm{Tr}\left(\frac{f(x_{k+1})}{s_k^T s_k} \left[\underbrace{(B - B_{k+1}) s_k}_{0} \right]^T \right) \\
&= 0
\end{aligned} \tag{38}$$

où on a utilisé $B \in \Delta_k \Rightarrow Bs_k = f(x_{k+1}) - f(x_k) = B_{k+1} s_k$.

On constate donc que B_{k+1} est bien le projeté orthogonal de B_k sur l'espace des matrices satisfaisant l'équation de la sécante. Cette construction de B_{k+1} est cruciale dans l'algorithme de Broyden, car l'eq. (35) est indéterminée pour $n > 1$, i.e. $\mathrm{card}(\Delta_k) > 1$ pour $n > 1$. La méthode de Broyden fixe donc le choix de la matrice B_{k+1} comme la matrice minimisant la norme de Frobenius $\|B_{k+1} - B_k\|_F$ parmi toutes les matrices satisfaisant (35). Elle assure ainsi une construction "de proche en proche" des approximations de la jacobienne.

Q3

Jusque là, nous avons travaillé avec la définition itérative des matrices de Broyden B_k. Cependant, les matrices nécessaires à l'itération (32) vers la solution de $f(x) = 0$ sont en fait les inverses des matrices de Broyden B_k^{-1}. Afin d'éviter la propagation d'erreurs numériques venant de l'itération (33) dans le calcul de B_k^{-1}, il est souvent préférable de construire directement les inverses des matrices de Broyden par itération.

On peut remarquer que les itérations de Broyden (33) font partie de la classe générale appelée "mise à jour de rang 1" (ou *rank-1 update*), qui correspond à l'opération $A + uv^T$ étant données une matrice A et deux vecteurs u et v. Ces opérations sont dites "de rang 1" car la matrice de "correction" uv^T est de rang 1, et sont très utilisées en calcul numérique en raison de l'existence d'algorithmes très efficaces pour les appliquer, ainsi que pour l'existence de la *formule de Sherman-Morrison*, qui permet d'exprimer l'inverse de cette mise à jour de rang 1 $(A + uv^T)^{-1}$ comme une perturbation de l'inverse A^{-1} : si $1 + v^T A^{-1} u \neq 0$,

$$(A + uv^T)^{-1} = A^{-1} - \frac{A^{-1} u v^T A^{-1}}{1 + v^T A^{-1} u}. \tag{39}$$

Appliquée à l'eq. (33), on obtient alors

$$\begin{aligned} B_{k+1}^{-1} &= \left(B_k + \frac{f(x_{k+1}) s_k^T}{s_k^T s_k} \right)^{-1} \\ &= B_k^{-1} - \frac{B_k^{-1} \frac{f(x_{k+1})}{s_k^T s_k} s_k^T B_k^{-1}}{1 + s_k^T B_k^{-1} \frac{f(x_{k+1})}{s_k^T s_k}} \\ &= B_k^{-1} - \frac{B_k^{-1} f(x_{k+1})}{s_k^T s_k + s_k^T B_k^{-1} f(x_{k+1})} s_k^T B_k^{-1}. \end{aligned} \tag{40}$$

Q4

Nous devons donc calculer à chaque itération de la méthode de Broyden

1. $s_k = -B_k^{-1} f_k$
   ↳ 1 produit matrice-vecteur: $O(n^2)$ opérations.

2. $x_{k+1} = x_k + s_k$
   ↳ 1 addition vecteur-vecteur: $O(n)$ opérations.

3. $f_{k+1} = f(x_{k+1})$
   ↳ 1 évaluation de f.

4. B_{k+1}^{-1} avec (40)
   ↳ 2 produits matrice-vecteur, 1 produit vecteur-vecteur, 2 produits scalaires
   $\rightarrow O(n^2)$ opérations.

On a donc finalement pour chaque itération $O(n^2)$ opérations, ainsi qu'une évaluation de f, ce qui est généralement moins coûteux que la méthode de Newton, qui requiert essentiellement le calcul de f, de Df, et la résolution d'un système linéaire, soit en général $O(n^3)$ opérations si Df n'a pas de propriétés particulières.