Chp. 6. Algorithmes de gradient

Avertissement! *Dans tout ce chapître, Ω désigne un ouvert de $\mathbb{R}^n$, et f une fonction de classe $\mathcal{C}^1$ sur Ω.*

6.1 Algorithme de gradient à pas fixe

L'algorithme du gradient à pas fixe est une méthode de descente utilisant un pas fixe et la stratégie de Cauchy pour le choix de la direction de descente :

```
GradFix(f,x_0, pas, tolerance)
x ← x_0
Tant que : || ∇f(x) || > tolerance
    x ← x − pas ⋆ ∇f(x)
 Retourner x
```

6.2 Théorème de convergence

Théorème 6.1

Supposons vérifiées les hypothèses (H_1) à (H_3) suivantes :

(H_1) *$S_0 = \{ x \in A \mid f(x) \leq f(x_0) \}$ est un fermé de $\mathbb{R}^n$ strictement contenu dans Ω.*

(H_2) *f est $\mathcal{C}$ [2] sur Ω et, pour tout x dans S_0 : $0 < c\,Id \leq \nabla^2 f(x) \leq K\,Id$*

(H_3) $0 < \texttt{pas} < 2/K$

Alors l'algorithme `GradFix` *converge vers un minimum local non dégénéré $x^\star$ de f dans Ω. On établit en outre les propriétés suivantes :*

- *La suite $f_k = f(x_k)$ des valeurs du critère aux points x_k calculés par l'algorithme est strictement décroissante.*
- *Pour tout indice $k \geq 0$, l'intervalle $[x_k, x_{k+1}]$ reste tout entier contenu dans l'ensemble : $S_k = \{ x \in \Omega \mid f(x) \leq f(x_k) \}$ de niveau $f(x_k)$ de f.*
- $\| x_0 - x^\star \| \leq c^{-1} \| \nabla f(x_0) \|$

La dernière propriété garantit en particulier que le dernier point x_k calculé par l'algorithme vérifie :

$$\| x_k - x^\star \| \leq c^{-1}\,\texttt{tolerance}$$

La démonstration de ce théorème fait l'objet du paragraphe 6.4 suivant.

Corollaire 6.2 *Si α est un niveau non critique de f tel que :*
(H_1') *L'ensemble : $S_\alpha = \{ x \in \Omega \mid f(x) \leq \alpha \}$ est non vide et compact.*
(H_2') *$\nabla^2 f(x)$ est D.P. en tout point x de l'intérieur $\overset{\circ}{S}_\alpha = \{ x \in \Omega \mid f(x) < \alpha \}$ de S_α.*
alors l'algorithme `GradFix` *converge pour toute initialisation x_0 dans S_α et tout pas suffisamment petit vers un minimum local non dégénéré de f.*

Preuve :

- Si $f(x_0) < \alpha$, $S_0 = \{ x \in \Omega \mid f(x) \leq f(x_0) \}$ est un fermé de $\mathbb{R}^n$ contenu dans S_α. C'est donc un compact, et, par continuité du Hessien, il existe des constantes c et K telles qu'en tout point de S_0 : $0 < c\,Id \leq \nabla^2 f(x) \leq K\,Id$. Ainsi les hypothèses (H_1) et (H_2) du théorème de convergence sont vérifiées.
- Si $f(x_0) = \alpha$, $u_O = -\nabla f(x_0)$ est une direction de descente au point x_0 et le premier point x_1 calculé par l'algorithme vérifiera : $f(x_1) < \alpha$: on est alors ramené au cas précédent.

□

Les hypothèses (H_1') et (H_2') du corollaire 6.2 sont en particulier vérifiées dès que S_α est non vide et Ω est un bassin d'ellipticité de f.

Si f est elliptique sur $\mathbb{R}^n$, elle atteint son minimum en un point unique $x^\star$. L'intérieur de tout ensemble de niveau : $\alpha > f(x^\star)$ est un bassin d'ellipticité de f, et l'algorithme converge, pour toute initialisation x_0, vers $x^\star$, pourvu que le pas soit choisi suffisamment petit :

Exemple 6.1 *La matrice Hessienne de : $f = x^2 + 2\,y^2$ est la matrice diagonale constante : $Q = \begin{pmatrix} 2 & 0 \\ 0 & 4 \end{pmatrix}$ qui vérifie : $0 < 2\,Id \leq Q \leq 4\,Id$. L'algorithme* `GradFix` *converge donc, pour toute initialisation (x_0, y_0) dans $\mathbb{R}^2$ vers l'unique minimum $(0,0)$ de f sur $\mathbb{R}^2$ dès que le pas choisi est (strictement) inférieur à 0.5.*

6.3 Choix du pas et vitesse de convergence

Lorsque $f = \frac{1}{2}\, x^T Q\, x + b^T x + c$ est une fonction quadratique elliptique, sa Hessienne Q est une matrice constante D.P.. La vitesse de convergence de l'algorithme `GradFix` est alors toujours linéaire, et le taux de convergence est optimal pour :

- `pas` $= \dfrac{2}{c+K}$

où c et K sont respectivement la plus petite et la plus grande des valeurs propres de Q.

Le taux de convergence optimal est :

- $\dfrac{K-c}{K+c} = \dfrac{\chi - 1}{\chi + 1}$

où : $\chi = K/c$ est le *conditionnement* de la matrice Q. Lorsque Q est mal conditionnée ($\chi \gg 1$) l'algorithme est lent. Lorsque : $\chi = 1$ au contraire (Q est un multiple de la matrice identité) l'algorithme calcule le minimum cherché en une seule itération.

Lorsque f n'est pas quadratique, mais satisfait les hypothèses (H_1) et (H_2) du théorème 6.1, la démonstration de ce théorème (voir paragraphe 6.4) montre que le taux de convergence reste inférieur à : $\dfrac{\chi-1}{\chi+1}$, où : χ est le conditionnement de la Hessienne de f, calculée au point $x^\star$ vers lequel l'algorithme converge. Le cas quadratique montre que cette majoration du taux est optimale.

Attention! *En pratique, on ignore le plus souvent même l'ordre de grandeur des valeurs de c et de K. Si le pas choisi est trop grand, l'algorithme peut diverger. S'il est trop petit, la convergence peut être extrêmement lente :*

Exemple 6.2 *Le tableau suivant donne le nombre d'itérations nécessaires pour approcher le minimum* $(0,0)$ *de la forme quadratique elliptique :* $f = x^2 + 2\,y^2$ *sur* $\mathbb{R}^2$ *à* 10^{-6} *près, à partir de l'initialisation :* $(x_0, y_0) = (1,1)$*, en fonction du pas choisi :*

pas	0.5	0.45	0.4	0.33	0.1	0.01
	divergence	60	30	13	60	685

Le taux de convergence optimal 1/3 est obtenu pour un pas égal à 1/3.

Exemple 6.3 *La Hessienne de la forme quadratique* $f = x^2 + 100\,y^2$ *est mal conditionnée :* $\chi = 100$*. Le taux de convergence optimal est obtenu pour un pas égal à :* $1/101 \simeq 0.0099$*. Pour cette valeur du pas, il faut près de 700 itérations pour approcher le minimum* $(0,0)$ *de* f *à* 10^{-6} *près. Si le pas est supérieur à* 0.01*, l'algorithme diverge.*

6.4 Démonstration du théorème de convergence

Tout point critique de f contenu dans S_0 est nécessairement un minimum local non dégénéré de f. Il suffit donc d'établir qu'étant donné un point non critique x_k $(k \geq 0)$ de S_0, le point $x_{k+1} = x_k - \mathtt{pas} \star \nabla f(x_k)$ est tel que :

(1) $\quad f_{k+1} < f_k, \quad \text{et :} \quad [x_k, x_{k+1}] \in S_k$

et, en supposant la suite x_k infinie, qu'elle converge toujours vers un point critique $x^\star$ de f.
On commence par établir (1) :

Posons $u_k = -\nabla f(x_k)$, et : $\varphi(t) = f(x_k + t\,u_k)$. Puisque, par hypothèse : $\varphi'(0) = -\,\|\,u_k\,\|^2 < 0$, $x_k + t\,u_k$ reste contenu dans S_k pour $t > 0$ suffisamment petit.

Tant que l'intervalle $[\,x_k, x_k + t\,u_k]$ reste contenu dans S_k, et donc, a fortiori, dans S_0, on déduit de la formule de Taylor-Lagrange appliquée à : $\varphi(t) = f(x_k + t\,u_k)$ sur l'intervalle $[\,0, t\,]$:

$$\varphi(t) = \varphi(0) + t\,\varphi'(0) + \frac{t^2}{2}\,\varphi''(\theta)$$

où : $\theta \in]\,0,1\,[$, soit :

$$\begin{aligned} f(x_k + T\,u_k) &= f(x_k) - t\,\|\,u_k\,\|^2 + \frac{t^2}{2}\,u_k \nabla^2 f(x + \theta\,u_k)\,u_k \\ &\leq f(x_k) - t\,\|\,u_k\,\|^2 + \frac{t^2}{2}\,K\,\|\,u_k\,\|^2 \end{aligned}$$

mais le second membre reste strictement inférieur à $f(x_k)$ pour : $0 < t < 2/K$. L' assertion (1) est donc une conséquence directe de l'hypothèse (H_3).

Il reste à montrer, en supposant la suite des points x_k calculés par l'algorithme infinie, qu'elle converge vers un point critique $x^\star$ de f. Pour cela, on commence par établir les deux lemmes suivants :

Lemme 6.1 *Pour toute $n \times n$ matrice symétrique S vérifiant : $\alpha\, Id \leq S \leq \beta\, Id$, et tout vecteur u de $\mathbb{R}^n$: $\| S\, u \| \leq \max(\, |\alpha|\, , |\beta|\,)\, \| u \|$.*

Preuve : Par hypothèse, les valeurs propres de S^2, qui sont les carrés des valeurs propres de S, sont toutes contenues dans l'intervalle $[\, 0, \gamma^2\,]$, où : $\gamma = \max(|\, \alpha\, |\, , |\, \beta\, |\,)$.

On en déduit : $\| S\, u \|^2 = u^T S^T S\, u \leq \gamma^2\, \| u \|^2$, d'où le lemme.

□

Lemme 6.2 *Soient g une fonction de classe $\mathcal{C}^2$ sur un ouvert Ω, et $[\, x, y\,]$ un intervalle contenu dans Ω. Si, en tout point z de $[\, x, y\,]$: $\alpha\, Id \leq \nabla^2 g(z) \leq \beta\, Id$, alors :*

$$\| \nabla g(x) - \nabla g(y) \| \leq \max(\, |\alpha|\, , |\beta|\,)\, \| x - y \|$$

Preuve : Posons $u = y - x$ et : $\varphi(t) = \| \nabla g(x + t\, u) - \nabla g(x) \|^2$. On vérifie que φ est dérivable en tout point de $[\, 0, 1\,]$, et :

$$\varphi'(t) = 2\, [\nabla g(x + t\, u) - \nabla g(x)]^T\, \nabla^2 g(x + t\, u)\, u$$

De l'inégalité de Cauchy-Schwarz, on déduit alors :

$$\forall t \in [\, 0, 1\,] \quad \varphi'(t) \leq 2\, \sqrt{\varphi(t)}\, \left\| \nabla^2 g(x + t\, u)\, u \right\|$$

Puisque $\nabla^2 g(x + t\, u)$ est une matrice symétrique vérifiant, par hypothèse : $\alpha\, Id \leq \nabla^2 g(x + t\, u) \leq \beta\, Id$, le lemme 6.1 implique :

$$\forall t \in [\, 0, 1\,] \quad \varphi(t) \neq 0 \;\Rightarrow\; \frac{\varphi'(t)}{2\, \sqrt{\varphi(t)}} \leq \max(\, |\alpha|\, , |\beta|\,)\, \| u \|$$

d'où, en intégrant de 0 à 1 :

$$\| \nabla g(y) - \nabla g(x) \| = \sqrt{\varphi(1)} - \sqrt{\varphi(0)} \leq \max(\, |\alpha|\, , |\beta|\,)\, \| y - x \|$$

□

On pose alors : $g(x) = \frac{1}{2} \| x \|^2 - \texttt{pas} \star f(x)$, de sorte que : $\nabla g(x) = Id - \texttt{pas} \star \nabla f(x)$, et on remarque que g satisfait les hypothèses du lemme 6.2 sur tout intervalle $[x, y]$ contenu dans S_0 avec :

$$\alpha = 1 - \texttt{pas} \star K, \quad \text{et :} \quad \beta = 1 - \texttt{pas} \star c$$

En appliquant le lemme 6.2 sur $[x_k, x_{k+1}]$, on déduit :

$$\| x_{k+2} - x_{k+1} \| = \| \nabla g(x_{k+1}) - \nabla g(x_k) \| \leq \gamma \| x_{k+1} - x_k \|$$

où : $\gamma = \max(|1 - \texttt{pas} \star K|, |1 - \texttt{pas} \star c|)$, et, par récurrence : $\| x_{k+1} - x_k \| \leq \gamma^k \| x_1 - x_0 \|$ pour tout indice $k \geq 0$.

Mais l'hypothèse (H_3) du théorème 6.1 implique : $\gamma < 1$ (le vérifier). En sommant, il vient :

$$k < l \;\Rightarrow\; \| x_l - x_k \| \leq \sum_{m=k}^{l-1} \gamma^m \| x_1 - x_0 \| \leq \frac{\gamma^k}{1-\gamma} \| x_1 - x_0 \|$$

qui montre que x_k est une suite de Cauchy de points de S_0, donc converge vers un point $x^\star$ de S_0.

Finalement : $\| x_{k+1} - x_k \| = \texttt{pas} \star \| \nabla f(x_k) \|$ implique, par continuité du gradient : $\nabla f(x^\star) = 0$, et, lorsque : $\texttt{pas} = \frac{2}{K+c}$, alors : $\gamma = \frac{K-c}{K+c}$, et :

$$\| x_0 - x^\star \| \leq \frac{1}{1-\gamma} \| x_1 - x_0 \| = \frac{\texttt{pas}}{1-\gamma} \| \nabla f(x_0) \| = c^{-1} \| \nabla f(x_0) \|$$

ce qui achève la démonstration.

□

6.5 Algorithme de gradient à pas optimal

L'algorithme du gradient à pas optimal combine la stratégie de Cauchy pour la détermination de la direction de descente avec, à chaque étape, une recherche du pas optimal minimisant : $\varphi(t) = f(x + t\,u)$, où : $u = -\nabla f(x)$ est la direction de descente au point x :

```
GradOpt(f, x_0, pas, tolerance)
x ← x_0
Tant que : ‖∇f(x)‖ > tolerance
  u = −∇f(x)
  Calculer le pas optimal t*
  x ← x + t*⋆ u
Retourner x
```

La recherche du pas optimal minimise $\varphi(t)$ en deux étapes. La première est la phase de « bracketing » : elle détermine un intervalle $[0, T]$ contenant le pas optimal $t^\star$:

```
StepBracket(φ)
T ← 1
Tant que : φ(T) < φ(0)
  T ← 2 T
Retourner T
```

Les conditions : $\varphi(T) \geq \varphi(0)$ et : $\varphi'(0) = \nabla f(x)^T u = -\| u \|^2 < 0$ garantissent alors que φ atteint son minimum en un point $t^\star$ de l'intervalle $[\,0, T\,]$.

La seconde étape est la phase de *recherche linéaire* : elle utilise une procédure de minimisation unidimensionnelle (`Goldensearch`, `Quadsearch`, ou `Newtonsearch` par exemple) pour déterminer une valeur approchée de $t^\star$ (chp. 4).

6.6 Convergence de l'algorithme de gradient à pas optimal

Théorème 6.3 *Si Ω est un bassin d'ellipticité de f, l'algorithme* `GradOpt` *converge, pour toute initialisation x_0 dans Ω vers l'unique minimum local $x^\star$ de f dans Ω, qui minimise f sur Ω. On a en outre les propriétés suivantes :*

- *La suite des valeurs : $f_k = f(x_k)$ du critère aux points x_k construits par l'algorithme est strictement décroissante.*
- *Pour tout indice $k \geq 0$: $\| x_k - x^\star \| \leq c^{-1} \| \nabla f(x_k) \|$, où c est une constante d'ellipticité de f sur l'ensemble $S_k = \{\, x \in \Omega \mid f(x) \leq f(x_k) \,\}$ de niveau f_k de f dans Ω.*

Preuve : Par construction, la suite f_k est strictement décroissante. La suite infinie des points x_k éventuellement construite par l'algorithme reste donc, à partir de tout rang k, contenue dans l'ensemble convexe compact S_k. Puisque f est, par hypothèse, elliptique sur S_k, il existe des constantes c et K telles qu'en tout point x de S_k : $c\,Id \leq \nabla^2 f(x) \leq K\,Id$. Les hypothèses (H_1) et (H_2) du théorème 6.1 sont donc satisfaites. On en déduit, pour tout indice $k \geq 0$:

$$(2) \qquad \| x_k - x^\star \| \leq c^{-1} \| \nabla f(x_k) \| \quad \text{et :} \quad 0 < t < 2/K \;\Rightarrow\; [\,x_k,\, x_k - t\,\nabla f(x_k)\,] \subset S_0$$

Mais en appliquant alors la formule de Taylor-Lagrange à : $\varphi(t) = f\,[\,x_k - t\,\nabla f(x_k)\,]$ sur l'intervalle $[\,0, 1/K\,]$, il vient :

$$f\left[\, x_k - \frac{1}{K}\,\nabla f(x_k) \,\right] \leq f(x_k) - \frac{1}{K} \| \nabla f(x_k) \|^2 + \frac{1}{2\,K} \| \nabla f(x_k) \|^2$$

d'où, par définition du pas optimal :

$$(3) \qquad f_{k+1} \leq f_k - \frac{1}{2\,K} \| \nabla f(x_k) \|^2$$

Finalement, la suite f_k est décroissante et minorée par $f(x^\star)$, donc convergente, et (3) implique alors la convergence de $\nabla f(x_k)$ vers 0, d'où, d'après (2), la convergence de x_k vers $x^\star$.

□

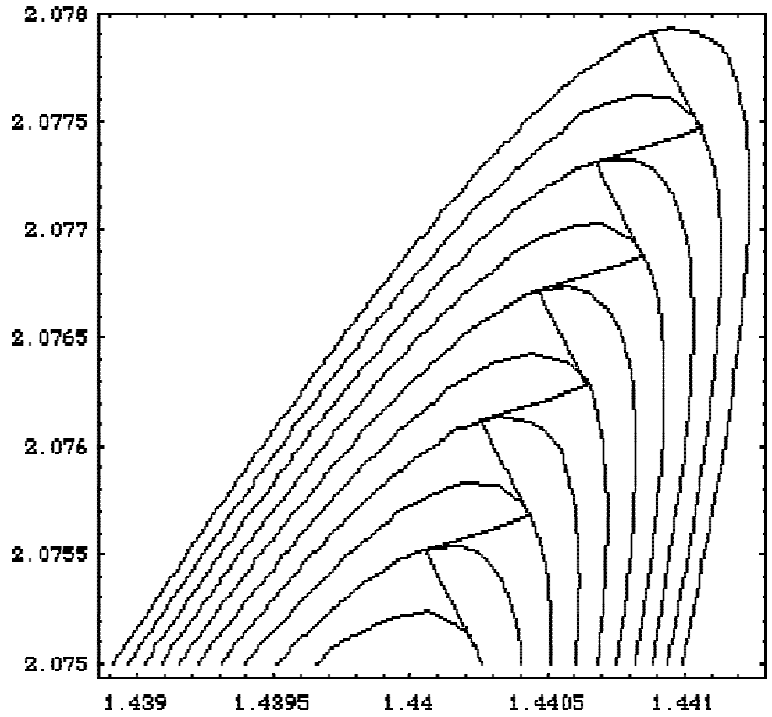

FIG. 6.7-1: *Zig-zags caractéristiques de l'algorithme* `GradOpt`

6.7 Comportement de l'algorithme de gradient à pas optimal

A la $k+1^{ème}$ itération, l'algorithme `GradOpt` minimise la fonction : $\varphi(t) = f(x_k + t\,u_k)$, où : $u_k = -\nabla f(x_k)$. Si t_k est le pas optimal calculé, on a donc :

$$\varphi'(t_k) = \nabla f(x_k + t_k u_k)^T u_k = -\nabla f(x_k + t_k u_k)^T \nabla f(x_k) = 0$$

et le point x_{k+1} vérifie : $\nabla f(x_{k+1})^T \nabla f(x_k) = 0$.

Deux directions de descente successives calculées par l'algorithme sont ainsi *orthogonales* . La figure 6.7-1 illustre les zig-zags caractéristiques correspondants de l'algorithme `GradOpt`.

6.8 Comparaison avec l'algorithme de gradient à pas fixe

En pratique, l'algorithme de gradient à pas optimal s'avère souvent plus efficace que l'algorithme de gradient à pas fixe :

Exemple 6.4 *Le paramètre* `tolerance` *de la procedure* `GoldenSearch`*, utilisée pour la phase de rechereche linéaire, étant fixé à* 10^{-8}*, il ne faut que 6 itérations à l'algorithme* `GradOpt`*, initialisé avec :* $(x_0, y_0) = (1, 1)$*, pour approcher le minimum* $(0, 0)$ *de :* $f = x^2 + 100\,y^2$ *à* 10^{-6} *près (comparer avec l'utilisation catastrophique de* `GradFix` *dans l'exemple 6.3).*

Il est cependant difficile de comparer objectivement les deux algorithmes, et le contre-exemple suivant montre l'impossibilité d'établir théoriquement, et pour *toute* initialisation donnée, la supériorité de l'algorithme du gradient à pas optimal, même lorque le critère est une forme quadratique elliptique simple :

Contre-exemple 6.5 *Si l'on cherche à minimise :* $f = x^2 + 2\,y^2$ *à partir de toute initialisation* (x_0, y_0) *située sur la droite d'équation :* $x = 2\,y$*, en utilisant l'algorithme* `GradOpt`*, le pas optimal effectué à chaque étape est constant égal à* $1/3$*, et la vitesse de convergence linéaire de taux :* $1/3$*.*

Contrairement à l'algorithme du gradient à pas fixe, cependant, l'algorithme du gradient à pas optimal ne requiert aucun encadrement *a priori* , jamais disponible en pratique, des valeurs propres de la Hessienne du critère, et garantit, par nature, la décroissance du critère. C'est donc un algorithme à la fois simple à mettre en oeuvre et robuste.

6.9 Règle d'Armijo

Lorsque l'évaluation du critère en un point est compliquée, la procédure de recherche linéaire utilisée par l'algorithme de gradient à pas optimal peut se révéler coûteuse. L'idée suggérée par Armijo consiste à chercher un pas qui permette simplement de faire décroître *suffisamment* la valeur du critère. On se donne un réel α strictement compris entre 0 et 1, et on cherche un pas t vérifiant :

$$f(x + t\,u) < f(x) + \alpha\,t\,\nabla f(x)^T u$$

Théorème 6.4 *Si :* $\nabla f(x)^T u < 0$*, il existe un* $t^\star$ *tel que :*

$$0 < t < t^\star \;\Rightarrow\; f(x + t\,u) < f(x) + \alpha\,t\,\nabla f(x)^T u$$

Preuve : Posons : $\psi(t) = f(x + t\,u) - f(x) - \alpha\,t\,\nabla f(x)^T u$, de sorte que : $\psi(0) = 0$, et :

$$\psi'(0) = (1 - \alpha)\,\nabla f(x)^T u < 0$$

Pour $t > 0$ suffisamment petit, on a donc : $\psi(t) < 0$, d'où le résultat.

□

6.10 Recherche linéaire rapide

La procédure `LinearSearch` implémente la règle d'Armijo :

`LinearSearch(`f`, `x`, `u`, `α`)`
$t \leftarrow 1$
`Tant que :` $f(x + t\,u) \geq f(x) + \alpha\,\nabla f(x)^T u$

$$t \leftarrow \frac{-t\,\nabla f(x)^T u}{2\,[f(x + t\,u) - f(x) - t\,\nabla f(x)^T u]}$$

`Retourner` $x + t\,u$

Théorème 6.5 *Si :* $\alpha < 1/2$*,* `LinearSearch` retourne toujours un pas vérifiant la règle d'Armijo.

Preuve : Si : $f(x + t\,u) < f(x) + \alpha\,t\,\nabla f(x)^T u$, on accepte le pas : $t = 1$. Sinon, on interpole $\varphi(t) = f(x + t\,u)$ par la parabole passant par les deux points $(0, f(x))$ et $(t, f(x + t\,u))$ dont le coefficient directeur de la tangente en $(0, f(x))$ est : $\varphi'(0) = \nabla f(x)^T u$, et on actualise t à la valeur de l'abscisse du minimum de cette parabole. On vérifie que, si : $\nabla f(x)^T u < 0$, cette stratégie d'actualisation de t conduit, à chaque étape, à remplacer t par un réel de l'intervalle : $]\,0, \frac{1}{2}\,\frac{t}{1-\alpha}\,[$. Pour : $\alpha < 1/2$, on réduit strictement la taille de l'intervalle $]\,0, t\,[$ à chaque étape. Le résultat est donc conséquence du théorème 6.4.

□