Méthodes numériques : optimisation.
L3 2016–2017 — 2e semestre.
Feuille de TD n° 2 : Méthodes de descente de gradient — Éléments de correction.

1 Retour en dimension un.

(a) Sachant qu'ici le gradient en x_k est $f'(x_k)$, les deux méthodes peuvent être vues sous la forme $x_{k+1} = x_k - \alpha_k f'(x_k)$.

Pour la méthode de Newton : $x_{k+1} = x_k - \frac{f'(x_k)}{f''(x_k)}$ on a $\alpha_k = \frac{1}{f''(x)}$, et pour la méthode de la sécante $x_{k+1} = x_k - \frac{(x_{k-1}-x_k)f'(x_k)}{f'(x_{k-1})-f'(x_k)}$ donc $\alpha_k = \frac{x_{k-1}-x_k}{f'(x_{k-1})-f'(x_k)}$ (dans ce cas-là, il faut avoir deux points de départ $x_0 \neq x_1$ pour que cela fonctionne).

Pour que la méthode soit vraiment une direction de descente, il faut que le pas soit positif, et que l'on ait $f(x_k + \alpha_k f'(x_k)) < f(x_k)$.

On peut essayer d'aller plus loin. Pour que le pas soit positif, une condition suffisante est que la fonction soit (un peu plus que) strictement convexe : si $f''(x) > 0$ pour tout x, alors le pas de la méthode de Newton est strictement positif. Si f' est strictement croissante, alors le pas de la méthode de la sécante est strictement positif (dans le cas où on a toujours $x_k \neq x_{k-1}$, ce que l'on peut supposer vrai, car si à une étape on a $x_k = x_{k-1}$, c'est qu'on a eu $f'(x_{k-1}) = 0$, et dans ce cas on arrête l'algorithme, puisqu'on a trouvé un point critique et qu'on ne peut pas calculer le point suivant, le dénominateur étant nul).

Pour que la deuxième condition soit satisfaite, il existe des conditions suffisantes, par exemple que x_k soit suffisamment proche d'un minimum x_* où $f''(x_*) > 0$, mais c'est plus compliqué à montrer (on ne s'attend pas à cette réponse).

(b) On a $x_{k+1} = x_k - 2\alpha x_k + \alpha x_k^3 = x_k(1 - 2\alpha + \alpha x_k^2)$. Pour $\alpha \in]0, 1[$, on a si x_0^2 est suffisamment petit avec $\rho = |1 - 2\alpha + \alpha x_0^2| < 1$, que $|x_k|$ est décroissante (donc $f(x_k)$ aussi) et que $|x_{k+1}| \leqslant \rho |x_k|$ donc la suite converge au moins linéairement vers 0.

Le taux de convergence linéaire est donné par $\lim_{k\to\infty} \frac{|x_{k+1}-0|}{|x_k-0|} = \lim_{k\to\infty} |1 - 2\alpha + \alpha x_k^2| = |1 - 2\alpha|$.

La convergence est donc superlinéaire quand ce taux est nul, c'est à dire pour $\alpha = \frac{1}{2}$. Essayons d'obtenir son ordre : on a dans ce cas-là $x_{k+1} = \frac{1}{2}x_k^3$, donc

on obtient que $\lim_{k\to\infty} \frac{|x_{k+1}-0|}{|x_k-0|^3} = \frac{1}{2}$. Comme la suite converge vers 0, on dit que l'ordre de convergence est 3. Bien remarquer que le fait que $\frac{1}{2} < 1$ n'est aucunement important, on aurait pu avoir 12 ça aurait très bien convenu.

Pour le cas $\alpha = 1$ c'est un peu plus subtil : on a $x_{k+1} = -x_k(1 - x_k^2)$ de sorte que $|x_k|$ est décroissante dès que $|x_0^2| < 2$. Elle converge donc vers une limite ℓ. En passant à la limite dans la relation de récurrence, on obtient que $\ell = \ell|1-\ell^2|$. Comme $\ell \geqslant 0$ on obtient que $\ell \in \{0, \sqrt{2}\}$. Comme on a démarré avec $|x_0| < \sqrt{2}$, on est assuré de la convergence vers 0.

Par contre dans ce cas là, on a $\lim_{k\to\infty} \frac{|x_{k+1}|}{|x_k|} = 1$, donc la convergence n'est pas linéaire.

Enfin dès que $\alpha > 1$, on obtient $|x_{k+1}| > |x_k|$ dès que $|x_k|$ est assez petit (par exemple dès que $|x_k| \leqslant \varepsilon$). Donc, à moins d'être constante égale à 0 à partir d'un certain rang, la suite x_k ne converge pas vers 0 : par l'absurde, il existerait un N_0 tel que pour tout $k \geqslant N_0$, on ait $|x_k| \leqslant \varepsilon$, ce qui donnerait que la suite $|x_k|$ serait croissante à partir du rang N_0.

On peut montrer (mais ce n'est pas demandé) que le cas où la suite stationne à zéro n'arrive presque jamais : si k_0 est le premier endroit où $x_k = 0$, alors $x_{k-1} = \pm\sqrt{2 - \frac{1}{\alpha}}$, puis x_{k-2} n'a qu'au plus trois solutions (on résout une équation de degré 3), ainsi de suite, si on remontait en arrière, on n'a qu'un nombre fini de solutions pour x_0, et ce pour chaque k_0. L'ensemble des points à partir duquel la suite peut stationner à 0 est donc dénombrable, et donc de mesure nulle.

(c) * En supposant que $0 < |x_0| < \sqrt{2}$, la suite $(x_k^2)_{k\in\mathbb{N}}$ est décroissante d'après la question **(b)** et on a

$$\frac{1}{x_{n+1}^2} - \frac{1}{x_n^2} = \frac{1}{x_n^2}\Big(\frac{1}{(1-x_n^2)^2} - 1\Big) = \frac{1}{x_n^2}\frac{2x_n^2 - x_n^4}{(1-x_n^2)^2} = \frac{2 - x_n^2}{(1-x_n^2)^2} \to 2.$$

En faisant la somme de $n = 0$ à $N-1$, les termes se téléscopent, et en divisant par N, on obtient

$$\frac{1}{Nx_N^2} = \frac{1}{N}\left(\frac{1}{x_0^2} + \sum_{n=0}^{N} a_n\right),$$

où $a_n = \frac{2-x_n^2}{(1-x_n^2)^2} \to 2$. On peut donc montrer (théorème des moyennes de Cesáro) que $Nx_N^2 \to 2$, autrement dit que $x_N \sim \frac{1}{\sqrt{2N}}$, et ce quelque soit $x_0 \in]-\sqrt{2}, \sqrt{2}[\setminus\{0\}$.

(d) On a $f(x) = 1 - \frac{1}{1+x^2}$. Donc $x_{k+1} = x_k - \alpha\frac{2x_k}{(1+x_k^2)^2} = x_k(1 - \frac{2\alpha}{(1+x_k^2)^2})$.

Si $\alpha \in]0, 1]$ et $x_0 \neq 0$, alors on a bien toujours $-1 < 1 - \frac{2\alpha}{(1+x_k^2)^2} < 1$, et on obtient que la suite $(|x_k|)_{k\in\mathbb{N}}$ est décroissante, et converge donc vers une limite ℓ telle que $\ell = \ell|1 - \frac{2\alpha}{(1+\ell^2)^2}|$ donc $\ell = 0$ ou $|1 - \frac{2\alpha}{(1+\ell^2)^2}| = 1$, qui n'a de solution non nulle que si $\alpha > 1$ (on obtient $\alpha = (1+\ell^2)^2$, soit $\ell = \sqrt{\sqrt{\alpha} - 1}$). Si on prend x_0 suffisamment petit, la suite x_k converge donc vers 0. De même on montre que le taux de convergence est linéaire est alors $|1 - 2\alpha|$ (la suite ne converge donc pas linéairement pour $\alpha = 1$).

Si $\alpha > 1$, lorsque $x_k \neq 0$ pour tout k, alors la suite ne peut pas converger vers 0 pour les mêmes raisons que dans la question **(b)**.

Encore une fois, pour k_0 fixé, on peut montrer (mais ce n'est pas demandé) que pour une suite telle que $x_{k_0} = 0$, il n'y a qu'un nombre fini de cas possibles pour x_0.

Si l'on prend $\alpha \in]0, 1]$ mais x_0 trop grand, le facteur $(1 - \frac{2\alpha}{(1+x_k^2)^2})$ sera très proche de 1 pour les premières itérations, et la décroissance de x_k va être très lente initialement. On ne verra pas numériquement le taux de convergence linéaire de $|1 - 2\alpha|$, il faut attendre que x_k soit suffisamment proche de 0 (et cela peut être long) pour pouvoir observer ce comportement.

2 Méthode de gradient à pas fixe, cas test en dimension 2.

(a) On a $\nabla f_a(x_0, x_1) = \frac{2}{(1+ax_0^2+x_1^2)^2}\begin{pmatrix} ax_0 \\ x_1 \end{pmatrix}$. Les itérées sont données par la formule de récurrence $X_{k+1} = X_k - \alpha \nabla f_a(X_k)$.

(b) Le code demandé pourrait ressembler à cela :

```
a=3

# On définit à la main la fonction gradient de f_a.
def gradfa(X):
    global compteur
    compteur=compteur+1
    d=(1+a*X[0]**2+X[1]**2)**2
    return 2/d*array([a,1])*X

# On écrit la descente de gradient.
# Noter que l'on ne fait qu'une seule évaluation de gradfa.
def methodeDescente(alpha,X0,eps):
    X=X0
    gX=gradfa(X0)
    while norm(gX)>eps:
        X=X-alpha*gX
        gX=gradfa(X)
    return X

X0=array([1,2]) # point de départ de la descente.
alpha=0.12
epsilon=logspace(-12,-1,20)
listen=[] # va contenir le nombre d'évaluations du gradient
          # pour chaque précision epsilon.
for eps in epsilon:
    compteur=0
    methodeDescente(alpha,X0,eps)
    listen.append(compteur)

semilogy(listen,epsilon,'-o')
xlabel("compteur")
ylabel("tolérance")
show()
```

(c) En 0, la hessienne de f_a est $\begin{pmatrix} 2a & 0 \\ 0 & 2 \end{pmatrix}$. Sa plus grande valeur propre est donc $\max(2, 2a)$. On s'attend donc à ce que la méthode converge, pour X_0 suffisamment proche de 0, lorsque $\alpha < \frac{2}{\max(2,2a)}$.

On s'attend à un taux de convergence inférieur à $\max(|1 - \alpha\ell|, |1 - \alpha L|)$,

où ℓ (resp. L) est la plus petite (resp. la plus grande) valeur propre de la hessienne au point de minimum. Donc ici, on s'attend à un taux inférieur à $\max(|1-2\alpha|, |1-2a\alpha|)$. Ce taux est minimisé pour $\alpha = \frac{2}{\ell+L}$ soit ici $\alpha = \frac{1}{1+a}$.

(d) La méthode par différences finies à droite permet de faire moins d'évaluations de fonctions : $n+1$ au lieu de $2n$. La méthode des différences finies centrées par contre est plus précise : on peut obtenir une erreur de l'ordre de $\eta^{\frac{2}{3}}$ au lieu de $\sqrt{\eta}$ pour les différences finies à droite (cf. TD1), si η est la précision relative du calcul de f, et si on se donne des hypothèses de régularité raisonnables sur f.

Dans le cas standard d'une précision η de l'ordre de 10^{-16}, un bon choix de ε est donc de 10^{-8} (soit $\sqrt{\eta}$) pour les différences finies à droite et 10^{-5} (soit $\eta^{\frac{1}{3}}$) pour les différences finies centrées.

(e) Si r est le taux de convergence linéaire de la méthode, on s'attend donc à un taux effectif de $r^{\frac{1}{n+1}}$ pour les différences finies à droite et $r^{\frac{1}{2n}}$ pour les différences finies centrées.

Étant donné que la précision sur le gradient est d'ordre $\sqrt{\eta}$ (ou $\eta^{\frac{2}{3}}$), cela n'a pas de sens de prendre pour critère d'arrêt que la norme du gradient est plus petite qu'une tolérance donnée si cette tolérance est elle-même plus petite que la précision de calcul du gradient. Pour que le calcul ait du sens, on doit donc prendre des tolérances plus grandes que $\sqrt{\eta}$ (ou $\eta^{\frac{2}{3}}$).

3 Problème d'application : recherche de trajectoires fermées sur un billard.

(a) Dessin

(b) Si $\theta_i \in [0,\pi]$ est l'angle entre la tangente dirigée par T_i et la trajectoire sortante en M_i, on a $\cos\theta_i = T_i \cdot U_i$. En effet comme T_i et U_i sont unitaires, on a $T_i \cdot U_i = \|T_i\|\|U_i\|\cos\theta_i$.

De même si $\varphi_i \in [0,\pi]$ est l'angle entre la trajectoire arrivant en M_i et la tangente, on a $\cos\varphi_i = (-T_i)\cdot(-U_{i-1})$. On a bien $-U_{i-1}$ qui correspond au vecteur unitaire dirigé de M_i à M_{i-1}.

On a donc que $\theta_i = \varphi_i$ si et seulement si leurs cosinus sont égaux, donc si $T_i \cdot U_i = T_i \cdot U_{i-1}$.

(c) On utilise la compacité de $[0,2\pi]^n$ pour montrer que L (continue) admet un maximum sur $[0,2\pi]^n$. Par périodicité ce point de maximum est inférieur à toute valeur de L sur $\mathbb{R}^n$, donc c'est un maximum global. Si n est pair, on peut montrer que le maximum est atteint en faisant $\frac{n}{2}$ allers-retours entre deux points du bord maximisant leur distance mutuelle.

(d) Par l'absurde, si on suppose que $M_i = M_i + 1$, on écrit

$$\begin{aligned} &L(t_0,\dots,t_{i-1},t_i+\varepsilon,t_{i+1},\dots,t_{n-1}) - L(t_0,\dots t_{n-1}\\ &\quad=\|\gamma(t_{i+1})-\gamma(t_i+\varepsilon)\| + \|\gamma(t_i+\varepsilon)-\gamma(t_{i-1})\| - \|\gamma(t_{i+1})-\gamma(t_{i-1})\|. \end{aligned}$$

Cette dernière quantité est positive (inégalité triangulaire), et elle n'est nulle que si $\gamma(t_i+\varepsilon)$ est situé sur le segment $[\gamma(t_{i-1}),\gamma(t_{i+1})]$. Or pour ε suffisamment petit, $\gamma(t_i+\varepsilon)$ est sur le bord du convexe, et différent de $\gamma(t_{i+1})$ (lui-même égal à $\gamma(t_i)$)) puisque $\gamma'(t_i) \neq 0$. De même, il est différent de $\gamma(t_{i-1})$. Comme le domaine est strictement convexe, alors les points du segment $[\gamma(t_{i-1}),\gamma(t_{i+1})]$ autres que les extrémités ne sont pas sur le bord. Faire un dessin pour bien voir tout ça !

Donc la quantité est strictement positive, en contradiction avec le fait qu'on a un maximum local en $(t_0,\dots t_{n-1})$.

On a donc toutes les normes dans l'expression de L qui sont strictement positives au voisinage du maximum local, donc on peut dériver.

On connait le gradient de la norme au carré : si $Q(x) = \langle x,x\rangle$, alors $\nabla Q(x) = 2x$. Donc par composition avec la racine, on obtient le gradient de la norme en un point non nul : si $N(x) = \|x\| = \sqrt{Q(x)}$ alors $\nabla N(x) = \frac{x}{\|x\|}$. Par composition, on obtient donc

$$\partial_{t_i} L = \gamma'(t_i)\cdot\frac{\gamma(t_i)-\gamma(t_{i-1})}{\|\gamma(t_i)-\gamma(t_{i-1})\|} - \gamma'(t_i)\cdot\frac{\gamma(t_{i+1})-\gamma(t_i)}{\|\gamma(t_{i+1})-\gamma(t_i)\|} = \gamma'(t_i)\cdot(U_{i-1}-U_i).$$

Ceci est vrai pour $i \neq 0$ et $i \neq n-1$, mais pour ces cas particuliers on obtient la même chose avec les notations $U_n = U_0$.

En un point de maximum local, on obtient donc bien que $\gamma'(t_i)\cdot(U_i - U_{i-1}) = 0$ pour tout i, donc on a bien affaire à une trajectoire de billard.