TP 1 : Algorithme de gradient

Soit $f : U \to \mathbb{R}$ de classe C^1, où U est un ouvert de $\mathbb{R}^n$. On veut calculer un minimiseur de f dans U i.e. une solution $\bar{x}$ du problème

$$f(\bar{x}) = \inf_{x \in U} f(x). \tag{1}$$

Un moyen de calculer $\bar{x}$ est donc de résoudre $\nabla f(\bar{x}) = 0$. Pour cela on peut utiliser par exemple l'algorithme de Newton ou un algorithme de point fixe. Mais ces algorithmes ne garantissent pas que l'on converge vers un minimiseur. Pour exploiter que $\bar{x}$ est un minimiseur de f, une idée est de construire des algorithmes de descente, dans lesquels f décroît à chaque étape.

Rappelons que $\nabla f(x)$ indique la direction et le sens dans lequel f croît le plus vite (au premier ordre) au voisinage de x. En effet, d'après le dvt de Taylor, pour $x \in \mathbb{R}^n$, $v \in \mathbb{R}^n$ avec $\|v\| = 1$, et $t \in \mathbb{R}$,

$$f(x + tv) = f(x) + t\langle \nabla f(x), v \rangle + t\varepsilon(t),$$

où $\varepsilon(t) \to 0$ dans $\mathbb{R}^n$ quand $t \to 0$. D'autre part, l'inégalité de Cauchy-Schwarz dit que

$$|\langle \nabla f(x), v \rangle \leq \|\nabla f(x)\| \quad \forall v \in \mathbb{R}^n,\ \|v\| = 1,$$

et qu'il y a égalité ssi $v = \pm \dfrac{\nabla f(x)}{\|\nabla f(x)\|}$ (pour $\nabla f(x) \neq 0$). Donc la meilleure direction v pour faire décroître f le plus vite sur la droite $x + tv$ est (au premier ordre en t), le choix $v = -\dfrac{\nabla f(x)}{\|\nabla f(x)\|}$. C'est le principe de l'algorithme du **gradient à pas fixe**, qui s'écrit

$$x^{k+1} = x^k - \mu \nabla f(x^k), \tag{2}$$

où $\mu > 0$ est un paramètre à fixer appelé *pas*. Comme il ne dépend pas de k, il s'agit d'un pas fixe.

On peut également voir cet algorithme comme le schéma d'Euler appliqué à l'EDO définissant la *plus grande pente* :

$$x'(t) = -\nabla f(x(t)). \tag{3}$$

On pourrait également considèrer l'algorithme de **gradient à pas optimal** : x_0 étant donné, pour $k = 0, 1, \ldots$ faire :

- trouver $\mu^k \geq 0$ qui réalise le minimum de la fonction $\mu \mapsto f(x^k - \mu \nabla f(x^k))$;
- poser $x^{k+1} = x^k - \mu^k \nabla f(x^k)$.

Le théorème suivant donne une condition suffisante sur la fonction f et sur le pas μ pour assurer la convergence de l'algorithme (2) vers un minimiseur de f.

Théorème 1 (Convergence de (2)) *On suppose que $f \in \mathcal{C}^1(\mathbb{R}^n, \mathbb{R})$ est α-convexe et que $\nabla f(x)$ est L-lipschitzien sur $\mathbb{R}^n$, i.e.*

$$\forall x, y \in \mathbb{R}^n, \quad \|\nabla f(x) - \nabla f(y)\| \le L\|x - y\|.$$

Alors pour tout $0 < \mu < 2\alpha/L^2$, l'algorithme de gradient à pas fixe converge : quel que soit x^0, la suite définie par (2) converge vers l'unique minimiseur $\bar{x}$ de f dans $\mathbb{R}^n$. De plus, la convergence est géométrique.

Exemple fondamental : Si $A \in M_n(\mathbb{R})$ est une matrice symétrique définie positive et $b \in \mathbb{R}^n$, alors la fonction

$$f(x) = \frac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle$$

définie $\forall x \in \mathbb{R}^n$ est α-convexe et son unique minimiseur dans $\mathbb{R}^n$ est également l'unique solution du système linéaire $Ax = b$.

Dans le cas d'un système linéaire, on a en fait le résultat plus précis suivant (cf. cours sur les systèmes linéaires) :

Théorème 2 *Soit $A \in M_n(\mathbb{R})$ symétrique définie positive, de valeurs propres $0 < \lambda_1 \le \lambda_2 \le \ldots \le \lambda_n$. La méthode du gradient à pas fixe converge si et seulement si $0 < \mu < 2/\lambda_n$, et la vitesse de convergence est optimale lorsque $\mu = 2/[\lambda_1 + \lambda_n]$.*

Exercices mathématiques

---

1. Pour $H = \mathbb{R}^2$ et $f(x, y) = ax^2 + by^2$ avec $a, b > 0$, étudier la convergence de l'algorithme à pas fixe : pour quelles valeurs de μ se produit-elle, pour quelle valeur est-elle la plus rapide ?
2. L'algorithme du gradient à pas fixe converge-t-il pour des matrices symétriques qui ne sont pas définies positives ? Pour des matrices non symétriques ?

Exercices informatiques

---

1. Programmer l'algorithme de gradient à pas fixe pour $f(x) = x^2 + \sin(x)$ sur $\mathbb{R}$. Cette fonction vérifie-t-elle les hypothèses du théorème 1 ? Quel pas peut-on choisir ? Vérifiez cela en pratique.

Essayez d'autres exemples : $ch(x)$ sur $\mathbb{R}$ (1-convexe), $x^4/4 - x^3/3 - x^2 + 1$ ou $\sin(x)$ (non convexes), $1/(1 - x^2)$ sur $]-1, 1[$ (domaine non défini partout)

2. Programmer l'algorithme de gradient à pas fixe pour l'exemple fondamental. Représenter l'erreur en fonction du nombre d'itérations pour différents pas.
3. Même question avec l'algorithme à pas optimal μ^k_{opt}. On montrera que

$$\mu^k_{opt} = \langle d^k, d^k\rangle / \langle Ad^k, d^k\rangle \text{ avec } d^k = Ax^k - b.$$

4. On choisit maintenant un pas $\mu^k = \theta^k \mu^k_{opt}$ où θ^k est choisi aléatoirement dans $[0, 2]$. Comparer la vitesse de convergence de cet algorithme par rapport aux précédents.