Optimisation et descente de gradient à pas fixe.

Soit $f : A \subset \mathbb{R}^d \to \mathbb{R}$, on considère le problème de minimisation suivant : trouver $x_* \in A$ tel que

$$f(x_*) = \min_{x \in A} f(x) \,. \tag{P}$$

Ce problème sous-tend plusieurs questions de natures assez différentes. Citons principalement :

(i) **Existence** de solution au problème (P) ?

(ii) **Caractérisation** des solutions : on parle souvent de conditions d'optimalité, leur forme varie selon la régularité de f ou la présence de contraintes par exemple.

(iii) **Calcul numérique** des solutions, garanties de convergence ?

1 Existence et unicité

On commence par rappeler les notions de minimum local/global.

Définition 1 (Minimum global, local)**.** On dit que f admet un *minimum global* sur A en $x_* \in A$ si

$$\forall x \in A,\ f(x_*) \leq f(x) \,.$$

On dit que f admet un *minimum local* sur A en $x_* \in A$ si

$$\exists \delta > 0,\ \forall x \in A, \|x - x_*\| \leq \delta \Rightarrow f(x_*) \leq f(x) \,.$$

On dira de façon équivalente que x_* est un *minimiseur local/global* de f sur A ou encore que x_* *réalise un minimum local/global* de f sur A.

Dans le cadre de la dimension finie, auquel on se restreint ici, les ensembles compacts sont les fermés bornés. Comme une fonction continue sur un compact atteint ses bornes, la question de l'existence admet diverses réponses assez simples à énoncer. Par exemple :

Proposition 1 (Existence en dimension finie)**.** *On suppose que $A \subset \mathbb{R}^d$ est fermé (non vide) et $f : A \to \mathbb{R}$ est continue*[1] *et coercive, i.e.*

$$\lim_{\substack{\|x\| \to +\infty \\ x \in A}} f(x) = +\infty \,,$$

alors f admet au moins un point de minimum global $x_* \in A$.

Preuve. En exercice. □

En ce qui concerne l'unicité, c'est du cas par cas, en dehors du cadre favorable de la stricte convexité :

Proposition 2 (Unicité et stricte convexité)**.** *On suppose que $A \subset \mathbb{R}^d$ est convexe et $f : A \to \mathbb{R}$ est strictement convexe, alors f admet au plus un point de minimum global $x_* \in A$.*

Preuve. En exercice. □

[1] semi-continue inférieurement est suffisant.

2 Conditions d'optimalité

On supposera dans toute la section que l'ensemble A est **ouvert** afin de pouvoir faire des petites variations dans toutes les directions autour de chaque point. On écarte ce faisant le cas des problèmes de minimisation sous contraintes égalités/inégalités (non strictes) par exemple.

2.1 Conditions nécessaires

Proposition 3. *Soit $f : A \subset \mathbb{R}^d \to \mathbb{R}$ différentiable au point $x_* \in A$. Si x_* réalise un minimum local de f alors,*

- Condition d'ordre 1 : équation d'Euler $\nabla f(x_*) = 0$.
- Conditions d'ordre 2*: Si f est deux fois différentiable en x_*, alors la Hessienne $\nabla^2 f(x_*)$ est une matrice positive, i.e. $\forall h \in \mathbb{R}^d$, $\left\langle \nabla^2 f(x_*)h, h \right\rangle \geq 0$.*

Preuve. Il suffit de se ramener aux résultats 1–dimensionnels bien connus, qu'on applique dans toutes les directions autour de x_*. Soit $h \in \mathbb{R}^d$, on considère

$$\begin{array}{rccl} \phi & : & [-\epsilon, \epsilon] & \to & \mathbb{R} \\ & & t & \mapsto & f(x_* + th)\,, \end{array}$$

où on a choisi $\epsilon > 0$ assez petit pour que ϕ soit bien défnie (possible car A ouvert). La fonction ϕ est dérivable en 0 et

$$\phi'(0) = Df(x_*)(h) =< \nabla f(x_*), h > \ .$$

De plus, 0 réalise un minimum local de ϕ, de sorte que $\phi'(0) = 0 =< \nabla f(x_*), h >$. Comme c'est vrai pour tout $h \in \mathbb{R}^d$, $\nabla f(x_*) = 0$. Si maintenant f est deux fois différentiable en x_*, alors ϕ est deux fois dérivable en 0 et

$$\phi''(0) = D^2 f(x_*)(h, h) = \left\langle \nabla^2 f(x_*)h, h \right\rangle \ .$$

Comme x_* réalise un minimum local de f, 0 réalise un minimum local de ϕ et par conséquent $\phi''(0) \geq 0$. □

2.2 Conditions suffisantes

On cherche à présent des conditions assurant qu'un candidat x_* réalise un minimum local de f. On suppose déjà que x_* satisfait la condition nécessaire d'ordre 1 $\nabla f(x_*) = 0$: on dit que x_* est un *point critique* de f. On se convaincra facilement que c'est loin d'être suffisant pour conclure que x_* réalise un minimum local de f (par exemple, 0 est un point critique de $t \mapsto t^3$). La condition nécessaire d'ordre 2 (positivité de la Hessienne) n'est toujours pas suffisante non plus (on pourra méditer $x_* = 0$ et $f : x \mapsto x^5$).
On peut donner les conditions suffisantes d'optimalité suivantes.

Proposition 4. *On suppose que $f : A \subset \mathbb{R}^d \to \mathbb{R}$ est différentiable en $x_* \in A$ et que x_* est un point critique de f. Si f est deux fois différentiable en x_* et si $\nabla^2 f(x_*)$ est une matrice définie postive, alors x_* réalise un minimum local (strict) de f.*

Preuve. On utilise correctement les formules de Taylor. □

Malheureusement, de même que les conditions nécessaires énoncées précédemment n'étaient pas suffisantes en général, ces conditions suffisantes d'optimalité ne sont pas nécessaires. Cependant, il existe un cas particulier, mais important, dans lequel être point critique est une condition nécessaire et suffisante de minimiseur local : lorsque la fonction f est convexe.

2.3 Le cas des fonctions convexes

On commence par rappeler une caractérisation d'ordre 2 des fonctions convexes, l'ensemble A est toujours ouvert et il est de plus supposé **convexe**.

Proposition 5 (Caractérisation des fonctions convexes)**.** *Soit $A \subset \mathbb{R}^d$ toujours ouvert, mais aussi convexe et $f : A \to \mathbb{R}$ différentiable.*

(*i*) *f est convexe ssi pour tout x, $y \in A$, $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$.*
Si de plus f est deux fois différentiable, alors

(*ii*) *f est convexe ssi pour tout $x \in A$, la matrice Hessienne $\nabla^2 f(x)$ est positive.*

(*iii*) ***Si** pour tout $x \in A$, la matrice Hessienne $\nabla^2 f(x)$ est définie postive, alors f est strictement convexe.*

Preuve. On admet (déjà vu dans un autre cours ?), attention, (*iii*) n'est pas une équivalence. □

Proposition 6. *Soit $f : A \to \mathbb{R}$ une fonction convexe. Alors,*

(*i*) *si f est différentiable, alors $x_* \in A$ réalise un minimum local de f si et seulement si x_* est un point critique : $\nabla f(x_*) = 0$,*

(*ii*) *si f admet un minimum local alors c'est automatiquement un minimum global.*

Preuve. On utilise la caractérisation précédente de convexité (d'ordre 1). Pour le (*i*), il reste à montrer que si $\nabla f(x_*) = 0$, alors x_* est bien un minimiseur local (dans l'autre sens c'est la condition nécessaire d'optimalité d'ordre 1 énoncée dans la proposition 3). Grâce à la proposition 5(*i*) appliquée avec $x = x_*$, on obtient que pour tout $y \in A$,

$$f(y) \geq f(x_*) + \langle \nabla f(x_*), y - x \rangle = f(x_*)\,,$$

on en déduit en même temps (*ii*) que tout minimum local de f est global. □

3 Un exemple d'algorithme numérique : la descente de gradient à pas fixe

On suppose dans cette section que $A = \mathbb{R}^d$, f est au moins de classe C^1 et $x_* \in \mathbb{R}^d$ est une solution de $f(x_*) = \min f$. Notre but est d'approcher numériquement x_*. On va se concentrer sur l'algorithme de *descente de gradient* qui appartient à la famille plus large des méthodes de descente.

3.1 Méthodes de descente

Principe général.

- On part de $x_0 \in \mathbb{R}^d$ ($k = 0$),
- While (*critère d'arrêt*)
  - choisir $d_k \in \mathbb{R}^d$ une *direction de descente*
  - choisir un *pas* τ_k
  - définir $x_{k+1} = x_k + \tau_k d_k$
  - $k \longleftarrow k + 1$.

Précisons ce qu'on entend par *direction de descente*.

Définition 2 (Direction de descente)**.** On dit que $h \in \mathbb{R}^d$ est une *direction de descente* (stricte) au point x_0 s'il existe $\tau_0 > 0$ tel que pour tout $\tau \in]0, \tau_0]$,

$$f(x_0 + \tau h) < f(x_0)\,.$$

Une direction de descente est une direction dans laquelle le fonction f décroît localement au voisinage de x_0

Exemple. Soit $x, h \in \mathbb{R}^d$.

- Si $\langle \nabla f(x), h\rangle < 0$ alors h est une direction de descente au point x.

  En effet, soit $g : t \in \mathbb{R} \mapsto f(x + th)$. On a $g'(t) = \langle \nabla f(x + th), h\rangle$ de sorte que $g'(0) = \langle \nabla f(x), h\rangle < 0$. Par ailleurs,

  $$g'(0) = \lim_{\substack{t \to 0 \\ t>0}} \frac{f(x + th) - f(x)}{t}\,,$$

  et il existe ainsi $t_0 > 0$ tel que $\forall 0 < t < t_0$,

  $$\frac{f(x + th) - f(x)}{t} < 0 \quad \Rightarrow \quad f(x + th) < f(x)$$

  et h est bien une direction de descente.

- Si $h = -\nabla f(x) \neq 0$, alors h est une direction de descente au point x, c'est même celle "de plus grande pente".

  En effet, comme $\langle \nabla f(x), -\nabla f(x)\rangle = -\|\nabla f(x)\|^2 < 0$ ($\nabla f(x) \neq 0$), c'est bien une direction de descente (en utilisant le point précédent). C'est la direction "de plus grande pente" au sens où $-\dfrac{d}{dt} f(x + th)\Big|_{t=0}$ est le plus grand possible pour $h \in \mathbb{R}^d$ de norme fixée : par Cauchy-Schwartz,

  $$-\frac{d}{dt} f(x + th)\Big|_{t=0} = \langle -\nabla f(x), h\rangle \leq \|\nabla f(x)\|\,\|h\|$$

  avec égalité si et seulement si h et $-\nabla f(x)$ sont positivement liés. Géométriquement, le gradient pointe dans la direction orthogonale aux lignes de niveaux de la fonction.

Il existe différentes méthodes de descente en fonction du choix de la direction de descente d_k et du pas τ_k. On va à présent examiner le choix $d_k = -\nabla f(x_k)$ et $\tau_k = \tau$ constant, la méthode qui en résulte est appelée *descente de gradient à pas fixe.*

3.2 Descente de gradient à pas fixe

On formalise l'algorithme de descente de gradient avant d'étudier ses propriétés de convergence.

Algorithme 1 : Gradient à pas fixe

Data : gradient de f `gradfun`, initialisation `x0`, nombre maximal d'itérations `maxIt`, pas `tau`, tolérance `epsilon` pour le critère d'arrêt.

initialisation `x` ⟵ `x0`
`nbIt` ⟵ `0`
while *critère d'arrêt pas satisfait* **and** *`nbIt < maxIt`* **do**
  mettre à jour `x` ⟵ `x - tau*gradfun(x)`
  incrémenter `nbIt` ⟵ `nbIt + 1`
return *`x` and ...*

On énonce à présent un cas de convergence qui sera largement suffisant pour traiter les problèmes considérés dans le TP.

Proposition 7 (Convergence du gradient à pas fixe)**.** *Soit $f : \mathbb{R}^d \to \mathbb{R}$ de classe C^2. On suppose que la matrice Hessienne $\nabla^2 f(x)$ est positive en tout $x \in \mathbb{R}^d$ et qu'il existe $0 < l \leq L$ tels que pour tout $x \in \mathbb{R}^d$, les valeurs propres de $\nabla^2 f(x)$ sont comprises entre l et L. Alors,*

1. *la fonction f est strictement convexe, coercive et admet un unique minimum global $x_* \in \mathbb{R}^d$ caractérisé par $\nabla f(x_*) = 0$,*
2. *pour tout $x_0 \in \mathbb{R}^d$ et pour tout pas $\tau \in \left]0, \frac{2}{L}\right[$, la suite $(x_k)_{k\in\mathbb{N}}$ donnée par la descente de gradient (à pas fixe τ partant de x_0) converge vers x_* et de plus $\exists C > 0$, $\forall k$,*

$$\|x_k - x_*\| \leq C r^k \quad \textit{avec} \quad r \leq \max(|1 - l\tau|, |1 - L\tau|) < 1 \,.$$

Preuve. On laisse la preuve du premier point en exercice. On définit $F : \mathbb{R}^d \to \mathbb{R}^d$, $x \mapsto x - \tau \nabla f(x)$ de sorte que la suite $(x_k)_k$ vérifie pour tout $k \in \mathbb{N}$, $x_{k+1} = F(x_k)$. On va montrer que F est contractante et appliquer le théorème du point fixe de Banach. La fonction F est de classe C^1 et pour tout $x, h \in \mathbb{R}^d$,

$$DF(x)h = h - \tau \nabla^2 f(x)h = (I_d - \tau \nabla^2 f(x))h \,.$$

Soit $x \in \mathbb{R}^d$ fixé et $M = I_d - \tau \nabla^2 f(x) \in \mathrm{M}_d(\mathbb{R})$ symétrique réelle, par le théorème spectral, soit $(v_1, \ldots, v_d)$ une base orthonormale de vecteurs propres : pour tout $k = 1 \ldots d$, $Mv_k = \lambda_k v_k$. On a alors pour $h \in \mathbb{R}^d$, qu'on écrit dans la base précédente :

$$h = \sum_{k=1}^{d} h_k v_k \quad \text{et ainsi,} \quad \|h\|^2 = \sum_{k=1}^{d} h_k^2 \quad \text{et} \quad \|Mh\|^2 = \sum_{k=1}^{d} \lambda_k^2 h_k^2 \leq \max_{k=1\ldots d} |\lambda_k|^2 \|h\|^2 \,.$$

On en déduit que la norme d'opérateur de M (norme matricielle) associée à la norme euclidienne dans $\mathbb{R}^d$,

$$\|M\|_{op} = \sup_{h \in \mathbb{R}^d \setminus \{0\}} \frac{\|Mh\|}{\|h\|} \leq \max_k |\lambda_k| \,.$$

Comme les valeurs propres de M sont $\{1 - \tau\lambda \,:\, \lambda \text{ valeur propre de } \nabla^2 f(x)\}$ et $0 < l \leq \lambda \leq L$, on conclut

$$\|DF(x)\|_{op} = \|I_d - \tau \nabla^2 f(x)\|_{op} \leq r(\tau) \quad \text{avec} \quad r(\tau) := \max(|1 - \tau l|, |1 - \tau L|) \,.$$

Enfin, par l'inégalité des accroissements finis, pour tout $y, z \in \mathbb{R}^d$,

$$\|F(z) - F(y)\| \leq \sup_{x \in [y,z]} \|DF(x)\|_{op} \|z - y\| \leq r(\tau) \|z - y\| \,.$$

Comme $\tau < \dfrac{2}{L}$, on vérifie que $r(\tau) < 1$ et F est bien contractante. On peut appliquer le théorème du point fixe de Banach et pour tout $k \in \mathbb{N}^*$,

$$\|x_k - x_*\| \leq \frac{r(\tau)^k}{1 - r(\tau)} \|x_1 - x_0\|$$

où le minimiseur global x_* de f est l'unique point fixe de F sur $\mathbb{R}^d$.

□