Cours optimisation

Benjamin Monmege

29 février 2012

On appelle problème d'optimisation la donnée d'une instance de la forme

$$\begin{aligned} &\text{minimiser/maximiser } f(x) \\ &\text{sous les conditions } \begin{cases} g_i(x) \leq 0 & \forall i \in \{1, \dots, m\} \\ h_j(x) = 0 & \forall j \in \{1, \dots, m'\} \end{cases} \end{aligned}$$

avec $f \colon \mathbb{R}^p \to \mathbb{R}$ une fonction objectif et $g_i, h_j \colon \mathbb{R}^p \to \mathbb{R}$ des fonctions de contraintes.

L'ensemble des points

$$\{x \in \mathbb{R}^p \mid g_i(x) \leq 0 \ \forall i \in \{1, \dots, m\}, h_j(x) = 0 \ \forall j \in \{1, \dots, m'\}\}$$

est appelé ensemble de candidats. S'il est non vide, on appelle solution du problème d'optimisation tout minimum/maximum global de la fonction f restreinte à cet ensemble de candidats, noté x^*.

Dans ce cours, on étudie uniquement les problèmes de minimisation sans contraintes, c'est-à-dire avec $m = m' = 0$. Ainsi, l'ensemble des candidats correspond à l'espace $\mathbb{R}^p$ tout entier. Remarquons qu'un problème de maximisation $\max f(x)$ peut naturellement se transformer en un problème de minimisation $\min -f(x)$.

Pourquoi s'intéresser à ces problèmes d'optimisation ? Mettons-nous un instant dans la situation où l'on souhaite *apprendre* des données :

$$\begin{aligned} \text{Données :} &\quad (x_i, y_i) \in \mathbb{R}^n \times \mathbb{R}, i \in \{1, \dots, N\} \\ \text{But :} &\quad \text{trouver un « modèle » } y = f(x) \text{ susceptible d'avoir généré les données} \end{aligned} \tag{1}$$

dans le but de pouvoir ensuite *prédire* au mieux le comportement sur une nouvelle donnée brute $x \in \mathbb{R}^n$. Un expert, sachant d'où sont extraites les données, choisit alors une catégorie de modèles potentiels, à savoir une famille $(f_\theta)_{\theta \in \mathbb{R}^p}$ de fonctions $\mathbb{R}^n \to \mathbb{R}$ régie par des paramètres $\theta \in \mathbb{R}^p$. Notre objectif est alors de trouver le meilleur paramètre θ, par exemple en minimisant la fonction d'erreur des sommes de carrés, c'est-à-dire de résoudre le problème d'optimisation :

$$\text{minimiser} \quad \frac{1}{2} \sum_{i=1}^{N} |f_\theta(x_i) - y_i|^2 \text{ pour } \theta \in \mathbb{R}^p.$$

Avant de détailler les outils qui vont nous permettre de résoudre ces problèmes d'optimisation, notons qu'il est possible de restreindre nos attentes en cherchant un optimum local (et non global comme précédemment), c'est-à-dire chercher une valeur x^* qui optimise la fonction f sur un voisinage de x^* et non sur l'espace tout entier. Dans ce cas, on peut utiliser des méthodes de descente (si la fonction objectif est différentiable), par exemple basée sur le gradient, que nous développerons en Section 2, ou des méthodes telles que la méthode du simplexe (si la fonction objectif est linéaire) ou de recherches de motifs (si la fonction objectif n'est pas différentiable). À noter que dans le cas d'une fonction objectif prenant des valeurs discrètes, les méthodes sont tout à fait différentes et constituent le domaine de l'optimisation combinatoire. Dans le cas des optima globaux, il est souvent nécessaire d'introduire du non-déterminisme pour s'échapper des optima locaux, à moins bien sûr de vouloir payer le prix d'une recherche *exhaustive*. À noter que le cas convexe que nous allons développer permet d'assurer que le minimum local trouvé est en fait un minimum global : on trouvera de plus amples détails sur l'optimisation convexe dans [1].

1 Outils mathématiques et conditions d'optimalité

Afin de développer les méthodes d'optimisation, il est nécessaire de connaître un minimum d'outils mathématiques relatifs au calcul différentiel. Pour de plus amples informations ou exercices, on se reportera par exemple à [2].

1.1 Différentielle et gradient

On se place dans des espaces vectoriels normés de dimension finie : on notera toujours $\|\cdot\|$ la norme, indépendamment de l'espace utilisé. Dans l'espace $\mathbb{R}^n$, on note $(e_1, \dots, e_n)$ la base canonique et on identifie le plus souvent une application linéaire et sa matrice dans la base canonique. Par ailleurs, on notera A^{T} la transposée d'une matrice A. L'espace $\mathbb{R}^n$ est naturellement muni d'un produit scalaire canonique, noté $\langle\cdot,\cdot\rangle$, et défini pour deux vecteurs $x, y \in \mathbb{R}^n$ par $\langle x, y\rangle = x^{\mathrm{T}}y$.

Définition 1 (différentielle)**.** Soient U un ouvert de $\mathbb{R}^n$ et $a \in U$. Une fonction $f\colon U \to \mathbb{R}^p$ est dite différentiable en a s'il existe une application linéaire continue $L\colon \mathbb{R}^n \to \mathbb{R}^p$ et une fonction $\varepsilon : \mathbb{R}^n \to \mathbb{R}^p$ telles que

- pour tout $h \in \mathbb{R}^n$ tel que $a + h \in U$, $f(a+h) - f(a) = L(h) + \|h\|\varepsilon(h)$;
- $\lim\limits_{h\to 0} \|\varepsilon(h)\| = 0$.

Si L existe, elle est unique : on la note $\mathrm{D}f(a)$ (ou $f'(a)$, $\mathrm{d}f(a)$, $\mathrm{D}_a f$...) et on l'appelle différentielle de f en a : dans la suite, on notera

$$f(a+h) = f(a) + \mathrm{D}f(a)h + o(\|h\|)\,.$$

La fonction f est dite différentiable, si elle est différentiable en tout point de U.

Exemple 1. Si f est une application linéaire sur $U = \mathbb{R}^n$ alors f est différentiable de différentielle f. Si $f\colon \mathbb{R}^n \to \mathbb{R}$ est définie pour tout $x \in \mathbb{R}^n$ par $f(x) = x^{\mathrm{T}}Ax$ avec A une matrice $n \times n$ à coefficients réels, alors f est différentiable de différentielle $\mathrm{D}f(a)h = a^{\mathrm{T}}(A + A^{\mathrm{T}})h$.

Théorème 1. *Soient U un ouvert de $\mathbb{R}^n$ et $a \in U$.*

- *Soient $f, g\colon U \to \mathbb{R}^p$ deux fonctions différentiables en a et $\lambda \in \mathbb{R}$. L'application $f + \lambda g$ est différentiable en a de différentielle $\mathrm{D}(f+\lambda g)(a) = \mathrm{D}f(a) + \lambda \mathrm{D}g(a)$.*
- *Soient $f\colon U \to \mathbb{R}^p$ une application différentiable en a et $g\colon f(U) \to \mathbb{R}^q$ une application différentiable en $f(a)$. L'application $g \circ f$ est différentiable en a de différentielle $\mathrm{D}(g\circ f)(a) = \mathrm{D}g(f(a)) \circ \mathrm{D}f(a)$.*

Exercice 1. Soient U un ouvert de $\mathbb{R}^n$, $a \in U$, $v \in \mathbb{R}^n$ et $f\colon U \to \mathbb{R}^p$ une application différentiable en a. Montrer que l'application $t \in \mathbb{R} \mapsto f(a+tv)$ est définie sur un voisinage de $t = 0$ et est différentiable en 0. Que vaut sa différentielle en 0 ?

Si $f\colon U \to \mathbb{R}^p$ est différentiable en a, on note $\partial f/\partial x_i(a)$ (ou $\partial_i f(a)$) le vecteur $\mathrm{D}f(a)e_i \in \mathbb{R}^p$. Grâce à la linéarité de la différentielle, on a pour tout $h = (h_1, \dots, h_n) \in \mathbb{R}^n$

$$\mathrm{D}f(a)h = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(a) h_i\,.$$

Définition 2 (gradient)**.** On se place ici dans le cas où $p = 1$. On considère à nouveau un ouvert U de $\mathbb{R}^n$. Si f est une application $U \to \mathbb{R}$ différentiable en un point $a \in U$, telle que $\mathrm{D}f(a)$ n'est pas l'application nulle, alors il existe un unique vecteur $\nabla f(a) \in \mathbb{R}^n$, appelé gradient de f en a, vérifiant pour tout $h \in \mathbb{R}^n$, $\mathrm{D}f(a)h = \langle \nabla f(a), h\rangle$.

On peut aisément vérifier que $\nabla f(a) = (\partial f/\partial x_1(a), \dots, \partial f/\partial x_n(a))$.

Exemple 2. Si f est une forme linéaire, de matrice $u^{\mathrm{T}} \in \mathbb{R}^{1\times n}$ (cela veut donc dire que $f(x) = u^{\mathrm{T}}x$ pour tout $x \in \mathbb{R}^n$), alors $\nabla f(a) = u$. Si $f(x) = x^{\mathrm{T}}Ax$, alors $\nabla f(a) = (A + A^{\mathrm{T}})a$.

Exercice 2. Soit $f\colon \mathbb{R}^n \to \mathbb{R}$ une application différentiable. Pour tout réel c, on appelle ligne de niveau c de f l'ensemble $\mathcal{N}_c = \{x \in \mathbb{R}^n \mid f(x) = c\}$. Soit $x_0 \in \mathcal{N}_c$. Montrer que $\nabla f(x_0)$ est orthogonal à $\mathcal{N}_c$ en x_0.

1.2 Différentielle seconde et Hessienne

On se limite dans cette sous-section aux fonctions numériques, pour simplifier l'exposé : il est bien sûr possible de généraliser les définitions et théorèmes suivants au cas général.

Définition 3 (différentielle seconde et Hessienne)**.** Soient U un ouvert de $\mathbb{R}^n$ et $a \in U$. Une fonction $f\colon U \to \mathbb{R}$ est dite deux fois différentiable en a si les conditions suivantes sont vérifiées :
- f est différentiable sur un voisinage V de a dans U ;
- l'application $\mathrm{D}f\colon V \to (\mathbb{R}^n)^*, u \mapsto Df(u)$ (on note $(\mathbb{R}^n)^*$ l'espace vectoriel des formes linéaires sur $\mathbb{R}^n$, qui s'identifie par dualité à $\mathbb{R}^n$ lui-même) est différentiable en a.

On note alors $\mathrm{D}^2 f(a) = \mathrm{D}(\mathrm{D}f)(a)$ (application linéaire de $\mathbb{R}^n$ dans $(\mathbb{R}^n)^*$) la différentielle seconde ainsi obtenue. Cette application linéaire est représentée dans les bases canoniques (base canonique de $\mathbb{R}^n$ et base duale canonique de $(\mathbb{R}^n)^*$) par une matrice carrée appelée matrice Hessienne de f en a, souvent notée $\nabla^2 f(a)$.

On note $\frac{\partial^2 f}{\partial x_i \partial x_j}(a)$ le vecteur $\frac{\partial}{\partial x_i}\left(\frac{\partial f}{\partial x_j}\right)(a)$, et on abrège $\frac{\partial^2 f}{\partial x_i \partial x_i}(a)$ en $\frac{\partial^2 f}{\partial x_i^2}(a)$. On peut alors aisément vérifier que

$$\nabla^2 f(a) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2}(a) & \frac{\partial^2 f}{\partial x_2 \partial x_1}(a) & \dots & \frac{\partial^2 f}{\partial x_n \partial x_1}(a) \\ \vdots & \vdots & & \vdots \\ \frac{\partial^2 f}{\partial x_1 \partial x_n}(a) & \frac{\partial^2 f}{\partial x_2 \partial x_n}(a) & \dots & \frac{\partial^2 f}{\partial x_n^2}(a) \end{pmatrix}.$$

De même qu'on a associé à $\mathrm{D}f(a)$ la forme linéaire sur $\mathbb{R}^n$

$$h \in \mathbb{R}^n \mapsto \mathrm{D}f(a)h = \sum_{i=1}^n \frac{\partial f}{\partial x_i}(a) h_i = \langle h, \nabla f a \rangle,$$

on associe à $\mathrm{D}^2 f(a)$ une forme bilinéaire sur $\mathbb{R}^n \times \mathbb{R}^n$ définie par

$$(h,k) \in \mathbb{R}^n \times \mathbb{R}^n \mapsto \mathrm{D}^2 f(a)(h,k) = \sum_{i,j=1}^n \frac{\partial^2 f}{\partial x_i \partial x_j}(a) h_i k_j = \langle h, \nabla^2 f(a) k \rangle.$$

Exercice 3. Calculer la Hessienne de la fonction $f\colon \mathbb{R}^n \to \mathbb{R}, x \mapsto x^{\mathrm{T}} A x$.

Théorème 2 (Schwarz)**.** *Soit U un ouvert de $\mathbb{R}^n$ et $a \in U$. Si $f\colon U \to \mathbb{R}$ est deux fois différentiable en a, alors pour tous $i, j \in \{1, \dots, n\}$,*

$$\frac{\partial^2 f}{\partial x_i \partial x_j}(a) = \frac{\partial^2 f}{\partial x_j \partial x_i}(a).$$

Autrement dit, la matrice $\nabla^2 f(a)$ est symétrique, ou bien la forme bilinéaire $\mathrm{D}^2 f(a)$ est symétrique.

Théorème 3 (Formule de Taylor-Young à l'ordre 2)**.** *Soit U un ouvert de $\mathbb{R}^n$ et $a \in U$. Si $f\colon U \to \mathbb{R}$ est deux fois différentiable en a, on a*

$$f(a+h) - f(a) - \mathrm{D}f(a)h - \frac{1}{2}\mathrm{D}^2 f(a)(h,h) = o(\|h\|^2)$$

lorsque h tend vers 0 dans $\mathbb{R}^n$.

Rappelons au passage la formule de Taylor-Lagrange à l'ordre 2 pour les fonctions $g : \mathbb{R} \to \mathbb{R}$ deux fois dérivables en $a \in \mathbb{R}$ (dont la preuve repose sur le théorème de Rolle) : pour tout $h \in \mathbb{R}$, il existe $\theta \in [0,1]$ tel que

$$g(a+h) = g(a) + g'(a)h + \frac{1}{2} g''(a+\theta h) h^2.$$

Si $f : U \to \mathbb{R}$ est une fonction différentiable sur un ouvert U de $\mathbb{R}^n$, cette formule est en particulier applicable à la fonction $g(t) = f(a+tv)$ pour tout $a \in U$ et $v \in \mathbb{R}^n$.

1.3 Conditions d'optimalité

Soient X une partie de $\mathbb{R}^n$, $a \in X$ et $f\colon X \to \mathbb{R}$ une fonction numérique. On dit que f admet en a un minimum global si $f(x) \geq f(a)$ pour tout $x \in X$.

On dit que f admet en a un minimum local s'il existe un voisinage V de a dans $\mathbb{R}^n$ tel que $f(x) \geq f(a)$ pour tout $x \in V \cap X$.

On dit que f admet en a un minimum strict si les inégalités précédentes sont strictes pour $x \neq a$.

Les définitions correspondantes pour maximum s'en déduisent en inversant les inégalités. Le terme extremum signifie maximum ou minimum.

Théorème 4. *Soient U un ouvert de $\mathbb{R}^n$, $a \in U$ et $f\colon U \to \mathbb{R}$ une fonction numérique.*

1. *Condition nécessaire (pas suffisante) du premier ordre : si f admet en a un extremum local et si f est différentiable en a, alors $\mathrm{D}f(a) = 0$.*
2. *Condition nécessaire (pas suffisante) du second ordre : si f admet en a un minimum local et si f est deux fois différentiable en a, alors $\mathrm{D}f(a) = 0$ et $\mathrm{D}^2 f(a)(h,h) \geq 0$ pour tout vecteur $h \in \mathbb{R}^n$.*
3. *Condition suffisante (pas nécessaire) du second ordre : si f est deux fois différentiable en a, $\mathrm{D}f(a) = 0$ et $\mathrm{D}^2 f(a)(h,h) > 0$ pour tout $h \in \mathbb{R}^n - \{0\}$, alors f admet en a un minimum local strict.*

2 Optimisation de fonctions quadratiques

2.1 Solution analytique

On considère le problème (1) dans le cas où l'expert choisit la catégorie des modèles affines : $f(x) = w^{\mathrm{T}}x + \beta$ avec $w \in \mathbb{R}^n$ et $\beta \in \mathbb{R}$. Afin de simplifier les notations, on introduit le paramètre $\theta = \begin{pmatrix} \beta \\ w \end{pmatrix} \in \mathbb{R}^{n+1}$, le vecteur $y = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix} \in \mathbb{R}^N$ et la matrice $X = \begin{pmatrix} 1 & x_1^{\mathrm{T}} \\ \vdots & \vdots \\ 1 & x_N^{\mathrm{T}} \end{pmatrix}$ de taille $N \times (n+1)$. Ainsi, on est ramené au problème de minimisation suivant :

$$\text{minimiser} \quad \frac{1}{2}\|X\theta - y\|_2^2 \text{ pour } \theta \in \mathbb{R}^{n+1}$$

où on a noté $\|\cdot\|_2$ la norme 2 sur $\mathbb{R}^N$, issue du produit scalaire canonique.

En utilisant le Théorème 4, on peut montrer que toute solution optimale θ^* de ce problème vérifie l'équation

$$X^{\mathrm{T}}X\theta^* = X^{\mathrm{T}}y \tag{2}$$

appelée équations normales.

Exercice 4. Montrer (2).

2.2 Régularisation

Comme toujours dans un tel problème d'apprentissage, la tentation est grande de « surapprendre » les données. Pour éviter ce désagrément, il est possible d'introduire un terme de régularisation, empêchant le paramètre θ de prendre des valeurs trop importantes :

$$\text{minimiser} \quad \frac{1}{2}\|X\theta - y\|_2^2 + \frac{\lambda}{2}\|\theta\|_2^2 \text{ pour } \theta \in \mathbb{R}^{n+1}$$

avec un paramètre λ de régularisation. Dans ce cas particulier, on trouve également une solution optimale de manière analytique, qui doit vérifier l'équation $(\lambda I + X^{\mathrm{T}}X)\theta^* = X^{\mathrm{T}}y$, avec I la matrice identité.

3 Optimisation de fonctions différentiables

3.1 Méthode de descente

On considère désormais une fonction numérique $f\colon \mathbb{R}^n \to \mathbb{R}$ deux fois différentiable. Le principe général des méthodes de descente est de construire une séquence $(x_k)_{k\leq 0}$ de points de $\mathbb{R}^n$ qui va converger vers un minimum local x^*. Cette séquence sera construite en choisissant un point initial $x_0 \in \mathbb{R}^n$ puis, pour $k \geq 0$, à répéter tant que nécessaire les étapes suivantes :

1. choisir une direction de descente $d_k \in \mathbb{R}^n - \{0\}$;
2. choisir une taille de pas $\sigma_k > 0$;
3. poser $x_{k+1} = x_k + \sigma_k d_k$ et continuer avec $k = k + 1$.

Le problème est de bien choisir x_0, les directions d_k et les tailles de pas σ_k. À chaque étape, d_k et σ_k sont choisis tels que $f(x_{k+1}) < f(x_k)$ (sauf si on a atteint le minimum, c'est-à-dire $x_k = x^*$).

3.2 Cas des fonctions convexes

On fixe dans toute la sous-section un ouvert convexe U de $\mathbb{R}^n$ et une fonction $f\colon U \to \mathbb{R}$. On dit que f est convexe sur U si, pour tous $x, y \in U$ et pour tout $t \in [0, 1]$,

$$f((1-t)x + ty) \leq (1-t)f(x) + tf(y)$$

où $(1-t)x+ty$ appartient à U, grâce à l'hypothèse de convexité de U. Ceci revient à dire que le graphe de f est au-dessous des cordes.

Exemple 3. Quelques exemples de fonctions convexes :
- les fonctions affines : $f(x) = \langle a, x\rangle + b$;
- les fonctions quadratiques : $f(x) = \frac{1}{2}\langle x, Ax\rangle + \langle a, x\rangle + b$ avec A une matrice symétrique positive ;
- les normes ℓ_p : $\|x\|_p = \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}$ pour $1 \leq p < \infty$ et $\|x\|_\infty = \max_i x_i$;
- l'opposé de la fonction entropie : $f(x) = -\sum_{i=1}^n x_i \log x_i$ (la fonction entropie est donc *concave*).

Théorème 5. – *Si f est différentiable, alors f est convexe sur U si et seulement si*

$$\forall x, y \in U \quad f(y) - f(x) \geq \mathrm{D}f(x)(y - x)$$

ce qui revient à dire (si $n = 1$) que le graphe de f est au-dessus des tangentes.

– *Si f est deux fois différentiable, alors f est convexe sur U si et seulement si $\mathrm{D}^2 f$ est une forme quadratique positive en tout point, c'est-à-dire*

$$\forall x \in U, h \in \mathbb{R}^n \quad \mathrm{D}^2 f(x)(h, h) \geq 0\,.$$

Exercice 5. Prouver le Théorème 5.

En utilisant ce théorème, il est donc aisé d'obtenir des conditions suffisantes du type de celles du Théorème 4 permettant de prouver l'existence d'un minimum global (et non local).

Exercice 6. Soit $f\colon U \to \mathbb{R}$ une fonction sur U convexe ouvert de $\mathbb{R}^n$.

1. Si f est convexe et différentiable sur U et que $\mathrm{D}f(a) = 0$, montrer que f admet en a un minimum global sur U.
2. En déduire que si f est deux fois différentiable et vérifie les conditions

$$\mathrm{D}f(a) = 0 \qquad \text{et} \qquad \forall x \in U, h \in \mathbb{R}^n \quad \mathrm{D}^2 f(x)(h, h) \geq 0$$

alors f admet en a un minimum global.

Dans le cas d'une fonction convexe f, on peut trouver une condition simple permettant de trouver une direction de descente dans l'algorithme de descente.

Exercice 7. Montrer, en utilisant le théorème 5, que la direction de descente d_k doit vérifier $\nabla f(x_k)^{\mathrm{T}} d_k < 0$.

Un choix raisonnable consiste donc à choisir $d_k = -\nabla f(x_k)$: dans ce cas, on parle de descente de gradient. Ce choix est aussi utilisable dans le cas général. Il reste à choisir la taille de pas σ_k. On peut faire une recherche exhaustive, à savoir chercher la taille σ minimisant $f(x_k - \sigma \nabla f(x_k))$. C'est parfois trop coûteux et donc on se permet souvent de faire une approximation sur cette étape.

Sous certaines conditions sur f (par exemple forte convexité), on peut prouver que cette méthode converge quel que soit le choix de x_0 et borner sa vitesse de convergence : il existe une constant $c \in]0, 1[$ (dépendant de f) tel que pour tout $k \leq 0$,

$$f(x_k) - f(x^*) \leq c^k (f(x_0) - f(x^*)) \,.$$

Cet algorithme est relativement lent, même s'il est abondamment utilisé car ne nécessite pas de calculs trop coûteux ou de mise en œuvre compliquée. Citons par ailleurs que l'algorithme rapide usuel est la méthode de Newton, une méthode de descente qui se base sur le calcul de la Hessienne (coûteux en général) plutôt que sur le gradient uniquement.

3.3 Gradient stochastique

Si l'on suppose que la fonction à minimiser s'écrit comme une somme de fonctions convexes $f = \sum_{i=1}^{N} f_i$, alors une méthode permettant de réduire potentiellement le coût en calcul consiste à considérer chaque fonction f_i aléatoirement (ou alternativement en cycle). Ainsi, la mise à jour devient $x_{k+1} = x_k - \sigma_k \Delta f_i(x_k)$ avec i tiré aléatoirement parmi $\{1, \ldots, N\}$. On peut également choisir un petit nombre de fonctions f_i à chaque mise à jour.

De retour à l'exemple mentionné en (1), cela revient à dire qu'on considère chaque donnée (x_i, y_i) aléatoirement à chaque étape.

Références

[1] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge University Press, 2004. Disponible en ligne à l'adresse http://www.stanford.edu/~boyd/cvxbook/.

[2] François Rouvière. *Petit guide de calcul différentiel à l'usage de la licence et de l'agrégation*. Cassini, 2003. Deuxième édition revue et augmentée.