

Programmation stochastique

Rappels d'optimisation

Fabian Bastin

IFT-6512 – Hiver 2011

Dérivées

- Soit $f : \mathbb{R}^n \rightarrow \mathbb{R}$. La dérivée directionnelle f' de f dans la direction d est

$$f'(x, d) = \lim_{\epsilon \rightarrow 0} \frac{f(x + \epsilon d) - f(x)}{\epsilon}.$$

- f est dite différentiable si cette dérivée directionnelle existe et a la même valeur pour tout $d \in \mathbb{R}^n$. L'unique valeur de cette dérivée est appelée le gradient de f en x , dénotée $\nabla_x f(x)$.
- Si f est différentiable, nous avons aussi

$$f'(x, d) = \nabla_x f(x)^T d.$$

- Mais toutes les fonction que nous considérerons ne seront pas différentiables.

- Un vecteur $\eta \in \mathbb{R}^n$ est un **sous-gradient** d'une fonction convexe f en un point x si et seulement si

$$f(y) \geq f(x) + \eta^T(y - x), \quad \forall y \in \mathbb{R}^n.$$

Le graphe de la fonction (linéaire) $h(y) = f(x) + \eta^T(y - x)$ est un hyperplan de support à l'ensemble convexe $\text{epi}(f)$ au point $(x, f(x))$.

- L'ensemble de tous les sous-gradients de f en x est appelé le sous-différentiel de f en x , dénoté par $\partial f(x)$.
- Théorème : $\eta \in \partial f(x)$ si et seulement si

$$f'(x, d) \geq \eta^T d, \quad \forall d \in \mathbb{R}^n.$$

- Pourquoi ne considère-t-on ici que des fonctions convexes ? Lien avec le cas différentiable : soit $\phi : \mathbb{R}^n \rightarrow \mathbb{R} \in \mathcal{C}^1$. ϕ est convexe ssi $\forall x, y \in \mathbb{R}^n$,

$$(y - x)^T \nabla_x \phi(x) \leq \phi(y) - \phi(x).$$

Hyperplan ? Généralement de la notion de plan. . .

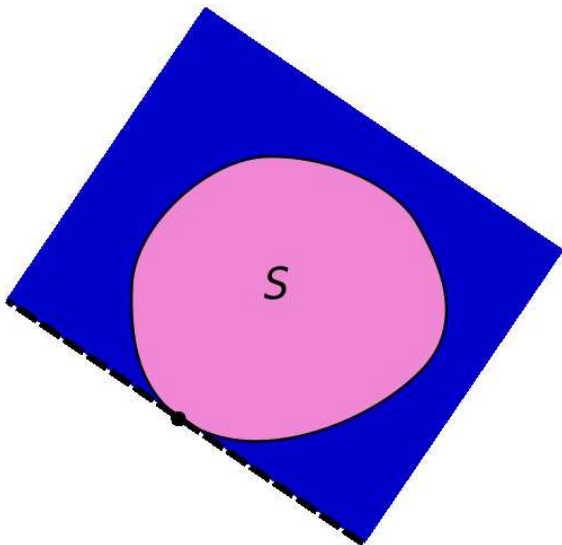
On parle ici d'hyperplan affine. Un hyperplan dans l'espace \mathbb{R}^n peut être décrit par l'équation

$$\sum_{i=1}^n a_i x_i = b.$$

Au moins un des a_i , $i = 1, \dots, n$, est non nul.

Hyperplan de support : généralisation de la notion de tangente. Un hyperplan de support à f en x est un hyperplan l'intersection avec f dans un voisinage de x est le singleton $\{x\}$, excepté dans les directions où f est linéaire, auquel cas la tangente se confond avec le graphe de f .

Hyperplan de support (2)



Source : Wikipedia.

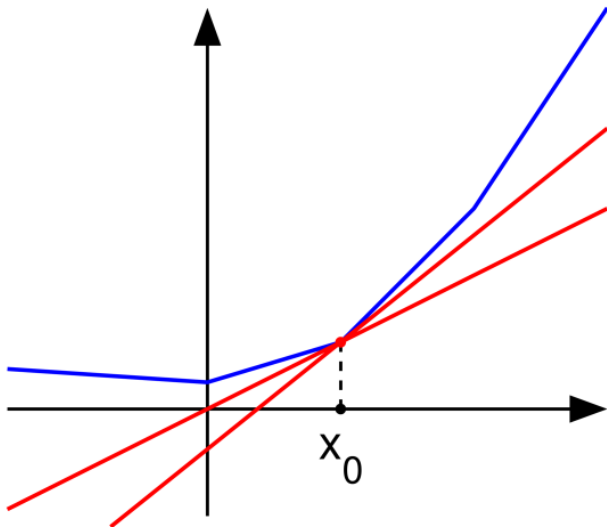
- Le sous-différentiel est toujours un ensemble compact non-vide et convexe.
- En une dimension, le sous-différentiel de f en x_0 est l'intervalle $[a, b]$, avec

$$a = \lim_{x \rightarrow x_0^-} \frac{f(x) - f(x_0)}{x - x_0},$$

$$b = \lim_{x \rightarrow x_0^+} \frac{f(x) - f(x_0)}{x - x_0}.$$

- Si f est différentiable en x , $\partial f(x) = \{\nabla_x f(x)\}$.

Illustration



Source : Wikipedia.

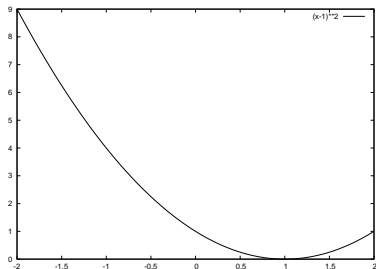
- Sous quelles conditions une solution est-elle optimale ?
- Beaucoup d'algorithmes visent à trouver des points qui satisfont ces conditions.
- On peut souvent apprendre beaucoup sur un problème en observant les propriétés de ses solutions optimales.
- Dans un premier temps, considérons la minimisation de fonction qui sont
 - unidimensionnelle ;
 - continue au sens de Lipschitz :

$$|f(a) - f(b)| \leq L|a - b|;$$

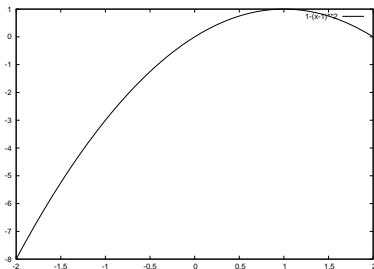
- différentiable.

Conditions nécessaires

Condition nécessaire pour que x soit une solution optimale :
 $f'(x) = 0$. Mais ce n'est pas une condition suffisante !



$$f(x) = (x - 1)^2$$



$$f(x) = 1 - (x - 1)^2$$

Condition suffisante pour que x soit (localement) optimal :
 $f''(x) > 0$.
Cela revient à exiger que $f(x)$ soit strictement convexe en x .

Soit $f : \mathbb{R}^n \rightarrow \mathbb{R} \in C^2$ (i.e. f est deux fois continûment différentiable).

- **Condition nécessaire** : si x^* est un optimum (local), $\nabla_x f(x^*) = 0$.
- Mais $\nabla_x f(x) = 0$ peut correspondre à un maximum (local) ou même à un point selle !
- **Condition suffisante** : f strictement convexe en x^* et $\nabla_x f(x^*) = 0$. Ou f convexe sur \mathbb{R}^n et $\nabla_x f(x^*) = 0$. Dans ce dernier cas, nous avons pour tout $x \in \mathbb{R}^n$

$$0 = (x - x^*)^T \nabla_x f(x^*) \leq f(x) - f(x^*) \quad (\text{convexité}),$$

et donc

$$f(x^*) \leq f(x), \quad \forall x \in \mathbb{R}^n.$$

Considérons le problème (unidimensionnel) suivant :

$$\min_{0 \leq x \leq u} f(x).$$

Il y a trois cas, pour lesquels une solution optimale pourrait être $x = 0$, $0 < x < u$, $x = u$.

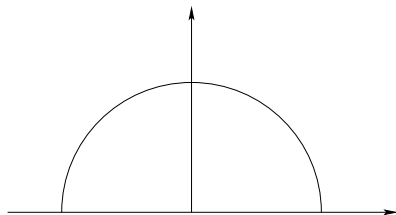
- Si $0 < x < u$, alors les conditions nécessaires et les conditions suffisantes d'optimalité sont les mêmes que pour le cas sans contraintes.
- Si $x = 0$, nous devons avoir $f'(x)|_{x=0} \geq 0$ (nécessaire), $f''(x)|_{x=0} > 0$.
- Si $x = u$, nous devons avoir $f'(x)|_{x=u} \leq 0$ (nécessaire), $f''(x)|_{x=u} > 0$.

- Généralisation à des problèmes à plus d'une variable et avec des contraintes plus générales.
- Intuition : si une contrainte est active (elle tient avec une égalité), alors le gradient de la fonction objectif doit pointer de manière que la fonction objectif serait améliorée en sortant de la région de faisabilité.
- Formellement : l'opposé du gradient de la fonction objectif doit être une combinaison linéaire des gradients des contraintes actives.
- KKT tient pour [Karush-Kuhn-Tucker](#).
Karush-Kuhn-Tucker ou Kuhn-Tucker ? Même chose, juste un problème de reconnaissance d'auteur, Karush ayant développé les conditions avec Kuhn-Tucker, mais son papier est resté longtemps ignoré.

Exemple ($x \in \mathbb{R}^2$)

Considérons le problème

$$\begin{aligned} \min f(x) &= x_1 + x_2 \\ \text{t.q. } x_1^2 + x_2^2 &\leq 2 \\ -x_2 &\leq 0. \end{aligned}$$



La solution optimale est $(-\sqrt{2}, 0)$. Le gradient de f est $(1, 1)$.

Nous considérons le problème général

$$\begin{aligned} \min_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x}) \\ \text{t.q. } c_i(\mathbf{x}) = 0, \quad i \in \mathcal{E}, \\ c_i(\mathbf{x}) \geq 0, \quad i \in \mathcal{I}. \end{aligned}$$

Ensemble réalisable :

$$\mathcal{C} = \{\mathbf{x} \mid \text{t.q. } c_i(\mathbf{x}) = 0, \quad i \in \mathcal{E}; \quad c_i(\mathbf{x}) \geq 0, \quad i \in \mathcal{I}\}.$$

Conditions critiques au premier ordre

Lagrangien du problème avec contraintes :

$$\mathcal{L}(\mathbf{x}, \lambda) = f(\mathbf{x}) - \sum_{i \in \mathcal{E} \cup \mathcal{I}} \lambda_i c_i(\mathbf{x}).$$

Conditions de Karush-Kuhn-Tucker :

$$\begin{aligned}\nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}^*, \lambda^*) &= \mathbf{0}, \\ c_i(\mathbf{x}^*) &= 0, \quad \forall i \in \mathcal{E}, \\ c_i(\mathbf{x}^*) &\geq 0, \quad \forall i \in \mathcal{I}, \\ \lambda_i^* &\geq 0, \quad \forall i \in \mathcal{I}, \\ \lambda_i^* c_i(\mathbf{x}^*) &= 0, \quad \forall i \in \mathcal{E} \cup \mathcal{I}.\end{aligned}$$

Conditions critiques au premier ordre (2)

La première équation des conditions KKT implique que

$$\nabla_x f(\mathbf{x}^*) = \sum_{i \in \mathcal{A}(\mathbf{x}^*)} \lambda_i^* \nabla_x c_i(\mathbf{x}^*),$$

où

$$\mathcal{A}(\mathbf{x}^*) = \mathcal{E} \cup \{i \in \mathcal{I} \mid c_i(\mathbf{x}) = 0\}.$$

$\mathcal{A}(\mathbf{x}^*)$ est l'ensemble des **contraintes actives**.

Géométriquement, on observe que si \mathbf{x}^* est une solution optimale, alors $\nabla_x f(\mathbf{x}^*)$ est combinaison linéaire des contraintes actives. Si la contraintes est non-active, son "poids" est nul.

Conditions critiques au premier ordre (3)

Les λ_i sont appelés multiplicateurs de Lagrange, ou variables duales. x est le vecteur des variables primales.

Les conditions KKT sont des conditions nécessaire si une **qualification de contraintes** tient en x^* .

La qualification de contraintes la plus utilisée est la LICQ (Linear Independence Constraint Qualification - qualification de contraintes d'indépendance linéaire).

La LICQ tient si l'ensemble des gradients des contraintes actives $\{\nabla_x c_i(x^*), i \in \mathcal{A}(x^*)\}$ est linéairement indépendant.

La condition

$$\lambda_i^* c_i(\mathbf{x}^*) = 0, \quad \forall i \in \mathcal{E} \cup \mathcal{I},$$

est connue sous le nom de condition de complémentarité, puisqu'elle implique que pour $i \in \mathcal{E} \cup \mathcal{I}$, $\lambda_i^* = 0$ ou $c_i(\mathbf{x}^*) = 0$.

Cas particulier : **complémentarité stricte**.

Etant donné une solution \mathbf{x}^* associée au vecteur de Lagrange λ^* , la condition de complémentarité stricte tient si exactement un des λ_i^* et $c_i(\mathbf{x}^*)$ est nul, pour chaque index $i \in \mathcal{I}$.

Autrement dit, nous avons $\lambda_i^* > 0, \forall i \in \mathcal{I} \cap \mathcal{A}(\mathbf{x}^*)$.

Un peu de maths...

Problème non-convexe : nous avons généralement besoin de connaître les dérivées seconde de f (... hypothèse : $f \in C^2$).
Soit $p = \#\mathcal{I}$ et $m = \#\mathcal{E}$. Pour des vecteurs donnés $\mathbf{x} \in \mathbb{R}^n$ et $\lambda \in \mathbb{R}^{p+m}$, l'ensemble $\mathcal{N}_+(\mathbf{x}, \lambda)$ est défini par

$$\mathcal{N}_+(\mathbf{x}, \lambda) \stackrel{\text{def}}{=} \left\{ \mathbf{w} \in \mathbb{R}^n \mid \begin{array}{l} \nabla_{\mathbf{x}} c_i(\mathbf{x})^T \mathbf{w} = 0 \quad \forall i \in \mathcal{E} \cup \{j \in \mathcal{A}(\mathbf{x}) \cap \mathcal{I} : \lambda_j > 0\} \\ \text{et } \nabla_{\mathbf{x}} c_i(\mathbf{x})^T \mathbf{w} \geq 0 \quad \forall i \in \{j \in \mathcal{A}(\mathbf{x}) \cap \mathcal{I} : \lambda_j = 0\} \end{array} \right\}.$$

Theorem (Conditions nécessaires au second-ordre)

Supposons que x^ est une solution locale de notre problème d'optimisation et qu'une qualification de contrainte tient en x^* . Soit λ^* un vecteur de multiplicateurs de Lagrange tel que les conditions KKT soient satisfaites. La courbure du Lagrangien le long des directions dans $\mathcal{N}_+(x^*, \lambda^*)$ doit être non-négative, c'est-à-dire*

$$w^T \nabla_{xx}^2 \mathcal{L}(x^*, \lambda^*) w \geq 0, \text{ for all } w \in \mathcal{N}_+(x^*, \lambda^*).$$

Conditions suffisantes au second ordre

En renforçant l'inégalité du théorème précédent, nous pouvons dériver des conditions suffisantes.

Theorem (Conditions suffisantes au second ordre)

Supposons que pour un certain point réalisable $x^ \in \mathbb{R}^n$, il existe un vecteur de Lagrange λ^* tel que les conditions KKT soient satisfaites. Supposons de plus que*

$$w^T \nabla_{xx}^2 \mathcal{L}(x^*, \lambda^*) w > 0, \forall w \in \mathcal{N}_+(x^*, \lambda^*) \text{ avec } w \neq 0.$$

Alors x^ est une solution locale stricte de notre problème d'optimisation.*

Note : ces conditions suffisantes sont valables sans nécessiter de qualification de contraintes.

$$\min x_1 + x_2,$$

tel que

$$\begin{aligned}x_1^2 + x_2^2 &\leq 2 && (\lambda), \\ -x_2 &\leq 0 && (\mu).\end{aligned}$$

Conditions KKT

Faisabilité primale

$$\begin{aligned}x_1^2 + x_2^2 &\leq 2, \\ -x_2 &\leq 0.\end{aligned}$$

Faisabilité duale

$$\begin{aligned}\lambda &\geq 0, \\ \mu &\geq 0, \\ -1 &= \lambda(2x_1), \\ -1 &= \lambda(2x_2) - \mu.\end{aligned}$$

Conditions de complémentarité

$$\begin{aligned}\lambda(2 - x_1^2 - x_2^2) &= 0, \\ \mu x_2 &= 0.\end{aligned}$$

Un point de vue géométrique

Rappel : ensemble réalisable dénoté par \mathcal{A} . On supposera simplement ici que \mathcal{A} est fermé.

Condition nécessaire : si \mathbf{x}^* est un minimum local,

$$-\nabla_{\mathbf{x}} f(\mathbf{x}^*) \in \mathcal{N}_{\mathcal{A}}(\mathbf{x}^*),$$

où $\mathcal{N}_{\mathcal{A}}(\mathbf{x}^*)$ est le cône normal à \mathcal{A} en \mathbf{x}^* .

Question : définition d'un cône ? (Encore des) définitions. . .

Un sous-ensemble C d'un espace vectoriel V est un **cône** (linéaire) si et seulement si αx appartient à C pour n'importe quel $x \in C$ et n'importe quel scalaire strictement positif α de V . Il est pointé si α peut être nul.

Un **cône convexe** est un cône qui est fermé sous les combinaisons convexe, i.e. si et seulement $\alpha x + \beta y \in C, \forall \alpha, \beta$ non-négatifs, avec $\alpha + \beta = 1$.

Cône tangent, cône normal

Cadre général. Nous dirons tout d'abord qu'un vecteur $w \in \mathbb{R}^n$ est tangent à \mathcal{A} en $x \in \mathcal{A}$ si pour toutes les séquences de vecteur $\{x_i\}$ avec $x_i \rightarrow x$, et $x_i \in \mathcal{A}$, et toutes les séquences de scalaires positifs $t_i \downarrow 0$, il existe une séquence $w_i \rightarrow w$ tel que $x_i + t_i w_i \in \mathcal{A}$ pour tout i .

Le **cône tangent** $T_{\mathcal{A}}(x)$ est la collection de tous les vecteurs tangents à \mathcal{A} en x .

Le **cône normal** $N_{\mathcal{A}}(x)$ est le complément orthogonal au cône tangent, c'est-à-dire

$$N_{\mathcal{A}}(x) = \{v \mid v^T w \leq 0, \forall w \in T_{\mathcal{A}}(x)\}.$$

Oui, mais, en pratique ???

Cône tangent, cône normal (2)

La définition du cône normal se simplifie fortement si \mathcal{A} est convexe. En effet, on a alors

$$N_{\mathcal{A}}(x) = \{v \mid v^T(x - x_0) \geq 0, \forall x_0 \in \mathcal{A}\}.$$

La condition nécessaire d'optimalité devient alors $\nabla_x f(\hat{x})(x - \hat{x}) \geq 0, \forall x \in \mathcal{A}$.

Si f est convexe (i.e. on est dans le cadre de la programmation convexe), cette condition est de plus suffisante.

Généralisation aux fonctions non-différentiables

- Les choses se compliquent fortement, en particulier dans le cas non-convexe. Nous n'aborderons que le cas convexe. . . sans entrer dans les détails de l'analyse convexe (sauf si vous y tenez vraiment!).
- Globalement, cela revient à remplacer $\nabla_x f(x) = 0$ par $0 \in \partial f(x)$.

Theorem

Soit une fonction convexe $f : \mathbb{R}^n \rightarrow \mathbb{R}$, et soit S un ensemble convexe non vide. $\hat{x} \in \arg \min_{x \in S} f(x)$ si (et seulement si) η est un sous-gradient de f en \hat{x} tel que $\eta^T(x - \hat{x}) \geq 0, \forall x \in S$.

Comparaison avec le cas différentiable : la condition est $\nabla_x f(\hat{x})(x - \hat{x}) \geq 0, \forall x \in S$ (cf interprétation géométrique).

Démonstration.

On ne prouvera que la partie facile. . .

Si η est un sous-gradient de f en \hat{x} tel que $\eta^T(x - \hat{x}) \geq 0$,
 $\forall x \in S$, alors

$$f(x) \geq f(\hat{x}) + \eta^T(x - \hat{x}) \geq f(\hat{x}), \quad \forall x \in S.$$

Par conséquent,

$$\hat{x} \in \arg \min_{x \in S} f(x).$$



Theorem

Soit une fonction convexe $f : \mathbb{R}^n \rightarrow \mathbb{R}$. $\hat{x} \in \arg \min_{x \in \mathbb{R}^n} f(x)$ si (et seulement si) $0 \in \partial f(\hat{x})$.

Démonstration.

- $\hat{x} \in \arg \min_{x \in \mathbb{R}^n} f(x)$ si et seulement si η est un sous-gradient avec $\eta^T(x - \hat{x}) \geq 0, \forall x \in \mathbb{R}^n$.
- Prenons $x = \hat{x} - \eta$. On a

$$0 \leq \eta^T(\hat{x} - \eta - \hat{x}) = -\eta^T\eta \leq 0.$$

Dès lors, $\eta = 0$.

- Ainsi, $\eta = 0 \in \partial f(\hat{x})$.



Theorem

Pour une fonction convexe $f : \mathbb{R}^n \rightarrow \mathbb{R}$ et des fonctions convexes $g_i : \mathbb{R}^n \rightarrow \mathbb{R}$, $i = 1, 2, \dots, m$, sous certaines conditions de régularité, \hat{x} est une solution optimale du problème

$$\min f(x) \text{ t.q. } g_i(x) \leq 0, \quad i = 1, 2, \dots, m$$

si et seulement si les conditions suivantes tiennent :

- $g_i(x) \leq 0$, $i = 1, 2, \dots, m$,
- $\exists \lambda_1, \dots, \lambda_m \in \mathbb{R}$ tels que
 - $0 \in \partial f(\hat{x}) + \sum_{i=1}^m \lambda_i \partial g_i(\hat{x})$,
 - $\lambda_i \geq 0$, $i = 1, \dots, m$,
 - $\lambda_i g_i(\hat{x}) = 0$, $i = 1, \dots, m$.

Que signifie $\sum_{i=1}^m \lambda_i \partial g_i(\hat{x})$ quand le sous-différentiel n'est pas réduit à un point (ce qui serait le cas si on pouvait parler de gradient) ?

Que signifie additionner des ensembles ? Etant donné deux ensembles C_1 et C_2 ,

$$C_1 + C_2 = \{x_1 + x_2 \mid x_1 \in C_1, x_2 \in C_2\}.$$

Cette opération est aussi appelée **somme de Minkowski**.

Le vendeur de journaux (le retour...)

Retour à l'exemple du cours précédent.

Nous voulons résoudre le problème

$$\max_{x \geq 0} -cx + Q(x),$$

où

$$Q(x) = qx - (q - r) \int_{-\infty}^x F(\omega) d\omega.$$

Les conditions KKT sont ici très simples...

Il suffit d'annuler le gradient de l'objectif, en notant que

$$Q'(x) = q - (q - r)F(x),$$

aussi, il faut résoudre

$$-c + q - (q - r)F(x) = 0.$$

Le vendeur de journaux (suite)

La solution x^* est dès lors

$$x^* = F^{-1} \left(\frac{q - c}{q - r} \right).$$

Un exemple : $c = 0.15$, $q = 0.25$, $r = 0.02$, $\omega \sim N(650, 80)$.

Alors

$$x^* = N^{-1}(-0.1/0.23) \approx 636.86.$$

Autre interprétation, plus intuitive : supposons que le vendeur a acheté t journaux. Quel est le revenu marginal espéré s'il achète un journal supplémentaire ? D'un point de vue économique, nous souhaiterions que ce revenu marginal soit égal à 0. . . comme explicité par les conditions KKT !

Dénotons le revenu marginal espéré par MR ; on a

$$\begin{aligned}MR(t) &= -c + qP[\omega \geq t] + rP[\omega \leq t] \\ &= -c + q(1 - F(t)) + rF(t).\end{aligned}$$

En annulant cette expression, on obtient

$$MR(t) = 0 \text{ ssi } F(t) = \frac{q - c}{q - r},$$

et donc on retrouve la solution précédente.

Pour des compléments, consultez Linderoth.