Cours Apprentissage - ENS Math/Info
Analyse Convexe

Francis Bach

8 Octobre 2015

Ce cours s'appuie sur le livre "Convex Optimization" de Stephen Boyd et Lieven Vandenberghe (disponible gratuitement : `http://www.stanford.edu/~boyd/cvxbook/`).

La convexité intervient dans de nombreuses branches des mathématiques et de l'informatique. Deux aspects seront vus dans le cours d'apprentissage : l'*analyse* convexe (propriétés des fonctions et problèmes d'optimisation convexes) et l'*optimisation* convexe (algorithmes de résolution).

Dans le cadre de l'apprentissage statistique, la convexité permet d'obtenir des problèmes d'optimisation *bien posés*, pour lesquels il est possible d'obtenir des solutions par des algorithmes efficaces.

Exemple classique en apprentissage : minimisation du risque empirique régularisé

$$\min_{f \in \mathcal{F}} \ \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f(x_i)) + \lambda \Omega(f),$$

avec
- $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$, $i = 1, \dots, n$ données d'apprentissage
- $\mathcal{F}$: ensemble convexe de prédicteurs $f : \mathcal{X} \to \mathbb{R}$
- $u \mapsto \ell(y, u)$ perte convexe pour tout $y \in \mathcal{Y}$
- Ω pénalité convexe.

La convexité servira typiquement à (a) analyser la solution et ses propriétés statistiques de généralisation et (b) déterminer des algorithmes d'estimation.

1 Ensembles convexes

On ne considère dans ce cours que la convexité dans un espace Euclidien de dimension finie (le plus généralement $\mathbb{R}^n$).

- **Définition** : $K \subset \mathbb{R}^n$ est convexe si et seulement si, pour tout $x, y \in K$, le segment $[x, y]$ est inclus dans K, i.e., $\forall \alpha \in [0, 1]$, $\alpha x + (1 - \alpha) y \in K$. Attention au domaine (qui doit être un ensemble convexe).

- **Exemples classiques** : hyperplan $a^\top x = b$ ($a \in \mathbb{R}^n$, $a \neq 0$, $b \in \mathbb{R}$), demi-espace $a^\top x \geqslant b$, sous-espace affine $Ax = b$, boules $\{\|x\| \leqslant 1\} \subset \mathbb{R}^n$, cone $\{\|x\| \leqslant t\} \subset \mathbb{R}^{n+1}$. Polyédres (intersections de demi-espaces) et polytopes (polièdres compacts).

— **Propriétés** : l'intersection d'une famille (non nécessairement dénombrables) de convexes est convexe ; la convexité est préservée par les applications affines (image et image inverse).

— **Enveloppe convexe** : Etant donné un ensemble A, l'enveloppe convexe est le plus petit ensemble convexe contenant A. Elle est égale à l'intersection de tous les convexes contenant A. Elle est égale à l'ensemble des barycentres à coefficients positifs ou nuls de familles finies de points de A (i.e., $\sum_{i=1}^p \alpha_i x_i$, pour $x_i \in A$, $\alpha_i \geqslant 0$ et $\sum_{i=1}^p \alpha_i = 1$).

— **Séparation des convexes** : Si C et D sont deux ensemble convexes disjoints ($C \cap D = \varnothing$), il existe un hyperplan séparant C et D, i.e., $\exists a \neq 0$ et $b \in \mathbb{R}$ tels que $C \subset \{a^\top x \geqslant b\}$ et $D \subset \{a^\top x \leqslant b\}$ (forme géométrique du théorème de Hahn-Banach). Si C et D sont compacts, alors il existe une séparation stricte, i.e., $C \subset \{a^\top x > b\}$ et $D \subset \{a^\top x < b\}$.

Exercice : Montrer le théorème de séparation stricte quand C et D sont compacts (indication : on utilisera la paire (x, y) minimisant $\|x - y\|^2$ pour $(x, y) \in C \times D$ et la médiatrice des points x et y).

2 Fonctions convexes

— **Définition** : Une fonction f définie sur $D \subset \mathbb{R}^n$ est convexe ssi (a) D est convexe et (b) pour tout $x, y \in D$, et $\alpha \in [0, 1]$, alors $f(\alpha x + (1 - \alpha)y) \leqslant \alpha f(x) + (1 - \alpha)f(y)$.

— **Convexité stricte** : même définition sauf : si $\alpha \in (0, 1)$, $f(\alpha x+(1-\alpha)y) < \alpha f(x)+(1-\alpha)f(y)$

— **Convexité forte** : même définition sauf : $f(\alpha x + (1 - \alpha)y) \leqslant \alpha f(x) + (1 - \alpha)f(y) - \frac{\mu}{2}\alpha(1 - \alpha)\|x - y\|^2$. NB : f est μ-fortement convexe ssi $x \mapsto f(x) - \frac{\mu}{2}x^\top x$ est convexe.

— **Exemples classiques en une dimension** : x, x^2, $-\log x$, e^x, $\log(1 + e^{-x})$, $|x|^p$ pour $p \geqslant 1$, $-x^p$ pour $p < 1$ et $x \geqslant 0$.

— **Exemples classiques en dimension supérieure** : fonctions linéaires $a^\top x$, fonctions quadratiques $\frac{1}{2}x^\top Qx$ pour Q symmétrique semidéfinie positive (rappel de la définition : toutes les valeurs propres positives ou nulles, ou $\forall x \in \mathbb{R}^n$, $x^\top Qx \geqslant 0$), normes, $\max\{x_1, \dots, x_n\}$, $\log(\sum_i e^{x_i})$.

— **Caractérisation pour f dérivable** : $\forall x, y \in D,\ f(x) \geqslant f(y) + f'(y)^\top(x - y)$.

— **Caractérisation pour f deux fois dérivable** : $\forall x \in D,\ f''(x)$ semidéfinie positive.

— **Opérations préservant la convexité** : supremum d'une famille de fonctions convexes $\sup_{i \in I} f_i(x)$, combinaison linéaires positives, minimisation partielle $\inf_{x \in C} f(x, y)$ (si f est convexe sur $C \times D$).

— **Comment vérifier qu'une fonction est convexe ou non-convexe ? :** Le plus souvent, aucun calcul n'est nécessaire. Une fonction est convexe ssi elle est convexe sur tout segment inclus dans son domaine.

— **Propriétés** : (a) f est continue sur l'intérieur de D, (b) f est convexe ssi l'épigraphe de f est convexe, i.e., $\{(x, t) \in \mathbb{R}^{n+1},\ f(x) \leqslant t\}$ est convexe.

— **Inégalité de Jensen** : $f(\sum_{i=1}^n \alpha_i x_i) \leqslant \sum_{i=1}^n \alpha_i f(x_i)$ et $f(\mathbb{E}X) \leqslant \mathbb{E}f(X)$. Attention au sens de l'inégalité.

— **Fonctions convexes étendues** (à valeurs dans $\mathbb{R} \cup \{+\infty\}$) : $\tilde{f} : \mathbb{R}^n \mapsto \mathbb{R} \cup \{+\infty\}$ finie sur son domaine, infinie sur son complément. Permet de gérer simplement les fonctions à domaine $D \neq \mathbb{R}^n$.

3 Problèmes d'optimisation non-contraints

Attention à la différence entre "min", "inf" et la définition d'un problème d'optimisation.

On suppose f convexe et finie sur $\mathbb{R}^n$. Alors, les trois cas exclusifs suivants sont possibles :
- $\inf_{x\in\mathbb{R}^n} f(x) = -\infty$: pas de minimum (exemple f linéaire)
- $\inf_{x\in\mathbb{R}^n} f(x) > -\infty$ non atteint (exemple $\log(1+e^{-x})$)
- $\inf_{x\in\mathbb{R}^n} f(x) > -\infty$ atteint, et alors égale à $\min_{x\in\mathbb{R}^n} f(x)$ (exemple le plus classique) : f est dite *coercive* ($\lim_{\|x\|\to+\infty} f(x) = +\infty$)

Minimas locaux vs. minimas globaux : x est minimum local ssi il existe un voisinage (ouvert) V de x tel que x est le minimum de f sur V. Lorsque f est convexe, tout minimum local est global.

Stricte convexité et minimum unique : si f est strictement convexe, alors il y a au plus un minimum.

Condition nécessaire et suffisante d'optimalité (cas dérivable) : Si f est convexe et dérivable, x est un minimum de f sur $\mathbb{R}^n$ si et seulement si $f'(x) = 0$. Attention à la différence entre minimum local et point stationnaire.

4 Problèmes d'optimisation contraints

On suppose f convexe et finie sur $D \subset \mathbb{R}^n$. On cherche à minimiser f sur un convexe $C \subset D$.

L'ensemble de contraintes C peut être spécifié par une intersection d'ensembles $h_i(x) = 0$ et $g_j(x) \leqslant 0$ (voir section suivante).

Minimisation d'une fonction linéaire sur une enveloppe convexe : Soit A un compact de $\mathbb{R}^n$ et $a \in \mathbb{R}^n$, $a \neq 0$. Alors

$$\min_{x\in A} a^\top x = \min_{x\in \text{ Enveloppe convexe}(A)} a^\top x$$

Exemple classique du problème d'affectation : on a p employés et p tâches, et à chaque paire employé/tâche (i,j), on a un coût c_{ij}, le but est de trouver une permutation $\sigma : \{1,\dots,p\} \mapsto \{1,\dots,p\}$ telle que $\sum_{i=1}^p c_{i\sigma(i)}$ est minimum. On a $\sum_{i=1}^p c_{i\sigma(i)} = \langle c, M_\sigma\rangle$ où M_σ est la matrice de permutation associée. L'enveloppe convexe des matrices de permutations est l'ensemble des matrices doublement stochastiques (théorème de Birkhoff), qui correspond à un problème d'optimisation comvexe contraint.

5 Dualité Lagrangienne

On s'intéresse au problème d'optimisation suivant (dit problème *primal*) :

$$\min_{x\in D} f(x) \quad \text{tel que } \forall i \in \{1,\dots,m\}, h_i(x) = 0 \text{ et } \forall j \in \{1,\dots,r\}, g_j(x) \leqslant 0.$$

On note D^* l'ensemble des $x \in D$ vérifiant les contraintes.

On utilisera l'exemple canonique de la projection orthogonale d'un point z sur le demi-espace $\{a^\top x \leqslant b\}$ ou l'hyper-espace $\{a^\top x = b\}$.

— **Définition du Lagrangien** : on appelle Lagrangien la fonction $\mathcal{L} : \mathbb{R}^m \times \mathbb{R}_+^r$ définie par

$$\mathcal{L}(x, \lambda, \mu) = f(x) + \lambda^\top h(x) + \mu^\top g(x).$$

λ et μ sont appelés multiplicateurs de Lagrange (ou variables duales).

— **Problème primal comme supremum du Lagrangien par rapport aux variables duales** : pour tout $x \in D$,

$$\sup_{(\lambda,\mu)\in\mathbb{R}^m\times\mathbb{R}_+^r} \mathcal{L}(x, \lambda, \mu) = \begin{cases} f(x) \text{ si } x \in D^* \\ +\infty \text{ sinon} \end{cases}$$

(Preuve utilisant le minimum d'une fonction linéaire sur $\mathbb{R}$ ou $\mathbb{R}_+$)
Le problème primal est donc équivalent à

$$p^* = \inf_{x\in D} \sup_{(\lambda,\mu)\in\mathbb{R}^m\times\mathbb{R}_+^r} \mathcal{L}(x, \lambda, \mu).$$

— **Fonction duale** : $q : \mathbb{R}^m \times \mathbb{R}_+^r \to \mathbb{R}$ définie par $q(\lambda, \mu) = \inf_{x\in D} \mathcal{L}(x, \lambda, \mu)$. Le problème dual est la minimisation de q sur $\mathbb{R}^m \times \mathbb{R}_+^r$, équivalent à

$$d^* = \sup_{(\lambda,\mu)\in\mathbb{R}^m\times\mathbb{R}_+^r} \inf_{x\in D} \mathcal{L}(x, \lambda, \mu).$$

— **Concavité du problème dual** : sans aucune hypothèses sur D, f, g, h, la fonction duale q est concave.

— **Dualité faible** : sans aucune hypothèses sur D, f, g, h, pour tout $(\lambda, \mu) \in \mathbb{R}^m \times \mathbb{R}_+^r$, et $x \in \mathcal{D}^*$,

$$\inf_{x'\in D} \mathcal{L}(x', \lambda, \mu) \leqslant \mathcal{L}(x, \lambda, \mu) \leqslant \sup_{(\lambda',\mu')\in\mathbb{R}^m\times\mathbb{R}_+^r} \mathcal{L}(x, \lambda', \mu')$$

ce qui implique $q(\lambda, \mu) \leqslant f(x)$. Ceci implique $d^* \leqslant p^*$.
NB : on peut toujours inverser sup et inf : "$\sup\inf \leqslant \inf\sup$"

— **Problèmes non faisables** ($D^* = \varnothing \Leftrightarrow d^* = \inf$), **non-bornés** ($p^* = -\infty$)

Interprétation géometrique : problème à une contrainte d'inégalité (intuition pour l'hypothèse de convexité) Pour le problème $\min_{x\in D} f(x)$ tel que $g(x) \leqslant 0$, on considère l'ensemble $A = \{(u, t) \in \mathbb{R}^2,\ \exists x \in D, f(x) \leqslant t, g(x) \leqslant u\}$.

— **Conditions de Slater** : si f et D sont convexes, h_i affines et g_j convexes et il existe un point strictement faisable ($\exists \bar{x} \in D^*$ tel que $\forall j,\ g_j(\bar{x}) < 0$), alors $d^* = p^*$ (dualité forte).

— **Conditions de Karush-Kühn-Tucker (KKT)** : Si il y a dualité forte, alors x^* est une variable primale optimale et (λ^*, μ^*) une paire duale optimale si et seulement si
  — *stationarité primale* : x^* minimise $x \mapsto \mathcal{L}(x, \lambda^*, \mu^*)$.
  — *faisabilité* : x^* et (λ^*, μ^*) sont faisables
  — *conditions de complémentarité* : $\forall j, \mu_j^* g_j(x^*) = 0$

— Preuve pour les conditions de KKT : soit $x^* \in D$ *faisable* (i.e., $x \in D^*$) et $(\lambda^*, \mu^*) \in \mathbb{R}^m \times \mathbb{R}_+^r$. Alors

$$\begin{aligned} q(\lambda^*, \mu^*) &= \inf_{x\in D} f(x) + (\lambda^*)^\top h(x) + (\mu^*)^\top g(x) \\ &\leqslant f(x^*) + (\lambda^*)^\top h(x^*) + (\mu^*)^\top g(x^*) \\ &\leqslant f(x^*). \end{aligned}$$

La paire (x^*, λ^*, μ^*) est alors optimale si et seulement si il y a égalité dans les deux inégalités précédentes, ce qui aboutit aux conditions de KKT.

— Remarques : (a) le dual du dual est le primal, (b) plusieurs problèmes duaux, dualité forte pas toujours vraie.

— **Exemple (Programmation linéaire)** : $\min_{Ax=b, x \geqslant 0} c^\top x = \max_{A^\top y \leqslant c} b^\top y$

— **Exemple (Problème quadratique avec contrainte d'égalité)** : $\min_{a^\top x=b} \frac{1}{2} x^\top Q x - q^\top x$

— **Exemple (Relaxation Lagrangienne de problème combinatoire - Max Cut)** : $\min_{x \in \{-1,1\}^n} x^\top W x$

— **Exemple (Dualité forte pour problème non convexe)** : $\min_{x^\top x \leqslant 1} \frac{1}{2} x^\top Q x - q^\top x$