Minimisation et fonctions convexes

E. Le Pennec

Novembre 2012

Pb Estimateurs souvent définis comme des "arg min" (M-estimateurs...) :
— Étude des propriétés (statistiques) de ces minimiseurs.
— Ici : Comment les calculer/approcher numériquement ? Optimisation !

Exemple $Y = AX_0 + \epsilon$

1. MCO : $\hat{X} = \arg\min \frac{1}{2}\|AX - Y\|^2$
2. Ridge : $\hat{X} = \arg\min \frac{1}{2}\|AX - Y\|^2 + \lambda\|X\|_2^2$
3. Sélection de modèle : $\hat{X} = \arg\min \frac{1}{2}\|AX - Y\|^2 + \lambda\|X\|_0$ avec $\|X\|_0 = \sum_{i=1}^d \mathbf{1}_{x\neq 0}$
4. Lasso : $\hat{X} = \arg\min \frac{1}{2}\|AX - Y\|^2 + \lambda\|X\|_1$

Algorithme :

1. Formule explicite si A^*A est inversible : $\hat{X} = (A^*A)^{-1}A^*X$ (Dérivée nulle)
2. Formule explicite si $\lambda > 0$: $\hat{X} = (A^*A + \lambda I)^{-1}A^*X$ (Dérivée nulle)
3. Exploration exhaustive des sous-ensembles $I \subset 1, \ldots, d$ tel que $X_I^*X_I$ est inversible, MCO sur ce support restreint et comparaison des différentes solutions. Nombre de sous-ensembles I=2^d $\implies$ inutilisable quand d est grand.
4. "Fonctionnelle convexe" $\implies$ facile à minimiser !

Fonctions convexes Une fonction f de $\mathbf{R}^d \to \mathbb{R} \cup +\infty$ est dite convexe si

$$\forall (x, y) \in (\mathbf{R}^d)^2, \forall \theta \in [0, 1] f(\theta x + (1 - \theta)y) \leq \theta f(x) + (1 - \theta)f(y).$$

Propriétés clés

1. Minimum local $\implies$ Minimum global

   $$\exists R > 0, \forall y \in B(x, R) f(x) \leq f(y) \implies \forall y \in \mathbf{R}^d f(x) \leq f(y)$$

   Recherche de minimum locaux suffisant.
2. Existence d'un sous-gradient sans hypothèse de régularité :

   $$\forall x \in \mathbf{R}^d, \exists R \text{ t.q. } \sup_{y \in B(x,R)} f(y) < +\infty \implies \exists \delta \in \mathbf{R}^d t.q. \forall y \in \mathbf{R}^d f(y) \geq f(x) + \langle \delta, y - x \rangle$$

   La sous différentielle en x $\partial f(x) = \{\delta, \forall y \in \mathbf{R}^d f(y) \geq f(x) + \langle \delta, y - x \rangle\}$ est donc non vide dès que $\exists R$ t.q. $\sup_{y \in B(x,R)} f(y) < +\infty$
3. Si $\exists R$ t.q. $\sup_{y \in B(x,R)} f(y) < +\infty$ alors x est un minimum global si et seulement si $0 \in \partial f(x)$.

1 Ensembles convexes et fonctions convexes

1.1 Ensembles convexes

E $\mathbb{R}$-espace vectoriel.

Définition $C \subset E$ est dit convexe si

$$\forall (x,y) \in C^2, \forall \theta \in [0,1] \theta x + (1-\theta) y \in C.$$

Exemples Espaces affines, segments, demi-espaces définis par des hyperplans, boules l_p pour $p \geq 1$ (inégalité triangulaire), ellipsoides

Cône K est un cône si $\forall \lambda \geq 0, x \in K \implies \lambda x \in K$.

Exemples de cônes convexes E, $\mathbb{R}^d_+$, l'ensemble des matrices semi-définies positives.

Si K est un convexe alors $\{(t,x) \in \mathbb{R} \times E, t \geq 0 et x \in tC\}$ est un cône convexe.

Le dual K^* d'un cône convexe défini par $\{x \in E, \forall y \in K, \langle x,y \rangle \geq 0\}$ est encore un cône convexe.

Stabilité Les convexes sont stables par intersections et sommes (de Minkowski).

Les polyèdres qui sont obtenus par intersection de demi-espaces définis par des hyperplans sont donc des convexes.

Résultat fondamental Théorème(s) de séparations de Hahn-Banach.

Soit C et D deux convexes tels que $\mathring{C} \cap \mathring{D} = \emptyset$

— $\exists$? un hyperplan séparant C et D ?

— $\exists$? une forme linéaire telle que

$$\forall x \in C, \forall y \in C, f(x) - f(y) \geq 0$$

Ces deux questions sont équivalentes et la réponse est obtenue à l'aide du théorème de prolongement des formes linéaires de Hahn-Banach...

La séparation est propre si il n'y a pas égalité partout, stricte si inégalité stricte et forte si on peut remplacer 0 par $\epsilon > 0$.

Th Si C est un ouvert convexe non vide et si $D = \{y\}$ avec $y \notin C$ alors il existe une forme linéaire f telle que

$$\forall x \in C, \forall y \in C, f(x) - f(y) > 0$$

Th Si C est un ouvert convexe non vide et si D est un convexe tel que $C \cap D = \emptyset$ alors il existe une forme linéaire f telle que

$$\forall x \in C, \forall y \in C, f(x) - f(y) > 0$$

Th Tout convexe fermé C est l'intersection des hyperplans qui le contiennent.

Les convexes fermés sont une généralisation des polyèdres.

1.2 Fonctions convexes

Def Une fonction f de $\mathbf{R}^d \to \mathbb{R} \cup +\infty$ est dite convexe si

$$\forall (x,y) \in (\mathbf{R}^d)^2, \forall \theta \in [0,1] f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta) f(y).$$

Soit $\mathrm{dom} f = \{x, f(x) < +\infty\}$, cette définition implique que $\mathrm{dom} f$ est convexe.

On retrouve, en posant $C = \mathrm{dom} f$, la définition plus classique :
Def Une fonction f de $C \subset \mathbf{R}^d \to \mathbb{R}$ est dite convexe si C est convexe et si

$$\forall (x,y) \in C^2, \forall \theta \in [0,1] f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta) f(y).$$

Jensen

— $\forall (x_1, \ldots, x_n) \in (\mathbb{R}^d)^n \forall (\theta_1, \ldots, \theta_n) \in [0,1]^n, \sum_{i=1}^n \theta_i = 1 \implies f\left(\sum_{i=1}^n \theta_i x_i\right) \leq \sum_{i=1}^n \theta_i f(x_i)$

— Soit X une v.a., $f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$

Epigraphe Il existe une caractérisation plus géométrique à l'aide de l'épigraphe de f,

$$\mathrm{epi}(f) = \{(x,t) \in \mathbb{R}^d \times \mathbb{R}, f(x) \leq t\}$$

Prop f est convexe si et seulement si $\mathrm{epi}(f)$ est convexe.

Exemples

— $x \mapsto \|x\|_2^2$, $x \mapsto \|x\|_p$ avec $p \geq 1$, $x \mapsto \langle a, x \rangle + b$

— Indicatrice de C : $\chi_C : x \mapsto \begin{cases} 0 & \text{si } x \in C \text{ convexe} \\ +\infty & \text{sinon} \end{cases}$

— Jauge de C t.q. $0 \in C$: $\mu_C : x \mapsto \inf\{t \geq 0, x \in tC\}$

— Si f_i est convexe $\forall i \in I$ alors

  — $\sum_{i \in I} \gamma_i f_i$ avec $\gamma_i \geq 0$ est convexe,

  — $x \mapsto \sup_{i \in I} f_i(x)$ est convexe.

Concavité On dit que f est concave si $-f$ est convexe.

Continuité Si f est convexe, f est continue sur $\overbrace{\mathrm{dom} f}^{\circ}$.

Sous-gradient (Hahn-Banach) Si f est convexe et $x \in \overbrace{\mathrm{dom} f}^{\circ}$,

$$\exists \delta \in \mathbf{R}^d, \forall y \in \mathbb{R}^d f(y) \geq f(x) + \langle \delta, y - x \rangle.$$

Sous-différentielle Si f est convexe et $x \in \overbrace{\mathrm{dom} f}^{\circ}$, la sous-différentielle de f en x

$$\partial f(x) = \{\delta \in \mathbf{R}^d, \forall y \in \mathbb{R}^d f(y) \geq f(x) + \langle \delta, y - x \rangle\}$$

est non-vide.

Condition de premier ordre Soit $x \in \overset{\circ}{\overbrace{\text{dom} f}}$, $x \in \arg\min f(y) \Leftrightarrow 0 \in \partial f(x)$.

Différentiabilité f convexe est différentiable en $x \in \overset{\circ}{\overbrace{\text{dom} f}}$ si et seulement si $\partial f(x) = \{\nabla f(x)\}$.

Monotonie Si f est convexe,

$$\forall (x,y) \in (\overset{\circ}{\overbrace{\text{dom} f}})^2, \forall \delta_x \in \partial f(x), \forall \delta_y \in \partial f(y), \langle \delta_y - \delta_x, y - x \rangle \geq 0.$$

Dans le cas $d = 1$, ceci correspond à la croissance des sous-gradients.

Transformée de Fenchel-Legendre Soit f une fonction non nécessairement convexe, $f_\star : x \mapsto \sup_y \langle x, y \rangle - f(y)$ est convexe.
Utilisation dans la méthode de Cramer-Chernoff.
Si f est convexe et semi continue inférieurement $f_{\star\star} = f$.
ement).

f et ses sous-gradients Si f convexe est s.c.i. alors epi$f$ est convexe fermé et donc l'union des demi-espaces définis par des hyperplans qui le contiennent. Ceci s'écrit pour $y \in \overset{\circ}{\overbrace{\text{dom} f}}$

$$f(y) = \sup\{f(x) + \langle \delta, y - x \rangle, x \in \overset{\circ}{\overbrace{\text{dom} f}} et \delta \in \partial f(x)\}.$$

2 Algorithmes de minimisation sans contraintes

Pb Déterminer $\min f(x)$ à ϵ près.

Méthode boîte noire Une boite noire (un oracle) fournit pour tout x :
— $f(x)$ (Ordre 0)
— $\delta \in \partial f(x)$ ou $\partial f(x)$ (Ordre 1)
— $\nabla^2 f$ (Ordre 2).

Coût d'un algorithme Nombre d'évaluations N nécessaire pour garantir une précision de ϵ.

Borne inférieure Minorant du coût minimale nécessaire pour une certaine classe de fonction quel que soit l'algorithme utilisé. (Approche minimax en statistique).

Robustesse Que se passe-t-il si les évaluations ne sont possible qu'avec une certaine précision...

2.1 Méthode de grille

Pb $\min\limits_{x \in [0,1]^d} f(x)$ avec f L-Lipschitzienne ($\forall (x,y) \in [0,1]^d, |f(x) - f(y)| \leq L\|x - y\|$).

Méthode naïve On évalue f sur une grille $\mathcal{G}$ de précision $2\epsilon/L\sqrt{d}$ ce qui garantit bien $\min_{x \in \mathcal{G}} f(x) \leq \min_{x \in [0,1]^d} f(x) + \epsilon$.

Coût $N \simeq \left(\frac{L\sqrt{d}}{2\epsilon}\right)^d$

Borne inférieure $N \geq \left(\frac{L}{2\epsilon}\right)^d$.

On ne peut pas faire (beaucoup) mieux que la méthode naïve.

Robustesse Pas de soucis.

2.2 Méthode de la bisection

Pb Déterminer $\min_{x\in[0,1]} f(x)$ avec f convexe et $\sup_{(x,y)\in[0,1]^2} f(x)-f(y) \leq V$ (ce qui implique $\forall x \in (0,1), \forall \delta \in \partial f(x) | \delta| \leq V$).

Principe Deux observations :

- $x \in]0,1[$ est un minimum si et seulement si $0 \in \partial f(x)$.
- $x > y \rightarrow \forall \delta_x \in \partial f(x), \forall \delta_y \in \partial f(y), \delta_x \geq \delta_y$

Algorithme On pose $m_0 = 0$ et $M_0 = 1$. On répète

- on pose $x_k = (m_k + M_k)/2$ et on obtient $\delta \in \partial f(x_k)$
- si
  - $\delta = 0$, on s'arrête puisque $x = (m_k + M_k)/2$ est un minimiseur de f.
  - $\delta < 0$, on sait que tous les minimiseurs sont plus grand que x_k, on pose donc $m_{k+1} = x_k$ et $M_{k+1} = M_k$.
  - $\delta > 0$, on sait que tous les minimiseurs sont plus petit que x_k, on pose donc $m_{k+1} = m_k$ et $M_{k+1} = x_k$.

A l'étape k, on peut garantir d'être à distance 2^{-k} du minimiseur, et donc à $2^{-k}V$ du minimum.

Coût $N \simeq \log_2\left(\frac{V}{\epsilon}\right)$

Borne inférieure $N \geq 1/5 \log_2(V/\epsilon)$

En dimension supérieure Généralisation au dimension supérieur en utilisant

$$\forall x \in \overbrace{\mathrm{dom} f}^{\circ}, \forall \delta \in \partial f(x), \forall y \in \mathbb{R}^d f(y) \geq f(x) + \langle \delta, y - x \rangle$$

qui permet d'éliminer un demi-espace défini par un hyperplan à chaque étape. La grande difficulté est de trouver un point intéressant où couper. La complexité de ce sous-problème est très grande et rend cette méthode inutilisable dès que $d \geq 4$.

$N \simeq d \log(V/\epsilon)$ (coïncide avec la borne inférieur à une constante multiplicative près.

Il existe un algorithme *similaire* basée sur des ellipsoïdes qui est utilisable en pratique et vérifie $N \simeq d^2 \log(V/\epsilon)$.

Absence de robustesse Si l'on me mesure le sous-gradient qu'à une certaine précision, on risque d'éliminer une partie contenant le minimum sans jamais pouvoir revenir sur cette décision...

2.3 Méthodes de descente de (sous) gradient

Pb Déterminer $\min f(x)$ avec f convexe et L-Lipchitzienne ($\forall (x,y) | f(x) - f(y)| \leq L\|x - y\|$)

Descente de gradient On a $f(y) = f(x) + \langle \nabla f(x), (y-x) \rangle + o(\|x-y\|)$. La direction de plus forte descente en x est donc $y - x \propto -\nabla f(x)$.

L'algorithme itératif suivant

$$x_{k+1} = x_k - \gamma_k \nabla f(x_k)$$

est donc naturel.

Th : $\exists C > 0$ universel tel que si $\gamma_k = \frac{D}{\sqrt{k}\|\nabla f(x_k)\|}$ avec D un majorant de $\|x_\star - x_0\|$ alors $\min_{k \leq N} f(x_k) - f(x_\star) \leq C \frac{LD}{\sqrt{N}}$.

$N \simeq \left(\frac{LD}{\epsilon}\right)^2$ qui est la vitesse optimale pour un problème en dimension quelconque (infini compris).

Descente de sous-gradient Si f est uniquement convexe, on obtient les mêmes performances en remplaçant $\nabla f(x)$ par $\delta \in \partial f(x)$.

Fonction substitut (Surrogate) Soit $\delta \in \partial f(x)$ et $\gamma > 0$, on pose

$$\phi(y) = f(x) + \langle \delta, y - x \rangle + \frac{1}{2\gamma}\|y - x\|_2^2.$$

On vérifie aisément que $x - \gamma\delta$ est l'unique minimiseur de cette fonction fortement convexe. On peut étendre ainsi la technique de descente en remplaçant la norme $\|\cdot\|_2$ et le produit scalaire associé par d'autre norme. La méthode de Newton peut ainsi s'obtenir en utilisant $\|h\|^2 = \langle \nabla^2(x)h, h \rangle$...

Le cas ∇f L-Lipschitzienne Sous cette hypothèse, si $\gamma \leq 1/L$ alors

$$\phi(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|_2^2 \geq f(y).$$

On peut donc interpréter dans ce l'algorithme de descente de gradient à pas fixe γ comme un algorithme MM (Majorization Minimization).

Th on montre que

$$\min_{k \leq N} f(x_k) - f(x_\star) \leq C \frac{DL^2}{N}$$

i.e. $N \simeq \frac{DL}{\epsilon}$.

Accélération La vitesse minimale est en $O(1/\sqrt{\epsilon})$ et il existe une technique d'accélération de la forme

$$x_{k+1} = \arg\min f(\widetilde{x}_k) + \langle \delta, y - \widetilde{x}_k \rangle + \frac{1}{2\gamma}\|y - \widetilde{x}_k\|_2$$
$$\widetilde{x}_{k+1} = x_k + \beta_{k+1}(x_{k+1} - x_k)$$

avec β_{k+1} bien choisi. Par exemple, $\beta_k = (t_k - 1)(t_{k+1})$ avec $t_0 = 1$ et $t_{k+1} = (1 + \sqrt{1 + 4t_k^2})/2$ permet d'obtenir la vitesse optimale tandis que $\beta_k = 1$ redonne la descente de gradient classique.

Le cas f strictement convexe Si $\exists \mu > 0$,

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|y - x\|_2^2$$

alors par le même algorithme, on arrive une vitesse optimale dans ce cas en $-\log \epsilon$.

Robustesse Ces méthodes sont robustes au sens où leur comportement est peu dégradé par des petites erreurs de mesures (par exemple sommable pour la descente de gradient non accélérée).

2.4 Méthodes proximales

Pb Déterminer $\min f(x)$ avec f convexe.

Opérateur proximal Soit f une fonction convexe, on définit

$$\text{prox}_{\gamma f}(y) = \arg\min_{x \in \mathbf{R}^d} \frac{1}{2\gamma}\|x - y\|^2 + f(x).$$

Le minimiseur de $x \mapsto \frac{1}{2\gamma}\|x - y\|^2 + f(x)$ est unique car cette fonction est strictement convexe.

Prox et méthode du gradient On a déjà utilisé implicitement cet opérateur avec la fonction convexe $y \mapsto f(x) + \langle \delta, y - x \rangle$ dans la méthode de descente.

Opérateur proximal et projection Si $f = \chi_C$ avec C convexe alors prox_f n'est rien d'autre que la projection orthogonale sur C.

Propriétés

- $u = \text{prox}_{\gamma f}(y) \Leftrightarrow y - u \in \gamma \partial f(u) \Leftrightarrow y \in u + \gamma \partial f(u)$
- En utilisant la monotonie des sous-gradients de f, on montre que $\text{prox}_{\gamma f}$ est un opérateur quasi-contractant :

$$\forall (x, y), \|\text{prox}_{\gamma f}(x) - \text{prox}_{\gamma f}(y)\|_2^2 \leq \|x - y\|_2^2 - \|(x - \text{prox}_{\gamma f}(x)) - (y - \text{prox}_{\gamma f}(y))\|_2^2.$$

Minimiseur de f et point fixe de $\text{prox}_{\gamma f}$ Soir $x^\star \in \mathring{\text{dom}} f$,

$$x^\star \in \arg\min_x f(x) \Leftrightarrow 0 \in \partial f(x^\star) \Leftrightarrow 0 \in \partial \gamma f(x^\star) \Leftrightarrow x^\star \in x^\star + \partial \gamma f(x^\star) \Leftrightarrow x^\star = \text{prox}_{\gamma f}(x^\star)$$

Minimisation possible par un algorithme de point fixe :

$$x_{k+1} = \text{prox}_{\gamma f}(x_k).$$

Pb Calculer un prox est en général aussi compliqué que de faire la minimisation...

Quelques prox pour des fonctions simples

- $f(x) = \chi_C(x) \Rightarrow \text{prox}_{\gamma f}(x) = P_C(x)$ avec P_C projection orthogonale.
- $f(x) = \lambda|x| \Rightarrow \text{prox}_{\gamma f}(x) = \begin{cases} 0 & \text{si } |x| \leq \gamma \\ \text{sign}(x)(|x| - \gamma) & \text{sinon} \end{cases}$ (seuillage doux)
- $f(x) = \sum_{i=1}^d f_i(x_i) \Rightarrow \text{prox}_{\gamma f}(x) = (\text{prox}_{\gamma f_1}(x_1), \ldots, \text{prox}_{\gamma f_d}(x_d))$
- mais en général $\text{prox}_{\gamma(f_1+f_2)}(x) \neq \text{prox}_{\gamma f_1}(x) + \text{prox}_{\gamma f_2}(x)$!

Le cas $f = f_1 + f_2$ avec f_1 à gradient L-Lipschitzien et γf_2 à prox connu

Algo Forward-Backward $x_{k+1} = \text{prox}_{\gamma f_2}(x_k - \gamma \nabla f_1(x))$ avec $\gamma < 1/L$

Vision MM $f(y) \leq \phi(y) = f(x) + \langle \nabla f_1(x), y - x\rangle + \frac{1}{2\gamma}\|y - x\|_2^2 + f_2(y)$ et $\arg\min \phi(y) = \arg\min \frac{1}{2\gamma}\|y - (x - \gamma\nabla f_1(x))\|_2^2 + f_2(y) = \text{prox}_{\gamma f_2}(x - \gamma \nabla f_1(x))$

Vision point fixe Si $x_\star \in \overset{\circ}{\overbrace{\text{dom} f}}$,

$$x_\star \in \arg\min f(x) \Leftrightarrow 0 \in \partial f(x_\star) \Leftrightarrow 0 \in \nabla f_1(x_\star) + \partial f_2(x_\star) \Leftrightarrow (x_\star - \gamma\nabla f_1(x_\star)) \in (x_\star + \gamma \partial f_2(x_\star)$$
$$\Leftrightarrow x_\star = \text{prox}_{\gamma f_2}(x_\star - \gamma \nabla f_1(x_\star))$$

Th $\min_{i \leq N} f(x_i) - f(x_\star \leq C\frac{DL^2}{N}$ (même vitesse que dans le cas $f_2 = 0$)

Accélération Même(s) accélération(s) que dans le cas gradient.

Le cas $f_2 = \chi_C$ correspond à $\min_{x\in C} f_1(x)$ et l'algo Forward-Backward est exactement la méthode du gradient projeté :

$$x_{k+1} = P_C\left(x_k - \gamma \nabla f_1(x_k)\right).$$

Autres cas Les cas $f = f_1 + f_2$ avec γf_1 et γf_2 à prox connu peut être traité par un algorithme (Douglas-Rachford) reposant sur l'itération

$$z_{k+1} = (1 - \alpha/2)z_k + \alpha/2\left(2\text{prox}_{\gamma f_2}\left(2\text{prox}_{\gamma f_1}(z_k) - z_k\right) - 2\text{prox}_{\gamma f_1}(z_k) - z_k)\right)$$
$$x_{k+1} = \text{prox}_{\gamma f_2}(z_{k+1})$$

tandis que le cas $f = f_0 + \sum_{i=1}^D f_i$ avec f_0 à gradient L-Lipschitz et γf_i connu peut être associé à l'itération

$$z_{k+1}^{(i)} = z_k^{(i)} + \text{prox}_{D\gamma f_i}\left(2x_k - z_k^{(i)} - \gamma \nabla f_0(x_k)\right) - x_k$$
$$x_{k+1} = \frac{1}{D}\sum_{i=1}^D z_{k+1}^{(i)}.$$

3 Dualité, Lagrangien et K.K.T.

3.1 Dualité

Pb (primal) Déterminer $p_\star = \inf f(x)$ t.q. $\begin{cases} g(x) \succeq_K b \\ x \in X \end{cases}$ où $a \succeq b \Leftrightarrow a - b \in K$ (K cône convexe donné)

Ex $K = \mathbb{R}_+^d$: $a \succeq_K b \simeq \forall i \in \{1, \ldots, d\} a_i \geq b_i$

Fonctions croissantes On définit l'ensemble $\mathcal{F}$ des fonctions K-croissantes par

$$\widetilde{\mathcal{F}} = \{F : \mathbb{R}^d \to \mathbb{R}, a \succeq_K b \Rightarrow F(a) \geq F(b)\}.$$

Pb dual Déterminer $d_{\star\star} = \sup F(b)$ t.q. $\begin{cases} \forall x \in X, F(g(x)) \leq f(x) \\ F \in \widetilde{\mathcal{F}} \end{cases}$

Dualité faible $p_\star \geq d_{\star\star}$
Preuve : $g(x) \succeq_K b \Rightarrow F(b) \leq F(g(x)) \leq F(x)$.

Dualité forte $p_\star = d_{\star\star}$

Preuve : On définit la fonction de sensibilité $\phi(a) = \inf f(x)$ t.q. $\begin{cases} g(x) \succeq_K a \\ x \in X \end{cases}$. Par construction $\phi(b) = p_\star$, $\phi \in \widetilde{\mathcal{F}}$ et $\forall x \in X, \phi(g(x)) \leq f(x)$.

Saut de dualité Estimer p et d et utiliser $p - d$ comme borne d'erreur entre $p - p_\star$ et $d_{\star\star} - d$.

Pb $\widetilde{\mathcal{F}}$ est une classe trop riche pour pouvoir étudier le problème dual.

Pb dual restreint Soit $\mathcal{F} \subset \widetilde{\mathcal{F}}$, déterminer $d_\star = \sup F(b)$ t.q. $\begin{cases} \forall x \in X, F(g(x)) \leq f(x) \\ F \in \mathcal{F} \end{cases}$

Dualité et transformée de Fenchel-Legendre $F \in \mathcal{F}$ est faisable $\Leftrightarrow \sup_{x \in X} F(g(x)) - f(x) \leq 0$. On vérifie alors que

$$\sup_{x \in X} F(g(x)) - f(x) = \sup_b \sup_{x \in X, g(x) \succeq b} F(b) - f(x) = \sup_b F(b) - \phi(b) = \phi^*(F)$$

où pour toute fonction g $g^*(F) = \sup_b F(b) - g(b)$ qui correspond bien à la transformée de Fenchel-Legendre classique lorsque $F(b) = \langle u, b \rangle$.

Dualité On conserve la dualité faible $p_\star \geq d_\star$ mais il n'y a aucune raison que la dualité forte soit conservée...

Complémentarité Si $\exists \widetilde{x} \in X, g(\widetilde{x}) \succeq_K b$ et $\exists \widetilde{F} \in \mathcal{F}, \forall x \in X, \widetilde{F}(g(x)) \leq f(x)$ tels que $f(\widetilde{x}) = \widetilde{F}(b)$ alors

$$p_\star = f(\widetilde{x}) = \widetilde{F}(g(\widetilde{x})) = \widetilde{F}(b) = d_\star$$

3.2 Lagrangien

Def On définit le Lagrangien de ce problème par

$$X \times \mathcal{F} \to \mathbb{R} : x \mapsto \mathcal{L}(x, f) = f(x) + F(b) - F(g(x))$$

Propriété minmax Si $\forall u - v \notin K, \exists F \in \mathcal{F}, F(u) < F(v)$ et si $\forall \alpha \geq 0, \alpha\mathcal{F} \subset \mathcal{F}$

$$\forall x \in X, \sup_{F \in \mathcal{F}} \mathcal{L}(x, F) = \begin{cases} f(x) & \text{si } g(x) \succeq b \\ +\infty & \text{sinon} \end{cases}$$

et donc

$$\inf_{x \in X} \sup_{F \in \mathcal{F}} \mathcal{L}(x, F) = p_\star.$$

Propriété maxmin Si $\forall \alpha \in \mathbb{R}, \mathcal{F} + \alpha \in \mathcal{F}$,

$$\inf_{x\in X} \mathcal{L}(x,F) = \begin{cases} -\infty & \text{si } \phi^*(F) = \sup_{x\in X} F(g(x)) - f(x) = \infty \\ \inf_{x\in X} \mathcal{L}(x, F - \phi^*(F)) & \text{avec } \forall x \in X, F(g(x)) - \phi^*(F) \leq f(x) \text{ sinon} \end{cases}$$

et donc

$$\sup_{F\in\mathcal{F}} \inf_{x\in X} \mathcal{L}(x,F) = d_\star.$$

Dualité faible On obtient sous les hypothèses sur $\mathcal{F}$ des propriétés minmax et maxmin

$$p_\star = \inf_{x\in X} \sup_{F\in\mathcal{F}} \mathcal{L}(x,F) \geq \sup_{F\in\mathcal{F}} \inf_{x\in X} \mathcal{L}(x,F) = d_\star.$$

Dualité forte Toujours sous les mêmes hypothèses sur $\mathcal{F}$, on a dualité forte si et seulement si

$$\inf_{x\in X} \sup_{F\in\mathcal{F}} \mathcal{L}(x,F) = \sup_{F\in\mathcal{F}} \inf_{x\in X} \mathcal{L}(x,F)$$

Dans ce cas, si les limites sont atteintes, $\mathcal{L}(x_\star, F_\star) = \inf_{x\in X} \mathcal{L}(x, F_\star)$.

Théorie max-min s'occupe d'obtenir de tels résultats.

Lagrangien usuel L'ensemble des fonctions K affine $\mathcal{F} = \{x \mapsto \langle u, x\rangle + u_0, u \in K^* \text{ et } u_0 \in \mathbb{R}\}$ satisfait les hypothèses des propriétés minmax et maxmin. Les termes en u_0 se simplifiant, on écrit le Lagrangien sous la forme

$$\mathcal{L}(x,u) = f(x) + \langle u, b - g(x)\rangle \quad \text{avec } u \in K^*.$$

Complémentarité Si $\exists \widetilde{x} \in X, g(\widetilde{x}) \succeq_K b$ et $\exists \widetilde{u} \in K^*, \exists \widetilde{u}_0 \in \mathbb{R}, \forall x \in X, \langle \widetilde{u}, g(x)\rangle + \widetilde{u}_0 \leq f(x)$ tels que $f(\widetilde{x}) = \langle \widetilde{u}, b\rangle + \widetilde{u}_0$

$$p_\star = f(\widetilde{x}) = \langle \widetilde{u}, g(\widetilde{x})\rangle + \widetilde{u}_0 = \langle \widetilde{u}, b\rangle + \widetilde{u}_0 = b_\star.$$

On en déduit $\langle \widetilde{u}, b - g(x)\rangle = 0$ et

$$\mathcal{L}(\widetilde{x}, \widetilde{u}) = \min_{x\in X} \mathcal{L}(x, \widetilde{u}) = \min_{x\in X} f(x) + \langle \widetilde{u}, b - g(x)\rangle.$$

3.3 Le cas convexe

Cadre On prend les fonctions K-affines $\mathcal{F} = \{x \mapsto \langle u, x\rangle + u_0, u \in K^* et u_0 \in \mathbb{R}\}$ et on suppose f convexe, X convexe et g K-concave (i.e. $-\langle u, g\rangle$ convexe $\forall u \in K^*$).

Lagrangien et minimisation On pose

$$\mathcal{L}(x,u) = f(x) + \langle u, b - g(x)\rangle \quad \text{avec } u \in K^*$$

et on vérifie que

— $u \mapsto \mathcal{L}(x,u)$ est trivial à maximiser en u ($f(x)$ ou $+\infty$)

— $x \mapsto \mathcal{L}(x,u)$ est convexe et donc *facile* à minimiser.

Condition de Slater Si $\exists x \in X$ t.q. $g(x) \succ b$ alors on a la propriété de dualité forte.

Théorème de K.K.T. (Karush-Kuhn-Tucker) Si on a la propriété de dualité forte alors $\widetilde{x}$ est un minimiseur de $f(x)$ sous $g(x) \succeq b$ et $x \in X$ si et seulement si $\exists \widetilde{u} \in K^*$ t.q.

— $g(\widetilde{x}) \succeq b$, $\widetilde{x} \in X$
— $\langle \widetilde{u}, b - g(\widetilde{x}) \rangle = 0$
— $0 \in \partial f(\widetilde{x}) + \partial(-\langle \widetilde{u}, g \rangle)(\widetilde{x})$

Si l'on admet le fait que $\widetilde{u}$ correspond à un maximiseur du problème dual. La première propriété correspond à la faisabilité de $\widetilde{x}$, la seconde à la propriété de complémentarité tandis que la troisième n'est alors rien d'autre que la condition d'optimalité de premier ordre de

$$\mathcal{L}(\widetilde{x}, \widetilde{u}) = \min_{x \in X} \mathcal{L}(x, \widetilde{u}) = \min_{x \in X} f(x) + \langle \widetilde{u}, b - g(x) \rangle.$$

En fait, par complémentarité, il suffit de montrer que $\widetilde{u}$ correspond bien à un point faisable du problème dual. On pose $F(a) = f(\widetilde{x}) + \langle \widetilde{u}, a - b \rangle$. Par construction $F(b) = f(\widetilde{x})$. On vérifie alors que F est faisable puisque

$$\begin{aligned}
F(g(x)) &= f(\widetilde{x}) + \langle \widetilde{u}, g(x) - b \rangle = f(\widetilde{x}) + \langle \widetilde{u}, g(x) - g(\widetilde{x}) \rangle \quad \text{(par complémentarité)} \\
&= f(\widetilde{x}) - ((-\langle \widetilde{u}, g)(x) - (-\langle \widetilde{u}, g)(\widetilde{x})) \\
&\leq f(\widetilde{x}) - \langle \delta, x - \widetilde{x} \rangle \quad \forall \delta \in \partial(-\langle \widetilde{u}, g \rangle)(\widetilde{x}) \text{ par } K\text{-concavité de } g \\
&\leq f(\widetilde{x}) + \langle \delta, x - \widetilde{x} \rangle \quad \text{avec } \delta \in \partial f(\widetilde{x}) \text{ par la condition d'optimalité de premier ordre} \\
&\leq f(x) \quad \text{par convexité de } f
\end{aligned}$$

Cas $K = \mathbb{R}^n_+$ alors $K^* = K$, la K-concavité de g se traduit par la K-concavité des g_i et les conditions K.K.T. se réécrivent $\exists \widetilde{u} \in \mathbb{R}^n_+$ t.q.

— $\forall i \in \{1, \dots, d\} g_i(\widetilde{x}) \geq b_i$, $\widetilde{x} \in X$
— $\langle \widetilde{u}, b - g(\widetilde{x}) \rangle = \sum_{i=1}^n \widetilde{u}_i (b_i - g(\widetilde{x})) = 0 \Leftrightarrow \forall i \in \{1, \dots, n\} \widetilde{u}_i (b_i - g(\widetilde{x})) = 0$
— $0 \in \partial f(\widetilde{x}) - \sum_{i=1}^n \widetilde{u}_i \partial(-g_i)(\widetilde{x})$

4 Algorithmes de minimisation avec contraintes

4.1 Méthodes primales/duales

Il s'agit d'une famille de méthodes où l'on utilise le problème dual pour résoudre le problème primal à travers la formulation Lagrangienne lorsqu'on a la propriété de dualité forte.

L'heuristique de cette méthode dans le cadre du Lagrangien usuel est la suivante.

1. Soit $\widetilde{u}_k \in K^*$, on peut minimiser sur X

   $$\mathcal{L}(x, \widetilde{u}_k) = f(x) + \langle \widetilde{u}_k, b - g(x) \rangle$$

   qui est une fonction convexe. Supposons ce minimum fini, on note x_k un minimiseur. Ceci est équivalent au fait que $\widetilde{u}_k$ correspond à la fonction affine F_k faisable $x \mapsto \langle \widetilde{u}_k, x \rangle + \delta_k$ avec

   $$\delta_k = \min_{x \in X} f(x) - \langle \widetilde{u}_k, g(x) \rangle = f(x_k) - \langle \widetilde{u}_k, g(x_k) \rangle.$$

2. Par construction $F_k(b) = f(x_k) + \langle \widetilde{u}_k, b - g(x_k) \rangle$. On vérifie alors la faisabilité de x_k.
   — Si x_k est faisable alors $F_k(b) \geq f(x_k)$ et par complémentarité, x_k est une solution du problème primal et F_k du problème dual.

— Sinon par dualité forte, c'est que $F_k(b) < d_\star$. Puisque $F_k(b) = f(x_k) + \langle \widetilde{u}_k, b - g(x_k)\rangle$, on va chercher à augmenter cette valeur en modifiant $\widetilde{u}_k$ en $\widetilde{u}_{k+1}$. Il faut assurer que $\widetilde{u}_{k+1}$ est encore associé à une fonction faisable. Une approche classique est de modifier $\widetilde{u}_k$ dans la direction opposé à celle du gradient $b - g(x_k)$ avec un pas suffisamment petit.

Les méthodes primales/duales sont des variations autour de ce principe.

4.2 Méthodes de points intérieurs

Les algorithmes précédents permettent d'optimiser des fonctions convexes mais ne garantissent pas que les itérés soient faisable. Ceci peut poser un problème pratique. Pour éviter ceci, une idée simple est de *renforcer* les contraintes $g(x) \succeq_K b \Leftrightarrow g(x) - b \in K$ par $C_K(g(x) - b) < \frac{1}{\mu} < +\infty$ avec C_K une fonction qui explose sur les bords de K. On remplace alors le problème initial

$$\inf_{x \in X} f(x) \text{ t.q. } g(x) \succeq_K b$$

en une famille de problème

$$\inf_{x \in X} f(x) + \frac{1}{\mu} C_K\left((g(x) - b)\right).$$

Il « suffit » alors de résoudre ce problème en faisant tendre μ vers 0 et en utilisant à chaque étape la solution précédente comme initialisation.

Les méthodes de type points intérieurs sont des méthodes de ce type dans lequel le choix de C_K (des fonctions barrières auto-concordantes), de la méthode de *résolution* de ce problème à chaque étape (Méthode de Newton) et de modification de μ ($\mu_{k+1} = \rho\mu_k$ avec $\rho < 1$) est justifiée théoriquement lorsque $f(x)$ est affine. Sa complexité est alors $N \propto \log \frac{1}{\epsilon}$. Si f n'est pas affine, le problème peut être déporté sur la construction de la fonction C_K en modifiant le problème en

$$\inf_{(t,X) \in \mathbb{R} \times X} t \text{ t.q. } (t, x) \in \mathrm{epi} f \text{ et } g(x) \succeq_K b.$$

Les techniques les plus efficace de points intérieur combinent ce principe avec l'approche primale/duale du paragraphe précédent en remplaçant $C_X(g(x) - b)$ par $-ln(\langle \tilde{u}, g(x) - b\rangle)$ avec $\tilde{u}$ (bien) choisi de sorte à garantir que le minimiseur reste faisable.

Références

[1] A. Beck et D. Teboulle. "Gradient-based algorithms with applications to signal recovery". In : *Convex Optimization in Signal Processing and Communications*. Sous la dir. d'Y. Eldar et D. Palomar. 2009.

[2] S. Boyd et L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[3] Y. Nesterov. *Introductory lectures on convex optimization: a basic course*. Applied Optimization Series 87. Kluwer Academic Publishers, 2003.

[4] Y. Nesterov et i A. Nemirovski. *Interior Point Polynomial Algorithms in Convex Programming*. SIAM, 1987.

[5] J. Tind et L. Wolsey. "An elementary survey of general duality theory in Mathematical Programming". In : *Mathematical Programming* 21 (1981), p. 241–261.