Optimisation et apprentissage statistique

Une introduction à l'optimisation

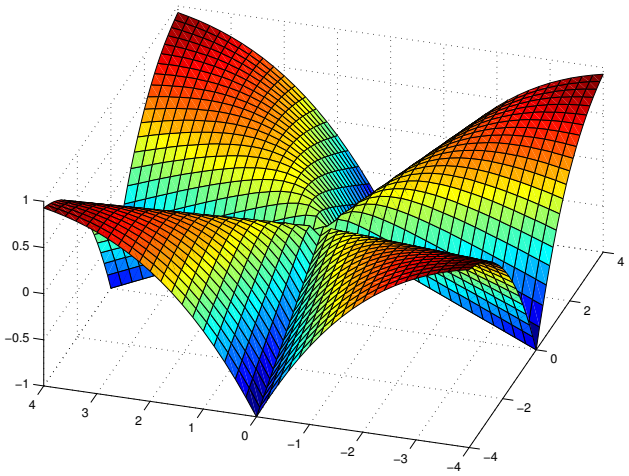
Stéphane Canu
asi.insa-rouen.fr/enseignants/~scanu

JES 2016, Fréjus

October 4, 2016

Plan

1. Exemples de problèmes d'optimisation en statistique
2. Optimisation différentiable, convexe et sans contraintes
3. Optimisation différentiable, convexe et avec contraintes
4. Optimisation sans contraintes non différentiable



Exemples de problèmes d'optimisation en statistique

Le LASSO, pour X, Y et un $\lambda > 0$ donné

$$\min_{\mathbf{w} \in \mathbb{R}^p} \quad \underbrace{\|\mathbf{X}\mathbf{w} - Y\|^2 + \lambda \|\mathbf{w}\|_1}_{J(\mathrm{w})} \tag{1}$$

Minimisation du risque empirique MRE pénalisé, pour $\lambda > 0$ donné

$$\min_{\mathbf{w} \in \mathbb{R}^d} \quad \underbrace{L(\mathbf{w}) + \lambda\, \Omega(\mathbf{w})}_{J(\mathrm{w})} \tag{2}$$

L est le risque empirique (un terme d'attache aux données)
Ω un terme de pénalisation (un a priori)

Optimisation sans contrainte

$$\min_{\mathbf{w} \in \mathbb{R}^d} \; J(\mathbf{w}) \qquad J : \begin{array}{ccc} \mathbb{R}^d & \longrightarrow & \mathbb{R} \\ \mathbf{w} & \longmapsto & J(\mathbf{w}) \end{array}$$

Avec contraintes de positivité (contraintes inégalité)

$$\begin{cases} \min\limits_{\mathbf{w}\in\mathbb{R}^d} & L(\mathbf{w}) + \lambda\,\Omega(\mathbf{w}) \\ \text{avec} & \underbrace{\mathbf{0} \leq \mathbf{w}}_{g_1(\mathbf{w})=-\mathbf{w}} \end{cases} . \tag{3}$$

pour deux paramètres $\lambda_l > 0$ et $\lambda_c > 0$ donnés (contraintes d'égalité)

$$\begin{cases} \min\limits_{\mathbf{w},\mathbf{w}^l,\mathbf{w}^c\in\mathbb{R}^d} & \|\mathbf{X}\mathbf{w} - Y\|^2 + \lambda_l\|\mathbf{w}^l\|_1 + \lambda_c\|\mathbf{F}\mathbf{w}^c\|_1 \\ \text{avec} & \underbrace{\mathbf{w} = \mathbf{w}^l + \mathbf{w}^c}_{h_j(\mathrm{w})=\mathbf{w}-\mathbf{w}^l-\mathbf{w}^c} \end{cases} \tag{4}$$

Optimisation (avec contraintes)

$$\mathcal{P} = \begin{cases} \min\limits_{\mathbf{w}\in\mathbb{R}^d} & J(\mathbf{w}) \\ \text{avec} & h_j(\mathbf{w}) = 0 \quad \forall\, j = 1,\dots,\ell \\ \text{et} & g_i(\mathbf{w}) \leq 0 \quad \forall\, i = 1,\dots,q \end{cases}$$

$$h_j, g_i : \begin{array}{ccc} \mathbb{R}^d & \longrightarrow & \mathbb{R} \\ \mathbf{w} & \longmapsto & J(\mathbf{w}) \end{array}$$

Exemples de problèmes d'optimisation en statistique

SVM, pour un $\lambda > 0$ donné

$$\begin{cases} \min\limits_{\mathbf{w}\in\mathbb{R}^d} & J(\mathbf{w}) & \\ \text{avec} & h_j(\mathbf{w}) = 0 & \forall\, j = 1, \ldots, \ell \\ \text{et} & g_i(\mathbf{w}) \leq 0 & \forall\, i = 1, \ldots, q \end{cases}$$

$$\begin{cases} \min\limits_{\mathbf{w}\in\mathbb{R}^p, b\in\mathbb{R}, \xi\in\mathbb{R}^n} & \frac{1}{2}\|\mathbf{w}\|^2 + \lambda \sum\limits_{i=1}^{n} \xi_i & \\ \text{avec} & Y_i(X_i\mathbf{w} + b) \geq 1 - \xi_i & i = 1, n \\ \text{et} & 0 \leq \xi_i & i = 1, n. \end{cases} \tag{5}$$

$$\begin{aligned} g_i(\mathrm{w}) &= 1 - \xi_i - Y_i(X_i\mathbf{w} + b) && i = 1, n \\ g_i(\mathrm{w}) &= -\mathbf{w} && i = n+1, 2n \end{aligned}$$

Le problème d'approximation de faible rang

$$\begin{cases} \min\limits_{\mathbf{W}\in\mathbb{R}^{n\times p}} & \|\mathbf{X} - \mathbf{W}\|_F^2 \\ \text{avec} & \text{rang}(\mathbf{W}) = k \end{cases} \tag{6}$$

Les principales questions liées à l'optimisation

$$\mathcal{P} = \begin{cases} \mathbf{w}^\star = \underset{\mathbf{w}\in\mathbb{R}^d}{\arg\min} & J(\mathbf{w}) \\ \text{avec} & h_j(\mathbf{w}) = 0 \quad \forall\, j = 1, \ldots, \ell \\ \text{et} & g_i(\mathbf{w}) \leq 0 \quad \forall\, i = 1, \ldots, q. \end{cases} \tag{7}$$

Le programme : [HU02]

- existence et unicité de la solution,
- conditions d'optimalité (caractérisation de $\mathbf{w}^\star$),
- calcul de $\mathbf{w}^\star$ (les aspects algorithmiques et informatiques),
- analyse et reformulation du problème.

Deux propriétés fondamentales

$\rightarrow$ convexité
$\rightarrow$ différentiabilité

Retour sur le LASSO

$$\min_{w \in \mathbb{R}^p} \frac{1}{2}\|Xw - y\|^2 + \lambda \sum_{j=1}^{p} |w_j|$$

CVX pour matlab cvxr.com/cvx/, CVX pour python cvxopt.org/

```
cvx_begin
variables w(p)
  minimize(  norm(Y - X*w,2)  + lambda *  sum(abs(w)) )
cvx_end
```

$$\begin{cases} \min\limits_{\mathbf{w} \in \mathbb{R}^p} & \frac{1}{2}\|Xw - y\|^2 + \lambda \sum\limits_{j=1}^{p} |w_j| \\ \text{avec} & \mathbf{0} \leq \mathbf{w} \end{cases}$$

```
cvx_begin
variables w(p)
  minimize(  norm(Y - X*w,2)  + lambda *  sum(abs(w)) )
  subject to
    w >= 0;
cvx_end
```

CVX : le couteau Suisse

Quelle solution utiliser ?

CVX	les solveurs du marché (Cplex, GLPK...)	code spécifique

Retour sur le LASSO avec CVX

$$\begin{cases} \min\limits_{\mathbf{w}\in\mathbb{R}^p} & \frac{1}{2}\|X\mathbf{w}-y\|^2+\lambda\sum\limits_{j=1}^{p}|w_j| \\ \text{avec} & \mathbf{0}\leq\mathbf{w} \\ \text{et} & \|X\mathbf{w}-y\|\geq k \end{cases}$$

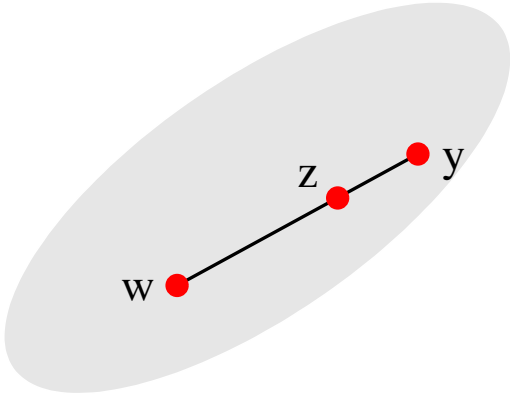
```
cvx_begin
variables w(p)
minimize(  norm(Y - X*w,2)  + lambda *  sum(abs(w)) )
subject to
  w >= 0;
  norm(Y-X*w) >= k;
cvx_end

Error using cvxprob/newcnstr (line 192)
Disciplined convex programming error:
   Invalid constraint: {convex} >= {real constant}
...
   Cannot minimize a(n) concave expression.
```

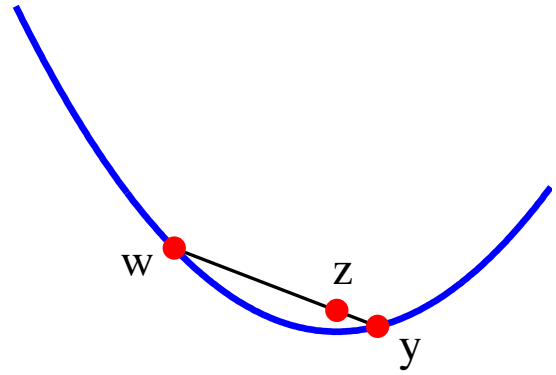
L'importance de la convexité

Convexité

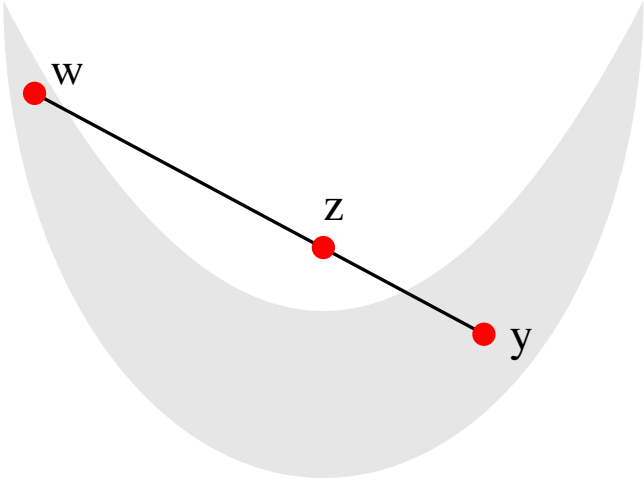
Exemple d'ensemble convexe



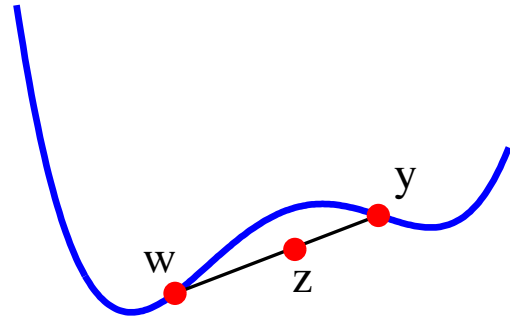
Exemple de fonction convexe



Exemple d'ensemble non convexe



Exemple de fonction non convexe



Une hiérarchie des problèmes d'optimisation convexes

programme linéaires (LP) programmes quadratiques (QP)

$$\text{(LP)} \quad \begin{cases} \min\limits_{\mathbf{w}\in\mathbb{R}^d} & \mathbf{c}^\top \mathbf{w} \\ \text{avec} & \mathbf{A}\mathbf{w} \leq \mathbf{b} \end{cases} \qquad \text{(QP)} \quad \begin{cases} \min\limits_{\mathbf{w}\in\mathbb{R}^d} & \frac{1}{2}\mathbf{w}^\top \mathbf{G}\mathbf{w} + \mathbf{c}^\top \mathbf{w} \\ \text{avec} & \mathbf{A}\mathbf{w} \leq \mathbf{b} \end{cases}$$

convexe (c) lorsque $\mathbf{G}$ définie positive.

La classe CVX

$\mathcal{P}$ est convexe si J et les g_i sont convexes et les contraintes d'égalités sont affines $h_j(\mathbf{w}) = a_j^\top \mathbf{w} + b_j = 0$

Disciplined Convex Programming hiérarchie

$$\text{LP} \subset \text{(c)QP} \subset \text{(c)QCQP} \subset \text{SOCP} \subset \text{SDP} \subset \text{CVX}$$

Lorsqu'un problème est exprimé dans ce cadre, il peut être résolu avec la boite à outil CVX [GB14].

Différentiabilité : le gradient

Supposons que les dérivées partielles $\frac{\partial J}{\partial x_i}$ existent.

Definition (Gradient)

Le gradient $\nabla J(\mathbf{w})$ d'une fonction J au point $\mathbf{w}$ est le vecteur dont les composantes sont les dérivées partielles de J.

Example

(Moindres carrées) Soit $\mathbf{X}$ une matrice $p \times n$ et Y un vecteur de réponses de talle n. Le gradient de la fonction coût des moindres carrés $J_1(\mathbf{w}) = \frac{1}{2}\|\mathbf{X}\mathbf{w} - Y\|^2$ est

$$\nabla J_1(\mathbf{w}) = \mathbf{X}^\top(\mathbf{X}\mathbf{w} - Y).$$

Règle de Fermat

Theorem

Si J est convexe et différentiable, alors

$$J(\mathbf{w}+\mathbf{h}) \geq J(\mathbf{w}) + \boldsymbol{\nabla} J(\mathbf{w})^\top \mathbf{h}, \qquad \forall\, \mathbf{h} \in \mathbb{R}^d.$$

conditions du premier ordre (règle de Fermat) :

$$\mathbf{w}^\star = \underset{\mathbf{w}\in\mathbb{R}^d}{\arg\min} \quad J(\mathbf{w}) \qquad \Leftrightarrow \qquad \boldsymbol{\nabla} J(\mathbf{w}^\star) = 0.$$

Un autre avantage du gradient est l'existence de règles générales facilitant son calcul (comme la règle d'additivité et la règle de chainage).

Le cas non différentiable

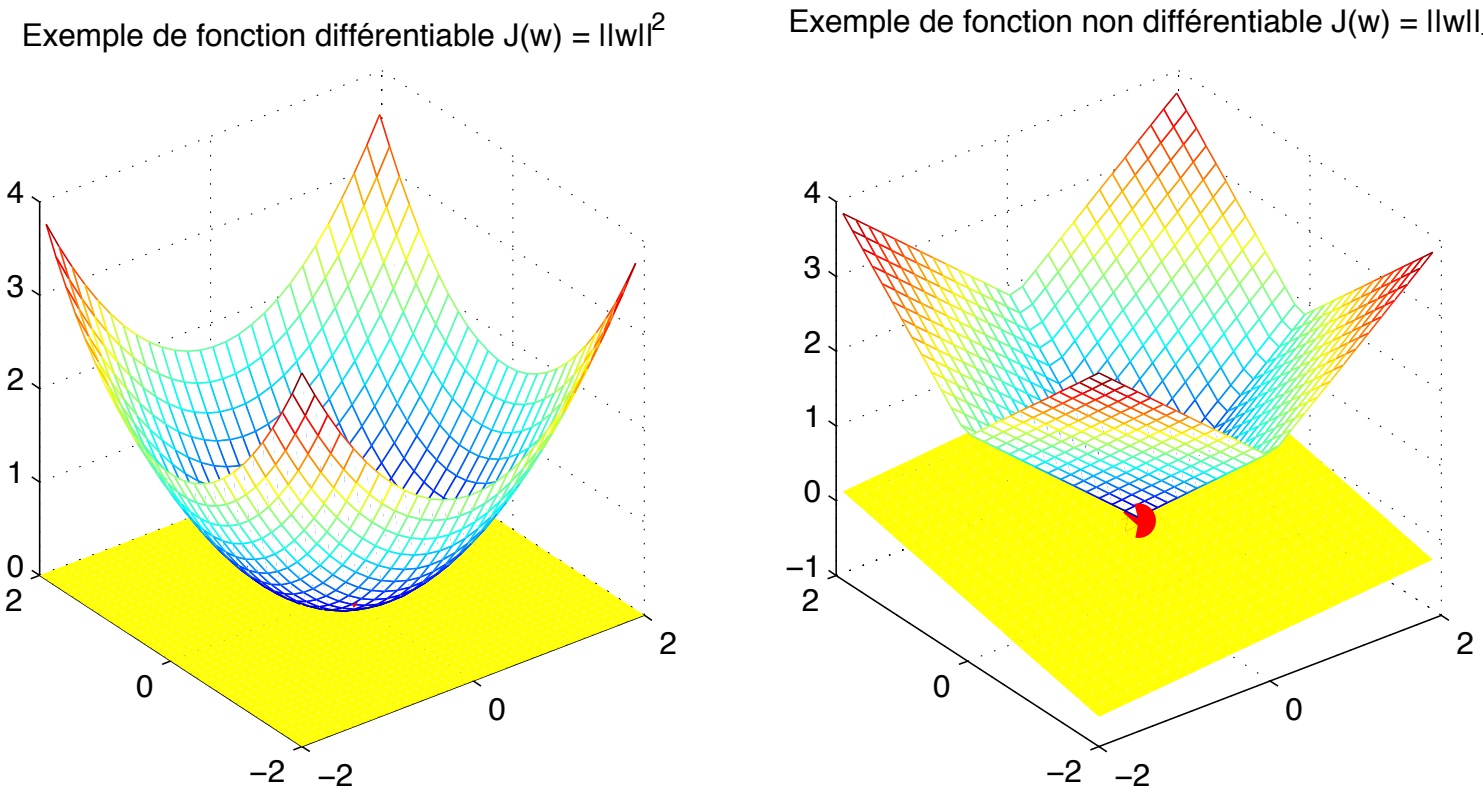


Figure : Exemple de sous gradient (en jaune) de fonctions différentiable (à gauche) et non-différentiable (à droite).

Definition (Sous gradients)

Un vecteur $\mathbf{d} \in \mathbb{R}^d$ est un sous gradient de la fonction J au point $\mathbf{w}$ si

$$J(\mathbf{w}+\mathbf{h}) \geq J(\mathbf{w}) + \mathbf{d}^\top \mathbf{h}, \qquad \forall \mathbf{h} \in \mathbb{R}^d.$$

Definition (Sous différentielle)

La sous différentielle $\partial J(\mathbf{w})$ de la fonction J au point $\mathbf{w}$ est l'ensemble (éventuellement vide) de tous ses sous gradients en ce point

$$\partial J(\mathbf{w}) = \left\{ \mathbf{g} \in \mathbb{R}^d \mid J(\mathbf{w}+\mathbf{d}) \geq J(\mathbf{w}) + \mathbf{g}^\top \mathbf{d}, \quad \forall \mathbf{d} \in \mathbb{R}^d \right\}.$$

Example

Supposons $d = 1$, nous avons :

$J_2(w) = |w|$ $\qquad$ $\partial J_2(0) = \{g \in \mathbb{R} \mid -1 \leq g \leq 1\}$,

$J_3(w) = \max(0, 1-w)$ $\qquad$ $\partial J_3(1) = \{g \in \mathbb{R} \mid -1 \leq g \leq 0\}$.

la règle de Fermat

$$\mathbf{w}^\star = \arg\min_{\mathbf{w}\in\mathbb{R}^d} J(\mathbf{w}) \quad \Leftrightarrow \quad \mathbf{0} \in \partial J(\mathbf{w}^\star). \tag{8}$$

Table : Règles de Fermat

	differentiable	non differentiable
cvx	$\mathbf{w}^\star = \arg\min_{\mathbf{w}} J(\mathbf{w}) \Leftrightarrow \nabla J(\mathbf{w}^\star) = 0$	$\mathbf{w}^\star = \arg\min_{\mathbf{w}} J(\mathbf{w}) \Leftrightarrow 0 \in \partial J(\mathbf{w}^\star)$
non cvx	$\mathbf{w}^\star = \arg\min_{\mathbf{w}\in V(\mathbf{w_0})} J(\mathbf{w}) \Rightarrow \nabla J(\mathbf{w}^\star) = 0$	$\mathbf{w}^\star = \arg\min_{\mathbf{w}\in V(\mathbf{w_0})} J(\mathbf{w}) \Rightarrow 0 \in \partial_c J(\mathbf{w}^\star)$

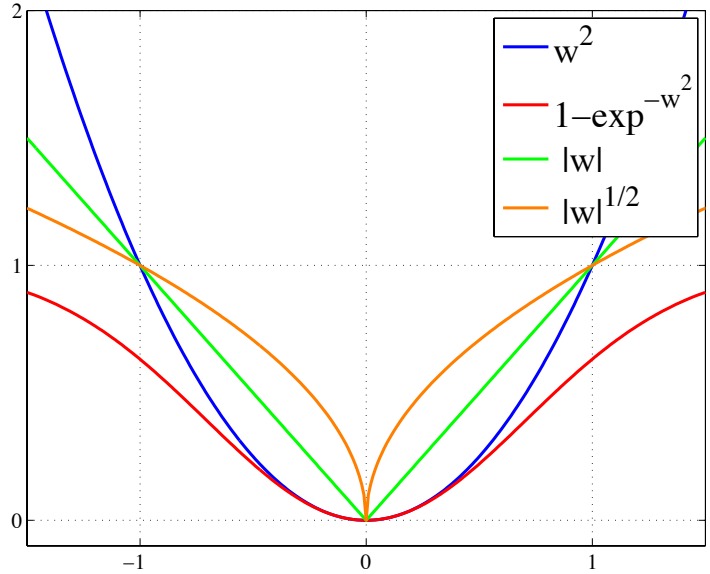
Optimum global vs. Optimum local

Example (Les pénalités décomposables)

$$\min_{\mathbf{w}\in\mathbb{R}^d} \quad L(\mathbf{w}) + \lambda\,\Omega(\mathbf{w}),$$

$$\Omega(\mathbf{w}) = \sum_{i=1}^{d} \omega(w_i). \tag{9}$$

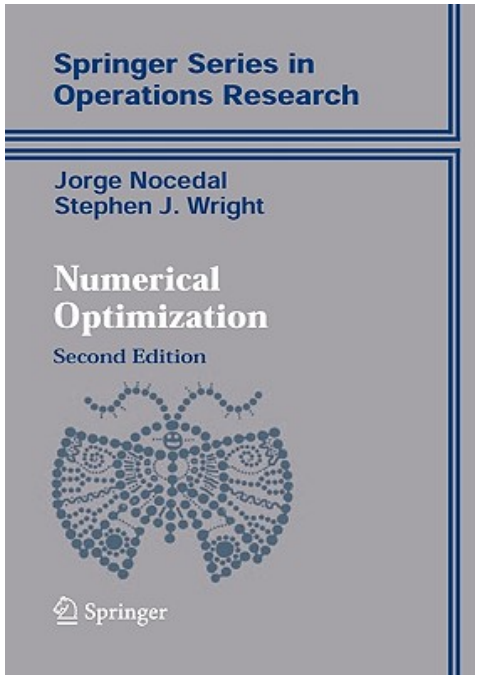
Pénalité	Convexe	Diff.
$\omega(w) = w^2$ $\Omega(\mathbf{w}) = \|\mathbf{w}\|^2$	✓	✓
$\omega(w) = 1 - \exp^{-w^2}$	✗	✓
$\omega(w) = \|w\|$ $\Omega(\mathbf{w}) = \|\mathbf{w}\|_1$	✓	✗
$\omega(w) = \sqrt{\|w\|}$ $\Omega(\mathbf{w}) = \|\mathbf{w}\|_{1/2}$	✗	✗

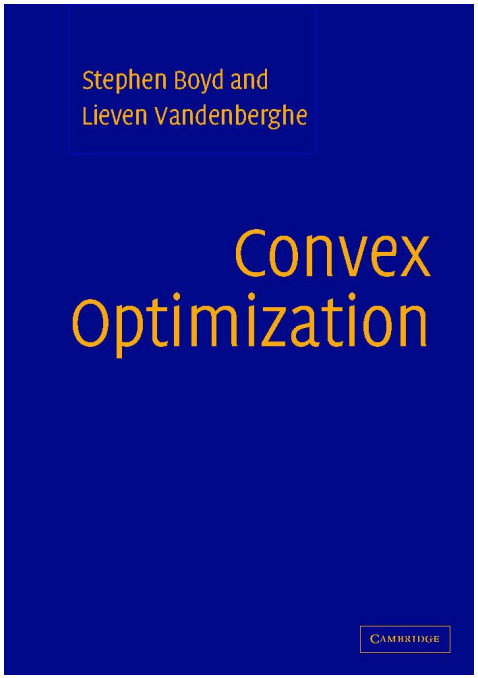


Non convexes mais parfois quasi convexes

Plan

1. Exemples de problèmes d'optimisation en statistique
2. Optimisation différentiable, convexe et sans contraintes
3. Optimisation différentiable, convexe et avec contraintes
4. Optimisation sans contraintes non différentiable





Optimisation différentiable sans contraintes

$$\min_{\mathbf{w}\in\mathbb{R}^d} J(\mathbf{w}), \tag{10}$$

avec J différentiable et convexe.

Definition (Méthode de descente de gradient)

Un algorithme de descente de gradient calcule la suite suivante :

$$\mathbf{w}^{(k+1)} = \mathbf{w}^{(k)} + \rho^{(k)}\mathbf{d}^{(k)}, \tag{11}$$

- $\mathbf{d}^{(k)} \in \mathbb{R}^d$ direction de descente $\boldsymbol{\nabla} J(\mathbf{w}^{(k)})^\top \mathbf{d}^{(k)} < 0$
- Un choix naturel $\mathbf{d}^{(k)} = -\boldsymbol{\nabla} J(\mathbf{w}^{(k)})$.
- $\rho^{(k)} \in \mathbb{R}^+$ un pas associé.
- Pour un choix judicieux de $\rho^{(k)}$ et $\mathbf{d}^{(k)}$ cette suite converge vers $\mathbf{w}^\star$ solution du problème (10) [Ber99] et [NW06]

Le choix du pas $\rho^{(k)}$ est un problème d'optimisation en une dimension (*line search*, méthode d'Armijo).

Illustration 2d

$$J(\mathbf{w}) = \frac{1}{2}\mathbf{w}^\top A\mathbf{w} - \mathbf{b}^\top \mathbf{w}$$

lignes d'iso cout : $\{\mathbf{w} \in \mathbb{R}^2 \mid J(\mathbf{w}) = \text{ Cte }\}$

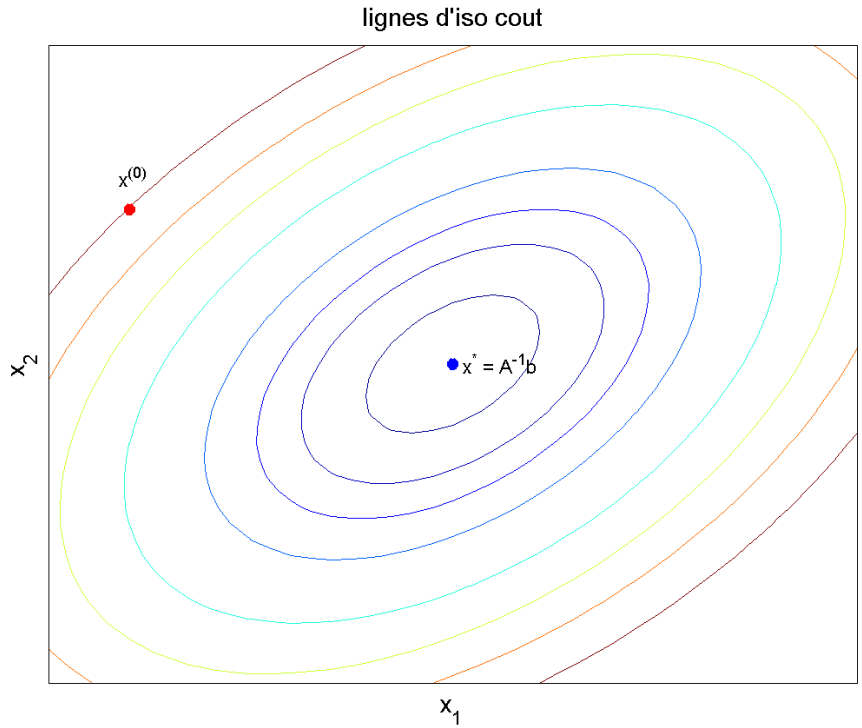
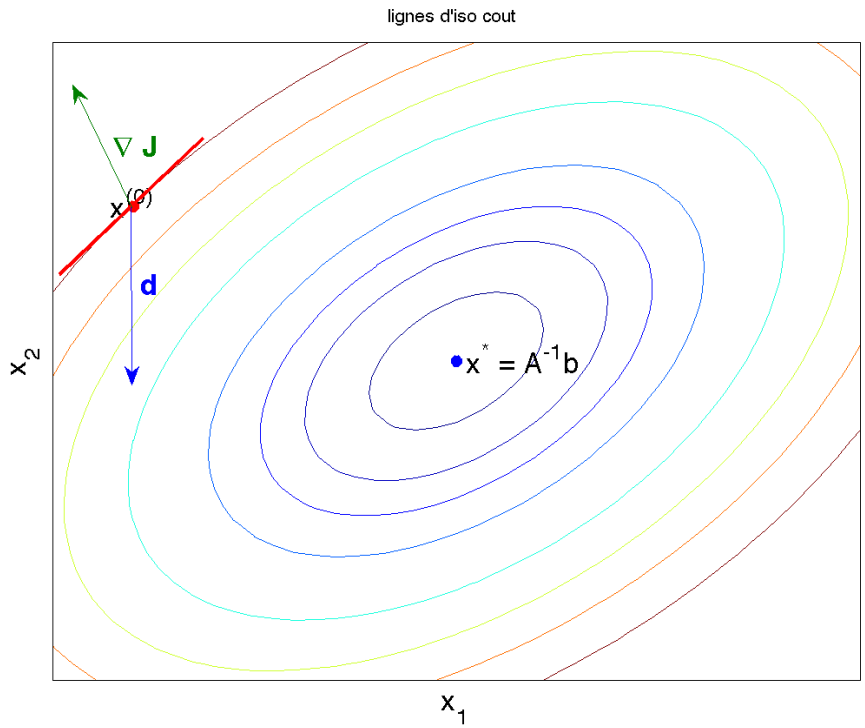


Illustration 2d

$$J(\mathbf{w}) = \frac{1}{2}\mathbf{w}^\top A\mathbf{w} - \mathbf{b}^\top \mathbf{w}$$

lignes d'iso cout : $\{\mathbf{w} \in \mathbb{R}^2 \mid J(\mathbf{w}) = \text{ Cte }\}$



une direction de descente $\mathbf{d}$ doit vérifier : $\mathbf{d}^\top \underbrace{(A\mathbf{w} - \mathbf{b})}_{\nabla J(\mathbf{w})} < 0$

Interprétation du gradient

Minimiser à chaque itération une approximation locale du problème

Definition

Une fonction J est gradient Lispchitz si il existe une constante L_J telle que

$$\|\nabla J(\mathbf{w}+\mathbf{d}) - \nabla J(\mathbf{w})\| \leq L_J \|\mathbf{d}\|, \qquad \forall \mathbf{d} \in \mathbb{R}^d, \forall \mathbf{w} \in \mathbb{R}^d. \tag{12}$$

La constante L_J est appelée la constante de Lipschitz de ∇J.

Example

Le coût des moindres carrés $J_1(\mathbf{w})$ est gradient Lispchitz avec une constante $L_J = \|\mathbf{X}^\top \mathbf{X}\|$. En effet, $\nabla J_1(\mathbf{w}+\mathbf{d}) = \mathbf{X}^\top(\mathbf{X}\mathbf{w} - Y) + \mathbf{X}^\top \mathbf{X}\mathbf{d}$, de sorte que

$$\nabla J_1(\mathbf{w}+\mathbf{d}) - \nabla J_1(\mathbf{w}) = \mathbf{X}^\top \mathbf{X}\mathbf{d},$$

et

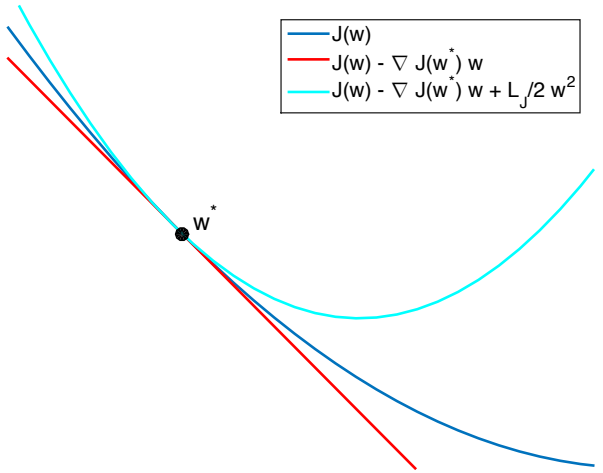
$$\|\nabla J_1(\mathbf{w}+\mathbf{d}) - \nabla J_1(\mathbf{w})\| \leq \|\mathbf{X}^\top \mathbf{X}\| \|\mathbf{d}\| = L_J \|\mathbf{d}\|.$$

Descent Lemma

[Ber99, Prop. A.24]

Si J est gradient Lispchitz, nous avons l'approximation suivante de J autour de $\mathbf{w}$ $\forall \mathbf{d} \in \mathbb{R}^d, \forall \mathbf{w} \in \mathbb{R}^d$.

$$J(\mathbf{w}+\mathbf{d}) \leq J(\mathbf{w}) + \boldsymbol{\nabla} J(\mathbf{w})^\top \mathbf{d} + \frac{L_J}{2}\|\mathbf{d}\|^2$$



Garantir la décroissance stricte du coût et donc la convergence :

- avec $\mathbf{d}^{(k)} = -\frac{1}{L_J}\boldsymbol{\nabla} J(\mathbf{w}^{(k)})$.
- et soit pas de descente $\rho^{(k)} \leq \frac{1}{L_J}$

Le Hessien

Supposons maintenant que J soit deux fois différentiable.

Definition (Hessien)

Le Hessien $\boldsymbol{\nabla}^2 J(\mathbf{w})$ d'une fonction J au point $\mathbf{w}$ est la matrice de taille $d \times d$ de composantes $\nabla^2_{ij} J(\mathbf{w}) = \frac{\partial^2 J}{\partial w_i \partial w_j}(\mathbf{w})$

Example

Le Hessien de J_1 le coût des moindre carrés de l'exemple (1) est

$$\boldsymbol{\nabla}^2 J_1(x) = \mathbf{X}^\top \mathbf{X}.$$

J est convexe si et seulement si

$$\forall \mathbf{w} \in \mathbb{R}^d,\ \boldsymbol{\nabla}^2 J(\mathbf{w}) \text{est une matrice définie posive}$$

La méthode de Newton

utiliser l'approximation locale du second ordre suivante pour minimiser J :

$$J(\mathbf{w}+\mathbf{d}) = J(\mathbf{w}) + \mathbf{d}^\top \boldsymbol{\nabla} J(\mathbf{w}) + \tfrac{1}{2}\mathbf{d}^\top \boldsymbol{\nabla}^2 J(\mathbf{w})\mathbf{d} + o(\|\mathbf{d}\|^2). \qquad (13)$$

Definition (la méthode de Newton)

minimiser, à chaque itération, l'approximation quadratique de J

$$\mathbf{d}^{(k)} = -(\boldsymbol{\nabla}^2 J)^{-1}(\mathbf{w}^{(k)})\boldsymbol{\nabla} J(\mathbf{w}^{(k)})$$

Example

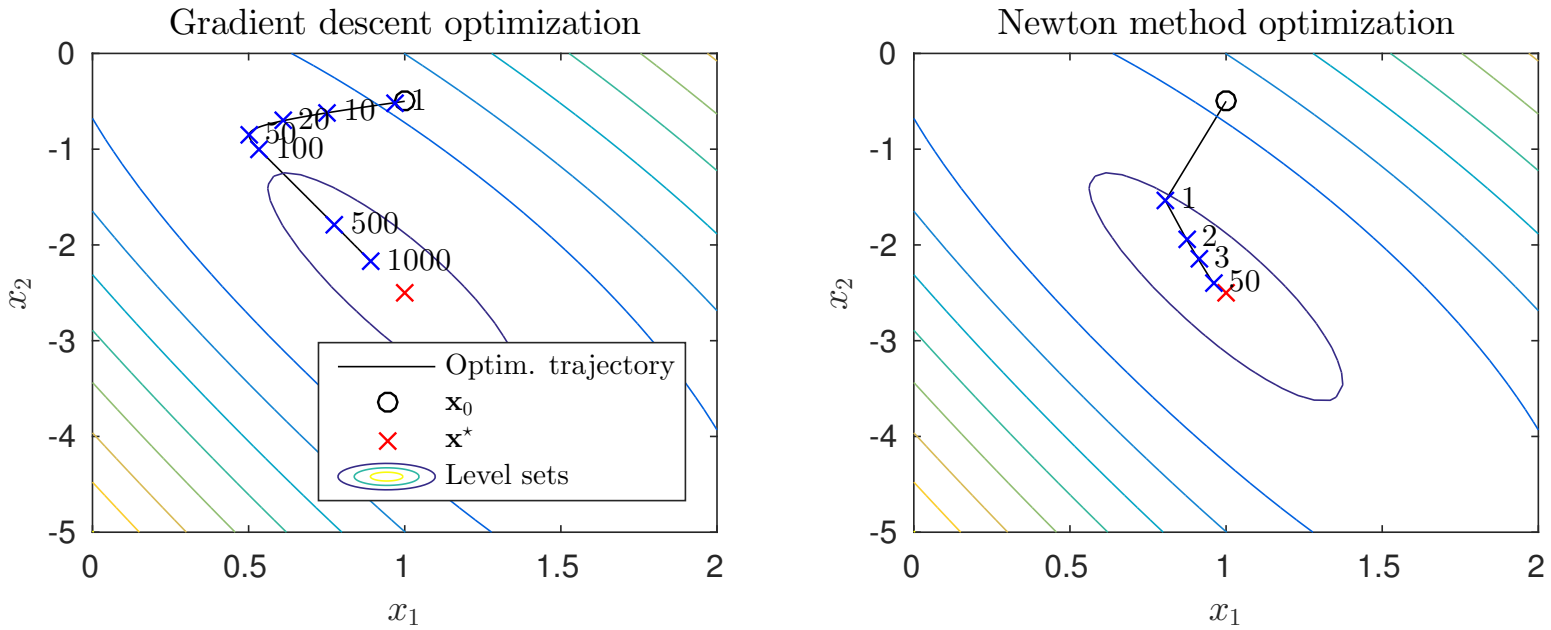
Les itérations des méthodes de descende de gradient et de Newton pour le problème de minimisation des moindres carrés sont :

$\mathbf{w}^{(k+1)} = \mathbf{w}^{(k)} - \rho^{(k)}\mathbf{X}^\top(\mathbf{X}\mathbf{w}^{(k)} - Y)$, Gradient

$\mathbf{w}^{(k+1)} = \mathbf{w}^{(k)} - (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top(\mathbf{X}\mathbf{w}^{(k)} - Y)$. Newton

Dans ce cas particulier, la méthode de Newton converge en une itération.

Illustration de la méthode de newton



méthode	*direction de descente*	*temps de calcul*	*convergence*
gradient	$d = -\nabla J$	$\mathcal{O}(n)$	lineaire
quasi Newton	$d = -B\nabla J$	$\mathcal{O}(n^2)$	super linéaire
Newton	$d = -H^{-1}\nabla J$	$\mathcal{O}(n^3)$	quadratic

le temps de calcul du pas optimal peut aussi varier

si n est très grand, utiliser une approximation stochastique du gradient comme direction de descente [pour plus de détails voir par exemple Bot10]

Exemple : la régression logistique

La régression logistique binaire avec $Y \in \{0,1\}^n$,

$$\min_{\mathbf{w}\in\mathbb{R}^p} J_\ell(\mathbf{w}) = \sum_{i=1}^{n} \left(-Y_i(\mathbf{X}\mathbf{w})_i + \log\left(1+\exp(\mathbf{X}\mathbf{w})_i\right)\right), \tag{14}$$

avec $(\mathbf{X}\mathbf{w})_i$ la $i^{\text{ème}}$ composante du vecteur $\mathbf{X}\mathbf{w}$.
La fonction coût $J_\ell(\mathbf{w})$ étant la composition de fonctions deux fois différentiables, son gradient et sa hessienne existent tous les deux et sont :

$$\begin{aligned} \boldsymbol{\nabla} J_\ell(x) &= \mathbf{X}^\top(\mathbf{p} - Y) \\ \boldsymbol{\nabla}^2 J_\ell(x) &= \mathbf{X}^\top \mathbf{P}\mathbf{X}, \end{aligned} \tag{15}$$

avec $\mathbf{p} \in \mathbb{R}^p$ le vecteur de composantes
$p_i(\mathbf{w}) = \exp((\mathbf{X}\mathbf{w})_i)/(1+\exp((\mathbf{X}\mathbf{w})_i))$ et $\mathbf{P}$ une matrice diagonale de terme général $\mathbf{P}_{ii} = p_i(1-p_i), i = 1, n$. Les itérations de Newton construisent la suite suivante :

$$\mathbf{w}^{(k+1)} = \mathbf{w}^{(k)} - (\mathbf{X}^\top \mathbf{P}\mathbf{X})^{-1}\mathbf{X}^\top(\mathbf{p} - Y).$$

La régression logistique : bricolage pratique

$$\begin{aligned}\mathbf{w}^{(k)} - (\mathbf{X}^\top \mathbf{P}\mathbf{X})^{-1}\mathbf{X}^\top(\mathbf{p} - Y) &= (\mathbf{X}^\top \mathbf{P}\mathbf{X})^{-1}\big((\mathbf{X}^\top \mathbf{P}\mathbf{X})\mathbf{w}^{(k)} - \mathbf{X}^\top(\mathbf{p} - Y)\big) \\ &= (\mathbf{X}^\top \mathbf{P}\mathbf{X})^{-1}\mathbf{X}^\top \mathbf{P}\mathbf{z},\end{aligned}$$

Algorithm 1 Méthode de Newton pour la régression logistique

Entrées : $\mathbf{X}$, Y données d'apprentissage.
Sortie : $\mathbf{w}$ paramètres du modèle
$\mathbf{w} \leftarrow 0$ initialisation des paramètres
while on n'a pas convergé **do**
  $\mathbf{p} \leftarrow \dfrac{\exp^{\mathbf{X}\mathbf{w}}}{1 + \exp^{\mathbf{X}\mathbf{w}}}$ division terme à terme
  $P_{ii} \leftarrow p_i(1 - p_i), \quad i = 1, n$ équation (15)
  $\mathbf{z} \leftarrow \mathbf{X}\mathbf{w} + \mathbf{P}^{-1}(Y - \mathbf{p})$
  $\mathbf{w} \leftarrow \left(\mathbf{X}^\top \mathbf{P}\mathbf{X}\right)^{-1}\mathbf{X}^\top \mathbf{P}\,\mathbf{z}$
end while

Plan

1. Exemples de problèmes d'optimisation en statistique
2. Optimisation différentiable, convexe et sans contraintes
3. Optimisation différentiable, convexe et avec contraintes
4. Optimisation sans contraintes non différentiable

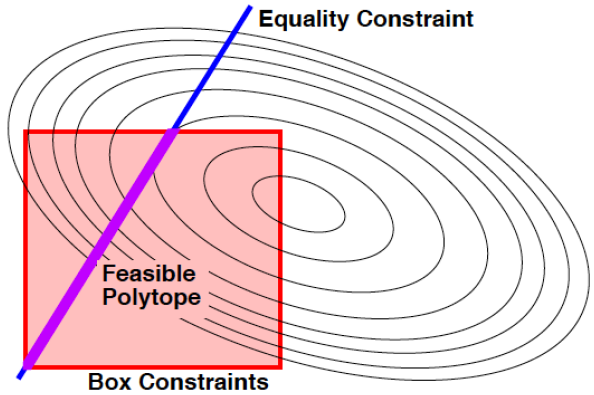
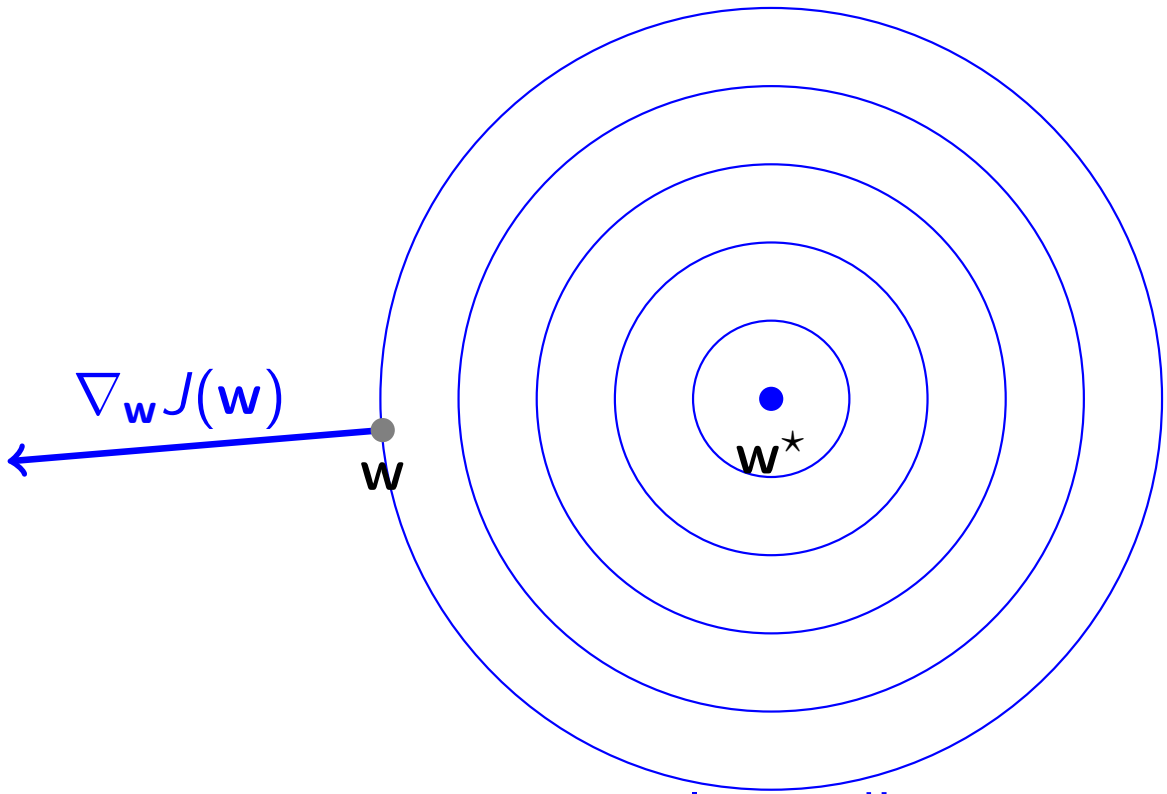


Figure from L. Bottou & C.J. Lin, Support vector machine solvers, in Large scale kernel machines, 2007.

Un exemple simple (pour commencer)

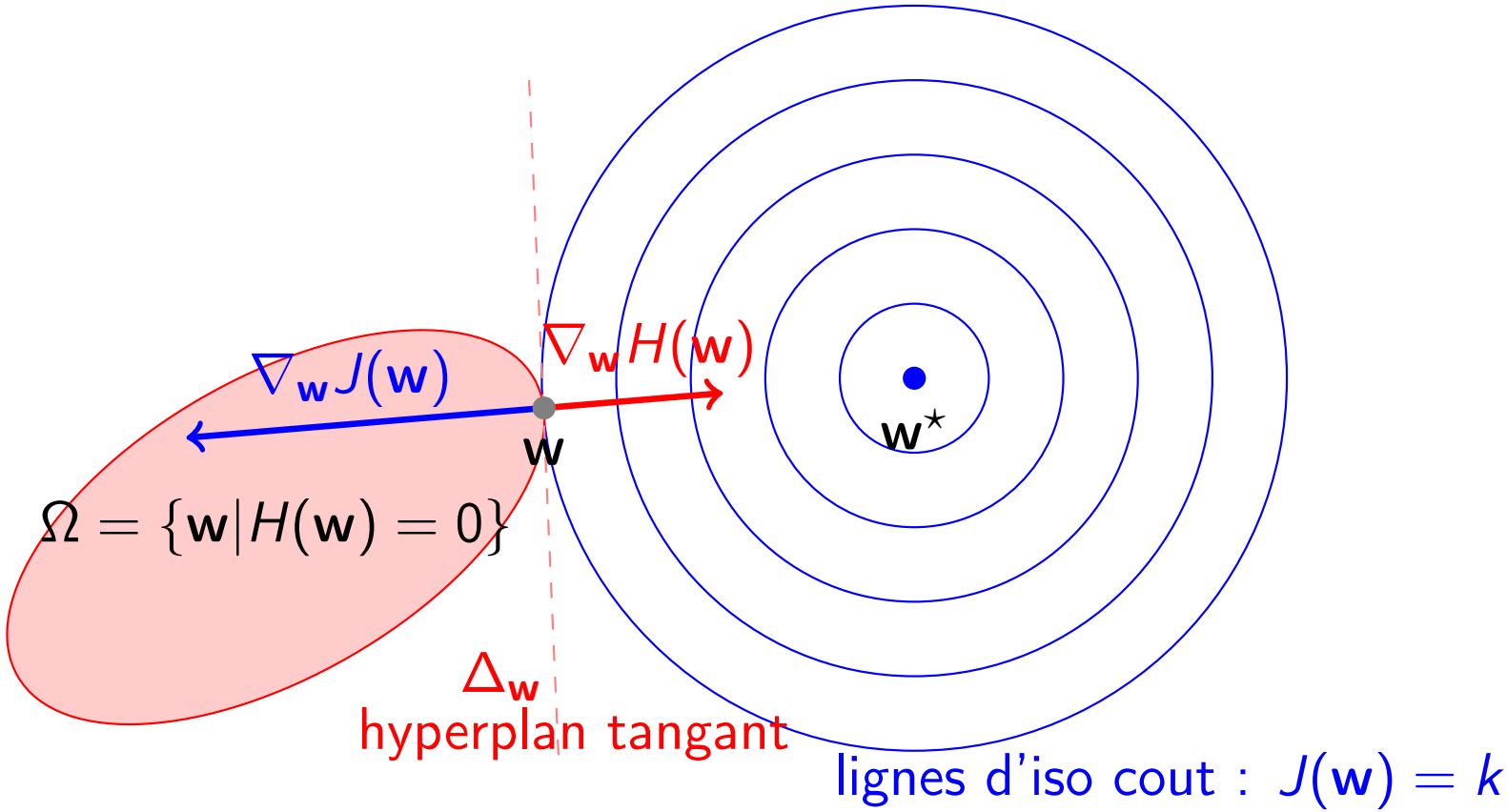
$$\begin{cases} \min\limits_{w_1, w_2} & J(\mathbf{w}) = (w_1 - a)^2 + (w_2 - b)^2 \\ \text{avec} \end{cases}$$



lignes d'iso cout : $J(\mathbf{w}) = k$

Un exemple simple (pour commencer)

$$\begin{cases} \min\limits_{w_1,w_2} & J(\mathbf{w}) = (w_1 - a)^2 + (w_2 - b)^2 \\ \text{avec} & H(\mathbf{w}) = \alpha(w_1 - c)^2 + \beta(w_2 - d)^2 + \gamma w_1 w_2 - 1 \end{cases}$$



$$\nabla_{\mathbf{w}} H(\mathbf{w}) = \mu \, \nabla_{\mathbf{w}} J(\mathbf{w})$$

Le cas d'une seule contrainte d'égalité

$$\begin{cases} \min\limits_{\mathbf{w}} & J(\mathbf{w}) & J(\mathbf{w}+\varepsilon\mathbf{d}) \approx J(\mathbf{w}) + \varepsilon\nabla_{\mathbf{w}} J(\mathbf{w})^\top \mathbf{d} \\ \text{with} & H(\mathbf{w}) = 0 & H(\mathbf{w}+\varepsilon\mathbf{d}) \approx H(\mathbf{w}) + \varepsilon\nabla_{\mathbf{w}} H(\mathbf{w})^\top \mathbf{d} \end{cases}$$

Cout J : $\mathbf{d}$ est une direction de descente s'il existe $\varepsilon_0 \in \mathbb{R}$ tel que $\forall \varepsilon \in \mathbb{R},\ 0 < \varepsilon \leq \varepsilon_0$

$$J(\mathbf{w}+\varepsilon\mathbf{d}) < J(\mathbf{w}) \quad \Rightarrow \quad \nabla_{\mathbf{w}} J(\mathbf{w})^\top \mathbf{d} < 0$$

Contrainte H : $\mathbf{d}$ est une direction de descente admissible s'il existe $\varepsilon_0 \in \mathbb{R}$ tel que $\forall \varepsilon \in \mathbb{R},\ 0 < \varepsilon \leq \varepsilon_0$

$$H(\mathbf{w}+\varepsilon\mathbf{d}) = 0 \quad \Rightarrow \quad \nabla_{\mathbf{w}} H(\mathbf{w})^\top \mathbf{d} = 0$$

si au point $\mathbf{w}^\star$, les vecteurs $\nabla_{\mathbf{w}} J(\mathbf{w}^\star)$ et $\nabla_{\mathbf{w}} H(\mathbf{w}^\star)$ sont colinéaires, il n'existe pas de direction $\mathbf{d}$ de descente admissible (sauf maximum).
$\rightarrow$ $\mathbf{w}^\star$ est donc une solution locale du problème.

Multiplicateurs de Lagrange

Supposons que J et les fonctions H_i sont continument différentiables (et indépendantes).

$$\mathcal{P} = \begin{cases} \min\limits_{\mathbf{w}\in\mathbb{R}^n} & J(\mathbf{w}) \\ \text{avec} & H_1(\mathbf{w}) = 0 \\ \text{et} & H_2(\mathbf{w}) = 0 \\ & \dots \\ & H_p(\mathbf{w}) = 0 \end{cases}$$

Multiplicateurs de Lagrange

Supposons que J et les fonctions H_i sont continument différentiables (et indépendantes).

$$\mathcal{P} = \begin{cases} \min\limits_{\mathbf{w}\in\mathbb{R}^n} & J(\mathbf{w}) & \\ \text{avec} & H_1(\mathbf{w}) = 0 & \mu_1 \\ \text{et} & H_2(\mathbf{w}) = 0 & \mu_2 \\ & \dots & \\ & H_p(\mathbf{w}) = 0 & \mu_p \end{cases}$$

chaque contrainte est associée avec μ_i : un multiplicateurs de Lagrange.

Multiplicateurs de Lagrange

Supposons que J et les fonctions H_i sont continument différentiables (et indépendantes).

$$\mathcal{P} = \begin{cases} \min\limits_{\mathbf{w}\in\mathbb{R}^n} & J(\mathbf{w}) & \\ \text{avec} & H_1(\mathbf{w}) = 0 & \mu_1 \\ \text{et} & H_2(\mathbf{w}) = 0 & \mu_2 \\ & \dots & \\ & H_p(\mathbf{w}) = 0 & \mu_p \end{cases}$$

chaque contrainte est associée avec μ_i : un multiplicateurs de Lagrange.

Theorem (Conditions d'optimalité du premier ordre)

Pour qu'un point $\mathbf{w}$ *soit un extremum local le* $\mathcal{P}$*, il est nécessaire que :*

$$\nabla_x J(\mathbf{w}^\star) + \sum_{i=1}^{p} \mu_i \nabla_x H_i(\mathbf{w}^\star) = 0 \qquad \textit{et} \qquad H_i(\mathbf{w}^\star) = 0, \quad i = 1, p$$

Un exemple où ça marche

$$\begin{cases} \min\limits_{\mathbf{w}\in\mathbb{R}^3} & J(\mathbf{w}) = -w_1 w_2 - w_1 w_3 - w_2 w_3 \\ \text{avec} & H(\mathbf{w}) = w_1 + w_2 + w_3 - 3 = 0 \end{cases}$$

Un exemple où ça marche

$$\begin{cases} \min\limits_{\mathbf{w}\in\mathbb{R}^3} & J(\mathbf{w}) = -w_1 w_2 - w_1 w_3 - w_2 w_3 \\ \text{avec} & H(\mathbf{w}) = w_1 + w_2 + w_3 - 3 = 0 \end{cases}$$

$$\nabla_x J(x) = -\begin{pmatrix} w_2 + w_3 \\ w_1 + w_3 \\ w_1 + w_2 \end{pmatrix} \qquad \nabla_x H(x) = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

Les conditions d'optimalité sont

$$\begin{cases} & -w_2 & -w_3 & +\mu & = 0 \\ -w_1 & & -w_3 & +\mu & = 0 \\ -w_1 & -w_2 & & +\mu & = 0 \\ w_1 & +w_2 & +w_3 & & = 3 \end{cases}$$

la résolution du système $(A\backslash b)$ donne :

$$w_1 = w_2 = w_3 = 1 \text{ et } \mu = 2$$

Un autre exemple où ça marche

$$\begin{cases} \min\limits_{\mathbf{w}\in\mathbb{R}^n} & J(\mathbf{w}) = -\frac{1}{2}\,\mathbf{w}^\top A\mathbf{w} - \mathbf{w}^\top \mathbf{b} \\ \text{avec} & C\mathbf{w} = \mathbf{d} \end{cases}$$

$$\begin{aligned} \nabla_{\mathbf{w}} J &= A\mathbf{w} - \mathbf{b} \\ \nabla_{\mathbf{w}} H &= C^\top \end{aligned}$$

$$\begin{aligned} \nabla_x J(x) + \sum_{i=1}^{p} \mu_i \nabla_x H_i(\mathbf{w}) = 0 \quad &\Rightarrow \quad A\mathbf{w} + C^\top \mu = \mathbf{b} \\ H(\mathbf{w}) = 0 \quad &\Rightarrow \quad C\mathbf{w} = \mathbf{d} \end{aligned}$$

...et on résout le système linéaire.

Attention si en plus on cherche $\mathbf{w} \geq 0$ ça devient plus compliqué

Un exemple où ça ne marche pas

$$\begin{cases} \min\limits_{w_1,w_2} & w_1 + w_2 \\ \text{avec} & (w_1+1)^2 + w_2^2 = 1 \\ \text{et} & (w_1-2)^2 + w_2^2 = 4 \end{cases}$$

le minimum est $(0,0)$ l'unique solution réalisable ! Dans ce cas, il n'existe pas de multiplicateurs de Lagrange

lagrangien

Une fonction bien pratique :

définition : lagrangien

On appelle lagrangien du problème $\mathcal{P}$ la fonction L définie par :

$$L(\mathbf{w}, \mu) = J(x) + \sum_{i=1}^{p} \mu_i H_i(\mathbf{w})$$

Grâce au lagrangien on retrouve les conditions d'optimalité :

$$\begin{cases} \nabla_x L(\mathbf{w}, \mu) = 0 & \Rightarrow \quad \nabla_x J(\mathbf{w}) + \sum_{i=1}^{p} \mu_i \nabla_x H_i(\mathbf{w}) = 0 \\ \nabla_{\mu_i} L(\mathbf{w}, \mu) = 0 & \Rightarrow \quad H_i(\mathbf{w}) = 0 \end{cases}$$

Interprétation graphique : optimisation multicritères.

Le cas d'une seule contrainte d'inégalité

$$\begin{cases} \min\limits_{\mathbf{w}} & J(\mathbf{w}) & J(\mathbf{w}+\varepsilon\mathbf{d}) \approx J(\mathbf{w}) + \varepsilon\nabla_{\mathbf{w}} J(\mathbf{w})^\top \mathbf{d} \\ \text{avec} & G(\mathbf{w}) \leq 0 & G(\mathbf{w}+\varepsilon\mathbf{d}) \approx G(\mathbf{w}) + \varepsilon\nabla_{\mathbf{w}} G(\mathbf{w})^\top \mathbf{d} \end{cases}$$

Cout J : $\mathbf{d}$ est une direction de descente, s'il existe un $\varepsilon_0 \in \mathbb{R}$ tel que $\forall \varepsilon \in \mathbb{R},\ 0 < \varepsilon \leq \varepsilon_0$

$$J(\mathbf{w}+\varepsilon\mathbf{d}) < J(\mathbf{w}) \quad \Rightarrow \quad \nabla_{\mathbf{w}} J(\mathbf{w})^\top \mathbf{d} < 0$$

Contrinate G : $\mathbf{d}$ est une direction de descente admissible, s'il existe un $\varepsilon_0 \in \mathbb{R}$ tel que $\forall \varepsilon \in \mathbb{R},\ 0 < \varepsilon \leq \varepsilon_0$

$$G(\mathbf{w}+\varepsilon\mathbf{d}) \leq 0 \quad \Rightarrow \quad \begin{array}{ll} G(\mathbf{w}) < 0 : & \text{no limit here on } \mathbf{d} \\ G(\mathbf{w}) = 0 : & \nabla_{\mathbf{w}} G(\mathbf{w})^\top \mathbf{d} \leq 0 \end{array}$$

Les deux alternatives pour caractériser l'optimalité

Si $x^\star$ se trouve sur la limite du domaine admissible ($G(\mathbf{w}^\star) = 0$), alors si les vecteurs $\nabla_{\mathbf{w}} J(\mathbf{w}^\star)$ et $\nabla_{\mathbf{w}} G(\mathbf{w}^\star)$ sont colinéaires et dans des directions opposées, il n'existe pas de direction de descente admissible $\mathbf{d}$ à ce point. $x^\star$ est donc un optimum local... Ou alors $\nabla_{\mathbf{w}} J(\mathbf{w}^\star) = 0$

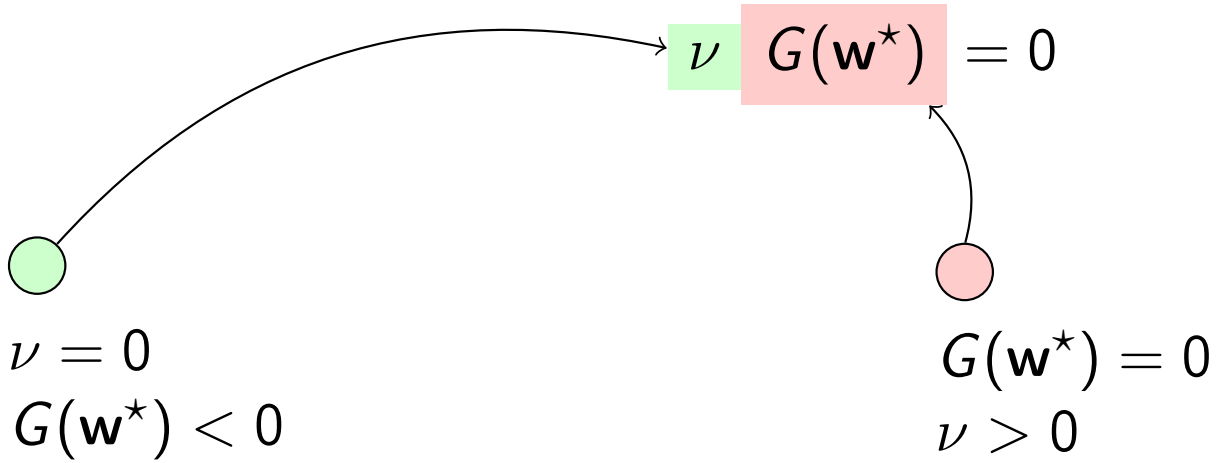
Les deux alternatives à l'optimum

$$\nabla_{\mathbf{w}} J(\mathbf{w}^\star) = -\nu \, \nabla_{\mathbf{w}} G(\mathbf{w}^\star) \qquad \text{et} \quad \nu > 0; G(\mathbf{w}^\star) = 0$$

or

$$\nabla_{\mathbf{w}} J(\mathbf{w}^\star) = 0 \qquad \text{et} \quad \nu = 0; G(\mathbf{w}^\star) < 0$$

Cette alternative est résumée par la condition de complémentarité



Conditions d'optimalité du premier ordre (1)

$$\text{problem } \mathcal{P} = \begin{cases} \min\limits_{\mathbf{w}\in\mathbb{R}^d} & J(\mathbf{w}) & \\ \text{avec} & h_j(\mathbf{w}) = 0 & j = 1,\dots,\ell \\ \text{et} & g_i(\mathbf{w}) \leq 0 & i = 1,\dots,q \end{cases}$$

Definition (Conditions de Karush, Kuhn and Tucker)

$$\text{stationarité} \quad \nabla J(\mathbf{w}^\star) + \sum_{j=1}^{\ell} \mu_j \nabla h_j(\mathbf{w}^\star) + \sum_{i=1}^{q} \nu_i \nabla g_i(\mathbf{w}^\star) = 0$$

$$\text{admissibilité primale} \quad h_j(\mathbf{w}^\star) = 0 \qquad j = 1,\dots,\ell$$
$$g_i(\mathbf{w}^\star) \leq 0 \qquad i = 1,\dots,q$$

$$\text{admissibilité duale} \quad \nu_i \geq 0 \qquad i = 1,\dots,q$$

$$\text{complementarité} \quad \nu_i g_i(\mathbf{w}^\star) = 0 \qquad i = 1,\dots,q$$

μ_j et ν_i sont les multiplicateurs de Lagrange du problème $\mathcal{P}$

Example (Les conditions KKT dans un exemple simple)

Considérons le problème d'optimisation en une dimension :

$$\min_{w \in \mathbb{R}} \quad \frac{1}{2}(w+1)^2 \quad \text{avec} \quad w \geq 0,$$

admet les conditions KKT suivantes :

stationnarité	$(w+1) - \nu = 0$
admissibilité primale	$-w \leq 0$
admissibilité duale	$\nu \geq 0$
complémentarité	$\nu w = 0.$

La condition de complémentarité impose que w ou ν soit nul. S'il s'agit de ν, alors la condition de stationnarité donne $w = -1$, valeur non admissible. Donc la solution est $w = 0$ et $\nu = 1$. On dit alors que la condition d'admissibilité primale (la contrainte) est saturée ou active.

Example (Moindres carrés positifs)

Il s'agit d'un cas particulier du problème (9) avec le coût des moindres carrés et sans pénalisation. Le problème s'écrit alors :

$$\begin{cases} \min\limits_{\mathbf{w}\in\mathbb{R}^n} & \frac{1}{2}\|\mathbf{X}\mathbf{w} - Y\|^2 \\ \text{avec} & 0 \leq w_i, \qquad\qquad i = 1, \ldots, p \end{cases}$$

Les conditions KKT de ce problème sont :

stationnarité	$\mathbf{X}^t(\mathbf{X}\mathbf{w} - Y) - \boldsymbol{\nu} = \mathbf{0}$
admissibilité primale	$-\mathbf{w} \leq \mathbf{0}$
admissibilité duale	$\boldsymbol{\nu} \geq \mathbf{0}$
complémentarité	$\mathrm{diag}(\boldsymbol{\nu})\mathbf{w} = \mathbf{0}.$

Conditions d'optimalité du premier ordre (2)

Théorème (12.1 Nocedal & Wright pp 321)

Si le vecteur $x^\star$ est un point stationnaire du problème $\mathcal{P}$
Alors il existe [a] des multiplicateurs de Lagrange $\left(\mathbf{w}^\star, \{\mu_j\}_{j=1:\ell}, \{\nu_i\}_{i=1:q}\right)$ vérifiant les conditions KKT

[a] sous certaines conditions sur les contraintes, par exemple leur indépendance

Si le problème est convexe, alors tout point stationnaire est solution du problème

Conditions KKT - Lagrangien (3)

$$\text{problem } \mathcal{P} = \begin{cases} \min\limits_{\mathbf{w}\in\mathbb{R}^n} & J(\mathbf{w}) \\ \text{avec} & h_j(\mathbf{w}) = 0 \quad j = 1, \ldots, p \\ \text{et} & g_i(\mathbf{w}) \leq 0 \; i = 1, \ldots, q \end{cases}$$

Définition : Lagrangien

Le lagrangien du problème $\mathcal{P}$ est la fonction suivante :

$$\mathcal{L}(\mathbf{w}, \mu, \nu) = J(x) + \sum_{j=1}^{\ell} \mu_j h_j(\mathbf{w}) + \sum_{i=1}^{q} \nu_i g_i(\mathbf{w})$$

L'importance du lagrangien

- la condition de stationnarité est donnée par : $\nabla \mathcal{L}(\mathbf{w}^\star, \mu, \nu) = 0$
- l'optimum est un point selle de lagrangien $\max\limits_{\mu,\nu} \min\limits_{\mathbf{w}} \mathcal{L}(\mathbf{w}, \mu, \nu)$

Variables primales : x et variables duales μ, ν (Les multiplicateurs de Lagrange)

Dualité – définitions (1)

Example (Lagrangien de l'exemple 18)

Il s'agit d'un cas particulier du problème (9) avec le coût des moindres carrés et sans pénalisation. Le problème s'écrit alors :

$$\begin{cases} \min\limits_{\mathbf{w}\in\mathbb{R}^n} & \frac{1}{2}\|\mathbf{X}\mathbf{w} - Y\|^2 \\ \text{avec} & 0 \leq w_i, \qquad\qquad i = 1, \ldots, p \end{cases}$$

Le Lagrangien est :

$$\mathcal{L}(\mathbf{w}, \boldsymbol{\nu}) = \frac{1}{2}\|\mathbf{X}\mathbf{w} - Y\|^2 - \boldsymbol{\nu}^t\mathbf{w}.$$

Example (L'exemple 2 et le lagrangien)

L et Ω convexes

$$\begin{cases} \min\limits_{\mathbf{w}\in\mathbb{R}^d} & L(\mathbf{w}) \\ \text{avec} & \Omega(\mathbf{w}) \leq k, \end{cases} \tag{16}$$

Le lagrangien de ce problème est

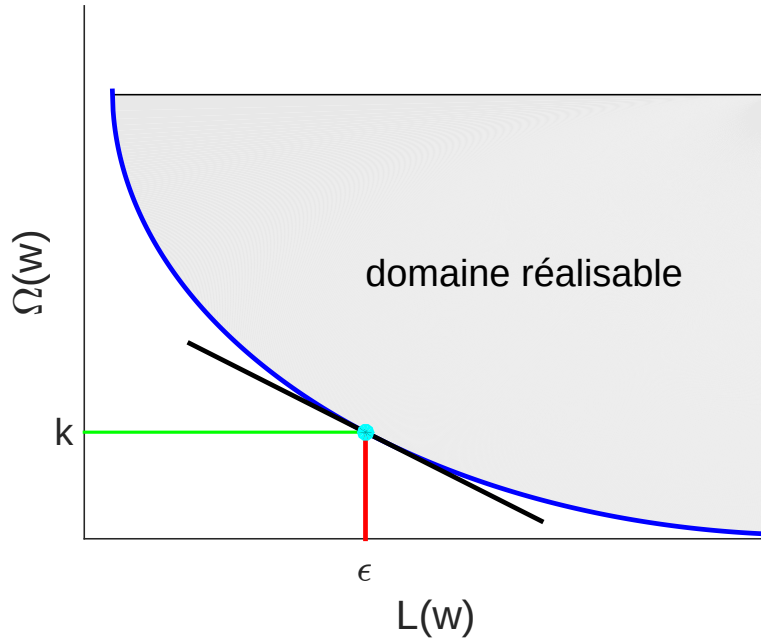
$$\mathcal{L}(\mathbf{w}, \nu) = L(\mathbf{w}) + \nu\big(\Omega(\mathbf{w}) - k\big),$$

pour un k donné, il existe un ν

$$\min_{\mathbf{w}\in\mathbb{R}^d} L(\mathbf{w}) + \nu\Omega(\mathbf{w}) \tag{17}$$

Ce même raisonnement peut être formulé à partir de problème suivant :

$$\begin{cases} \min\limits_{\mathbf{w}\in\mathbb{R}^d} & \Omega(\mathbf{w}) \\ \text{avec} & L(\mathbf{w}) \leq \epsilon, \end{cases} \tag{18}$$



Dualité – définitions (1)

Problème primal et son dual lagrangien

$$\mathcal{P} = \left\{ \begin{array}{ccc} \min\limits_{\mathbf{w}\in\mathbb{R}^n} & J(\mathbf{w}) & \\ \text{avec} & h_j(x) = 0 & j = 1, \ell \\ \text{et} & g_i(x) \leq 0 & i = 1, q \end{array} \right. \qquad \mathcal{D} = \left\{ \begin{array}{ccc} \max\limits_{\mu\in\mathbb{R}^p, \nu\in\mathbb{R}^q} & Q(\mu, \nu) & \\ \text{avec} & \nu_j \geq 0 & j = 1, q \end{array} \right.$$

La fonction objectif duale :

$$\begin{aligned} Q(\mu, \nu) &= \inf_x \ \mathcal{L}(\mathbf{w}, \mu, \nu) \\ &= \inf_x \ J(x) + \sum_{j=1}^{p} \mu_j h_j(x) + \sum_{i=1}^{q} \nu_i g_i(x) \end{aligned}$$

Le dual de Wolf

$$\mathcal{W} = \left\{ \begin{array}{ccc} \max\limits_{\mathbf{w},\mu\in\mathbb{R}^p, \nu\in\mathbb{R}^q} & \mathcal{L}(\mathbf{w}, \mu, \nu) & \\ \text{avec} & \nu_j \geq 0 & j = 1, q \\ \text{et} & \nabla J(x) + \sum\limits_{j=1}^{p} \mu_j \nabla h_j(x) + \sum\limits_{i=1}^{q} \nu_i \nabla g_i(x) = 0 & \end{array} \right.$$

Example (Calcul du problème dual de l'exemple 18)

$$Q(\boldsymbol{\nu}) = \inf_{\mathbf{w}\in\mathbb{R}^d} \quad \tfrac{1}{2}\|\mathbf{X}\mathbf{w} - Y\|^2 - \boldsymbol{\nu}^t\mathbf{w}$$

Le gradient du lagrangien par rapport à $\mathbf{w}$ est $\mathbf{X}^t(\mathbf{X}\mathbf{w} - Y) - \boldsymbol{\nu}$ ce qui nous donne $\mathbf{w}^\star = (\mathbf{X}^t\mathbf{X})^{-1}(\boldsymbol{\nu} + \mathbf{X}^t Y)$.

$$Q(\boldsymbol{\nu}) = -\tfrac{1}{2}(\mathbf{X}^t Y + \boldsymbol{\nu})(\mathbf{X}^t\mathbf{X})^{-1}(\mathbf{X}^t Y + \boldsymbol{\nu})$$

et le problème dual est

$$\begin{cases} \max\limits_{\boldsymbol{\nu}\in\mathbb{R}^q} & -\tfrac{1}{2}\boldsymbol{\nu}^t(\mathbf{X}^t\mathbf{X})^{-1}\boldsymbol{\nu} - Y^t\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\boldsymbol{\nu} \\ \text{avec} & \nu_j \geq 0, \qquad j = 1, \ldots, q, \end{cases}$$

la condition de stationnarité de l'exemple (18) $\boldsymbol{\nu} = \mathbf{X}^t(\mathbf{X}\mathbf{w} - Y)$, le problème dual peut être reformulé à travers les variables primales comme :

$$\begin{cases} \max\limits_{\mathbf{w}\in\mathbb{R}^d} & -\tfrac{1}{2}\mathbf{w}^t\mathbf{X}^t\mathbf{X}\mathbf{w} \\ \text{avec} & \mathbf{X}^t(\mathbf{X}\mathbf{w} - Y) \geq 0. \end{cases}$$

Dualité – théorèmes (2)

Théorème (12.12, 12.13 and 12.14 Nocedal & Wright pp 346)

Si J, et g_i sont convexes et continument différentiables et les h_j affines, [a], alors les solutions des problèmes primale et dual ont le même cout

[a] sous certaines conditions sur les contraintes, par exemple leur indépendance

$$\begin{aligned}(\mu^\star, \nu^\star) &= \text{solution du problem } \mathcal{D} \\ \mathbf{w}^\star &= \arg\min_{\mathbf{w}} \mathcal{L}(\mathbf{w}, \mu^\star, \nu^\star)\end{aligned}$$

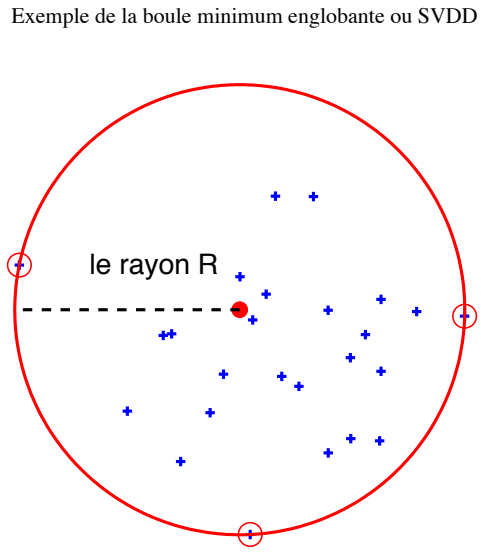
$$\begin{aligned}Q(\mu^\star, \nu^\star) = \arg\min_{\mathbf{w}} \mathcal{L}(\mathbf{w}, \mu^\star, \nu^\star) &= \mathcal{L}(\mathbf{w}^\star, \mu^\star, \nu^\star) \\ &= J(\mathbf{w}^\star) + \mu^\star H(\mathbf{w}^\star) + \nu^\star G(\mathbf{w}^\star) = J(\mathbf{w}^\star)\end{aligned}$$

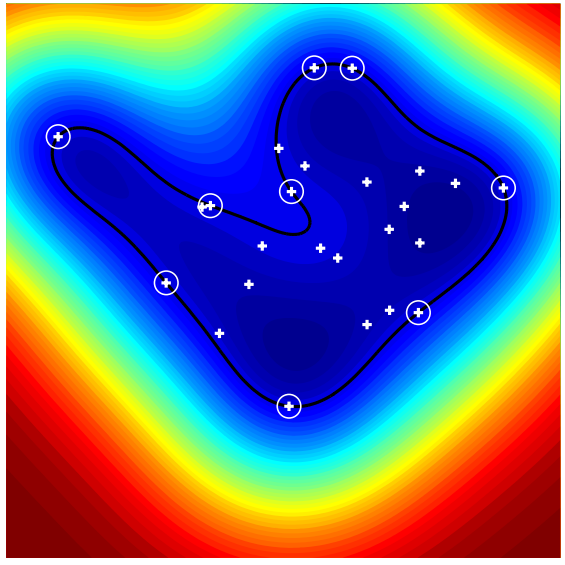
pour tout point $\mathbf{w}$ admissible

$$Q(\mu, \nu) \leq J(\mathbf{w}) \qquad \rightarrow \qquad 0 \leq J(\mathbf{w}) - Q(\mu, \nu)$$

le saut de dualité est la différence entre les couts primal et dual

Exemple : la boule minimum englobante ou SVDD





Étant donné un ensemble de n observations $\{X_i \in \mathbb{R}^p, i = 1, \ldots, n\}$ ce problème consiste à trouver la boule p dimensionnelle de centre $\mathbf{c}$ de rayon minimum R contenant tous les points X_i.

$$\begin{cases} \min\limits_{R\in\mathbb{R},\mathbf{c}\in\mathbb{R}^p} & R^2 \\ \text{avec} & \|X_i - \mathbf{c}\|^2 \leq R^2, \quad i = 1, \ldots, n. \end{cases} \tag{19}$$

Les KKT des SVDD

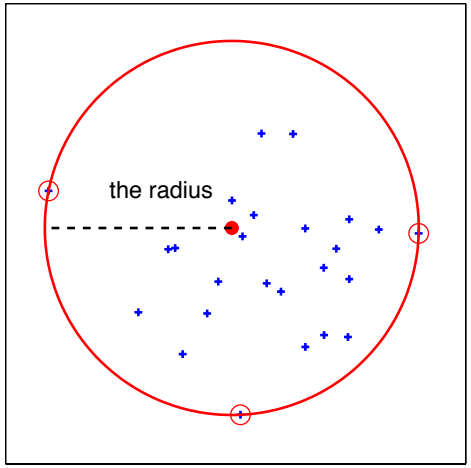
$$\mathcal{L}(\mathbf{c}, R, \nu) = R^2 + \sum_{i=1}^{n} \nu_i \left(\|X_i - \mathbf{c}\|^2 - R^2 \right)$$

KKT conditionns :

stationarté	$2\mathbf{c} \sum_{i=1}^{n} \nu_i - 2 \sum_{i=1}^{n} \nu_i X_i = 0$	$\leftarrow$ théorème de représentation
	$1 - \sum_{i=1}^{n} \nu_i = 0$	
primal admiss.	$\|X_i - \mathbf{c}\|^2 \leq R^2$	
dual admiss.	$\nu_i \geq 0$	$i = 1, n$
complementarité	$\nu_i \left(\|X_i - \mathbf{c}\|^2 - R^2 \right) = 0$	$i = 1, n$

Les KKT des SVDD

$$\mathcal{L}(\mathbf{c}, R, \nu) = R^2 + \sum_{i=1}^{n} \nu_i \left(\|X_i - \mathbf{c}\|^2 - R^2 \right)$$



KKT conditionns :

stationarté
- $2\mathbf{c} \sum_{i=1}^{n} \nu_i - 2 \sum_{i=1}^{n} \nu_i X_i = 0$ $\leftarrow$ théorème de représentation
- $1 - \sum_{i=1}^{n} \nu_i = 0$

primal admiss. $\|X_i - \mathbf{c}\|^2 \leq R^2$

dual admiss. $\nu_i \geq 0$ $\quad i = 1, n$

complementarité $\nu_i \left(\|X_i - \mathbf{c}\|^2 - R^2 \right) = 0$ $\quad i = 1, n$

La complémentarité définie deux groupes de points

les vecteurs support $\|X_i - \mathbf{c}\|^2 = R^2$ et les points de l'intérieur $\nu_i = 0$

MEB: Dual

Le théorème de représentation

$$\mathbf{c} = \frac{\sum_{i=1}^{n} \nu_i X_i}{\sum_{i=1}^{n} \nu_i} = \sum_{i=1}^{n} \nu_i X_i$$

$$\mathcal{L}(\nu) = \sum_{i=1}^{n} \nu_i \left(\|X_i - \sum_{j=1}^{n} \nu_j \mathbf{w}_j\|^2 \right)$$

$$\sum_{i=1}^{n}\sum_{j=1}^{n} \nu_i \nu_j X_i^\top x_j = \nu^\top G \nu \qquad \text{and} \qquad \sum_{i=1}^{n} \nu_i X_i^\top X_i = \nu^\top \text{diag}(G)$$

avec $G = XX\top$ la matrice de Gram : $G_{ij} = X_i^\top x_j$,

$$\begin{cases} \min\limits_{\nu \in \mathbf{R^n}} & \nu^\top G \nu - \nu^\top diag(G) \\ \text{avec} & e^\top \nu = 1 \\ \text{et} & 0 \leq \nu_i, \qquad\qquad i = 1 \dots n \end{cases}$$

MEB: Dual

Le théorème de représentation

$$\mathbf{c} = \frac{\sum_{i=1}^{n} \nu_i X_i}{\sum_{i=1}^{n} \nu_i} = \sum_{i=1}^{n} \nu_i X_i$$

$$\mathcal{L}(\nu) = \sum_{i=1}^{n} \nu_i \left(\|X_i - \sum_{j=1}^{n} \nu_j \mathbf{w}_j\|^2 \right)$$

$$\sum_{i=1}^{n}\sum_{j=1}^{n} \nu_i \nu_j X_i^\top x_j = \nu^\top G \nu \qquad \text{and} \qquad \sum_{i=1}^{n} \nu_i X_i^\top X_i = \nu^\top \text{diag}(G)$$

avec $G = XX\top$ la matrice de Gram : $G_{ij} = X_i^\top x_j$,

$$\begin{cases} \min\limits_{\nu \in \mathbb{R}^n} & \nu^\top G \nu - \nu^\top diag(G) \\ \text{avec} & e^\top \nu = 1 \\ \text{et} & 0 \leq \nu_i, \qquad\qquad i = 1 \dots n \end{cases}$$

SVDD primal vs. dual

Primal

$$\begin{cases} \min\limits_{R\in\mathbb{R},\mathbf{c}\in\mathbb{R}^p} & R^2 \\ \text{avec} & \|X_i - \mathbf{c}\|^2 \leq R^2, \\ & i = 1, \ldots, n \end{cases}$$

- $p + 1$ inconnues
- n contraintes
- peut être reformulé QP
- bien quand $d << n$

Dual

$$\begin{cases} \min\limits_{\nu} & \nu^\top G\nu - \nu^\top diag(G) \\ \text{avec} & e^\top \nu = 1 \\ \text{et} & 0 \leq \nu_i, \\ & i = 1 \ldots n \end{cases}$$

- n inconnues avec G la matrice des influences des couples
- n contraintes de boite
- moins difficiles à résoudre
- à utiliser lorsque $d > n$

SVDD primal vs. dual

Primal

$$\begin{cases} \min\limits_{R\in\mathbb{R},\mathbf{c}\in\mathbb{R}^p} & R^2 \\ \text{avec} & \|X_i - \mathbf{c}\|^2 \leq R^2, \\ & i = 1, \ldots, n \end{cases}$$

- $p + 1$ inconnues
- n contraintes
- peut être reformulé QP
- bien quand $d << n$

Dual

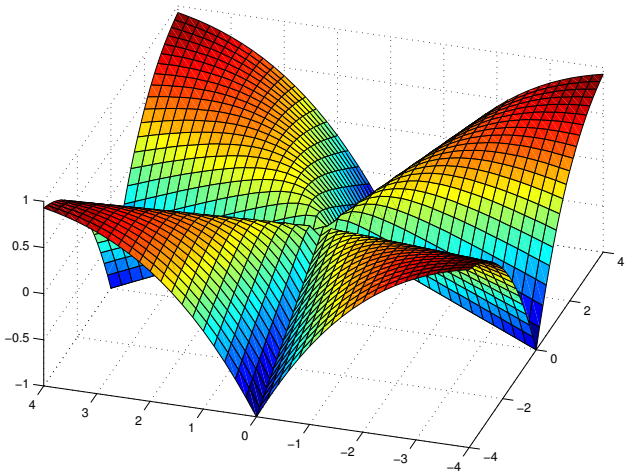
$$\begin{cases} \min\limits_{\nu} & \nu^\top G\nu - \nu^\top diag(G) \\ \text{avec} & e^\top \nu = 1 \\ \text{et} & 0 \leq \nu_i, \\ & i = 1 \ldots n \end{cases}$$

- n inconnues avec G la matrice des influences des couples
- n contraintes de boite
- moins difficiles à résoudre
- à utiliser lorsque $d > n$

Et ce sont tous les deux des QP convexes

Plan

1. Exemples de problèmes d'optimisation en statistique
2. Optimisation différentiable, convexe et sans contraintes
3. Optimisation différentiable, convexe et avec contraintes
4. Optimisation sans contraintes non différentiable

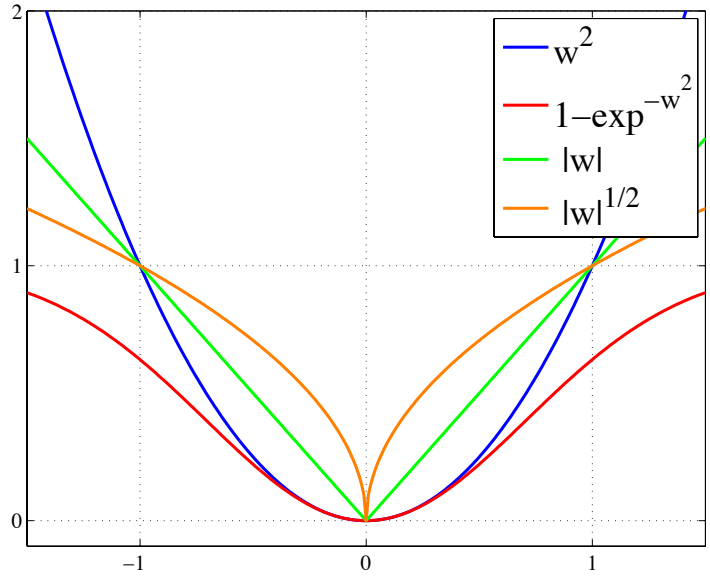


Formulation des problèmes d'optimisation non différentiable

Example (Les pénalités décomposables)

$$\min_{\mathbf{w}\in\mathbb{R}^d} \quad L(\mathbf{w}) + \lambda\,\Omega(\mathbf{w}), \qquad\qquad \Omega(\mathbf{w}) = \sum_{i=1}^{d} \omega(w_i).$$

Pénalité	Convexe	Diff.
$\omega(w) = w^2$ $\Omega(\mathbf{w}) = \|\mathbf{w}\|^2$	✓	✓
$\omega(w) = 1 - \exp^{-w^2}$	✗	✓
$\omega(w) = \|w\|$ $\Omega(\mathbf{w}) = \|\mathbf{w}\|_1$	✓	✗
$\omega(w) = \sqrt{\|w\|}$ $\Omega(\mathbf{w}) = \|\mathbf{w}\|_{1/2}$	✗	✗



Le cas non différentiable en zéro

L'approche proximale

Supposons $L(\mathbf{w})$ Lipschitz gradient, le *descent lemma* (22) stipule

$$J(\mathbf{w}) \leq L(\mathbf{w}^{(k)}) + \boldsymbol{\nabla} L(\mathbf{w}^{(k)})^\top (\mathbf{w} - \mathbf{w}^{(k)}) + \frac{1}{2\rho} \|\mathbf{w} - \mathbf{w}^{(k)}\|^2 + \lambda \Omega(\mathbf{w}),$$

$\rho \leq \frac{1}{L_L}$ avec L_L la constante de Lipschitz de la fonction $L(\mathbf{w})$.

Minimisation de cette majoration

$$\min_{\mathbf{w} \in \mathbb{R}^d} \quad L(\mathbf{w}^{(k)}) + \boldsymbol{\nabla} L(\mathbf{w}^{(k)})^\top (\mathbf{w} - \mathbf{w}^{(k)}) + \frac{1}{2\rho} \|\mathbf{w} - \mathbf{w}^{(k)}\|^2 + \lambda \Omega(\mathbf{w}),$$

est équivalent à

$$\min_{\mathbf{w} \in \mathbb{R}^d} \quad \tfrac{1}{2} \|\mathbf{w} - \mathbf{u}\|^2 + \lambda \rho \Omega(\mathbf{w}),$$

avec $\mathbf{u} = \mathbf{w}^{(k)} - \rho \boldsymbol{\nabla} L(\mathbf{w}^{(k)})$.

L'opérateur proximal

$$\min_{\mathbf{w}\in\mathbb{R}^d} \quad \tfrac{1}{2}\|\mathbf{w}-\mathbf{u}\|^2 + \lambda\rho\Omega(\mathbf{w}),$$

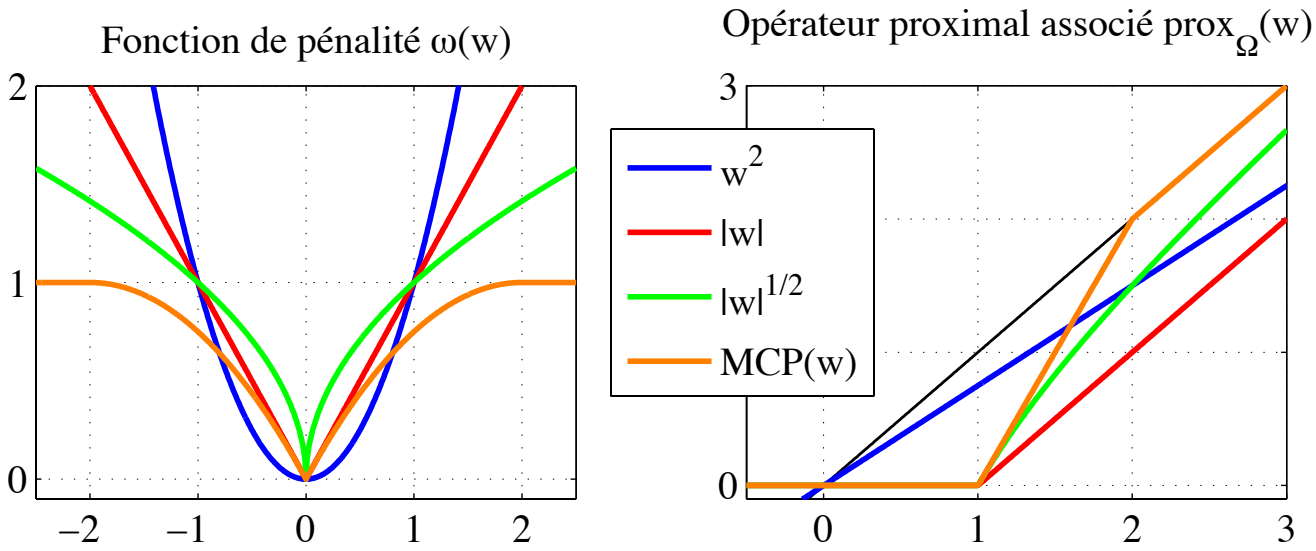
L'expression ci-dessus est associée à l'opérateur proximal :

Definition (Opérateur proximal)

L'opérateur proximal de la fonction Ω est :

$$\begin{aligned} \text{prox}_\Omega : \mathbb{R}^d &\longrightarrow \mathbb{R}^d \\ \mathbf{w} &\longmapsto \text{prox}_\Omega(\mathbf{w}) = \arg\min_{\mathbf{u}\in\mathbb{R}^d} \ \Omega(\mathbf{u}) + \tfrac{1}{2}\|\mathbf{u}-\mathbf{w}\|^2. \end{aligned}$$

$\Omega(\mathbf{w})$ est séparable



Example (Opérateurs proximaux utilisés en apprentissage)

$\Omega(\mathbf{w}) = 0$	$\text{prox}_\Omega(\mathbf{w}) = \mathbf{w}$	identité
$\Omega(\mathbf{w}) = \lambda\|\mathbf{w}\|_2^2$	$\text{prox}_\Omega(\mathbf{w}) = \frac{1}{1+\lambda}\mathbf{w}$	rétrécissement
$\Omega(\mathbf{w}) = \lambda\|\mathbf{w}\|_1$	$\text{prox}_\Omega(\mathbf{w}) = s_w \max(0, \lvert\mathbf{w}\rvert - \lambda)$	seuillage doux
$\Omega(\mathbf{w}) = \lambda\|\mathbf{w}\|_{1/2}^{1/2}$	[XCXZ12, Equation 11]	*bridge or power family*
$\Omega(\mathbf{w}) = \mathbb{I}_C(\mathbf{w})$	$\text{prox}_\Omega(\mathbf{w}) = \underset{\mathbf{u}\in C}{\arg\min}\ \frac{1}{2}\|\mathbf{u}-\mathbf{w}\|^2$	projection

Algorithme de descente de gradient proximal

Definition (Algorithme de descente de gradient proximal)

Un principe général pour résoudre le problème de MRE pénalisé consiste à mettre en œuvre les itérations suivantes

$$\mathbf{w}^{k+1} = \mathrm{prox}_{\rho^{(k)}\lambda\Omega}\left(\mathbf{w}^k - \rho^{(k)}\nabla L(\mathbf{w}^k)\right).$$

C'est l'algorithme de descente de gradient proximal, aussi connu sous le nom de *Forward Backward splitting* dans la communauté de traitement du signal.

- Si le pas ρ est bien choisi, il converge vers un minimum
- utiliser la règle de Barzilai-Borwein [WNF09].
- il est aussi possible d'accélérer l'algorithme [N$^+$07, BT09].
- code Matlab : `github.com/rflamary/nonconvex-optimization`

L'exemple du LASSO

$L(\mathbf{w}) = \frac{1}{2}\|\mathbf{X}\mathbf{w} - Y\|^2$ et $\Omega(\mathbf{w}) = \sum_{i=1}^{p} |w_i|$

$$\begin{aligned} \text{prox}_{\Omega}(\mathbf{w}) &= \underset{\mathbf{u}\in\mathbb{R}^n}{\arg\min} \sum_{i=1}^{p} |u_i| + \tfrac{1}{2}\|\mathbf{u} - \mathbf{w}\|^2 \\ &= \underset{\mathbf{u}\in\mathbb{R}^n}{\arg\min} \sum_{i=1}^{p} \big(|u_i| + \tfrac{1}{2}(u_i - w_i)^2\big). \end{aligned}$$

Calcul de sa sous différentielle

$$\partial(|u_i| + \tfrac{1}{2}(u_i - w_i)^2) = \begin{cases} sign(u_i) + u_i - w_i & \text{si } u_i \neq 0 \\ g + u_i - w_i \quad \text{avec } -1 \leq g \leq 1 & \text{si } u_i = 0, \end{cases}$$

de sorte que

$$0 \in \partial(|u_i| + \tfrac{1}{2}(u_i - w_i)^2) \quad \Leftrightarrow \quad u_i = \begin{cases} sign(u_i)(|w_i| - 1) & \text{si } |w_i| > 1 \\ 0 & \text{si } |w_i| \leq 1. \end{cases}$$

la méthode du gradient proximal pour le LASSO

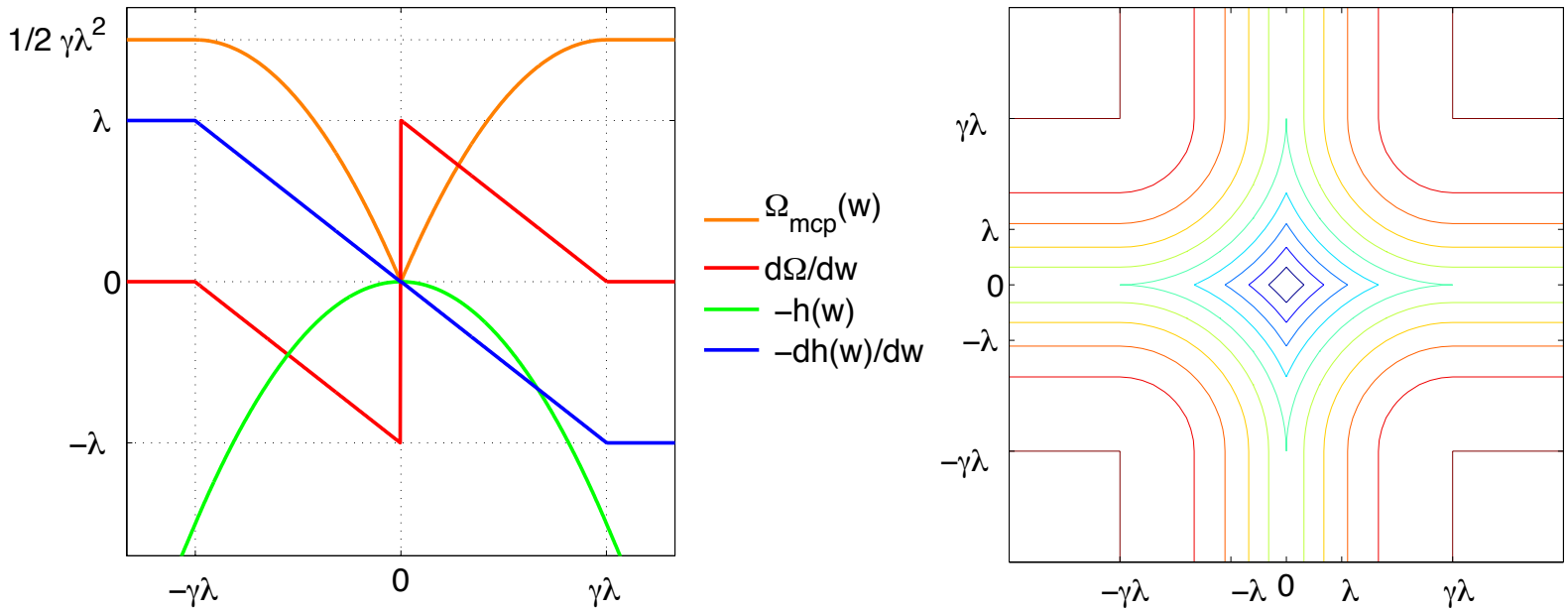
Algorithm 2 L'algorithme de gradient proximal pour le LASSO

1: **Entrées** : $\mathbf{X}, Y$ données d'apprentissage.
2: **Sortie** : $\mathbf{w}$ paramètres du modèle
3: $\rho \leftarrow 1/\|\mathbf{X}^\top \mathbf{X}\|$ initialisation du pas
4: **while** on n'a pas convergé **do**
5: $\quad \mathbf{w} \leftarrow \mathbf{w} - \rho \mathbf{X}^\top (\mathbf{X}\mathbf{w} - Y)$ pas de gradient (*forward*)
6: $\quad \mathbf{w} \leftarrow \mathit{sign}(\mathbf{w}) \max(0, |\mathbf{w}| - \rho\lambda)$ opérateur proximal (*backward*)
7: **end while**

Un exemple d'optimisation non convexe : la régression logistique parcimonieuse

La pénalité de type *minimax concave penalty* (MCP) pour $(\lambda \geq 0, \gamma \geq 1)$

$$\Omega_{\lambda,\gamma}(t) = \begin{cases} \lambda t - \dfrac{t^2}{2\gamma} & \text{si } t \leq \gamma\lambda \\ \dfrac{\gamma\lambda^2}{2} & \text{sinon.} \end{cases}$$



MCP avec le cout logistique J_ℓ

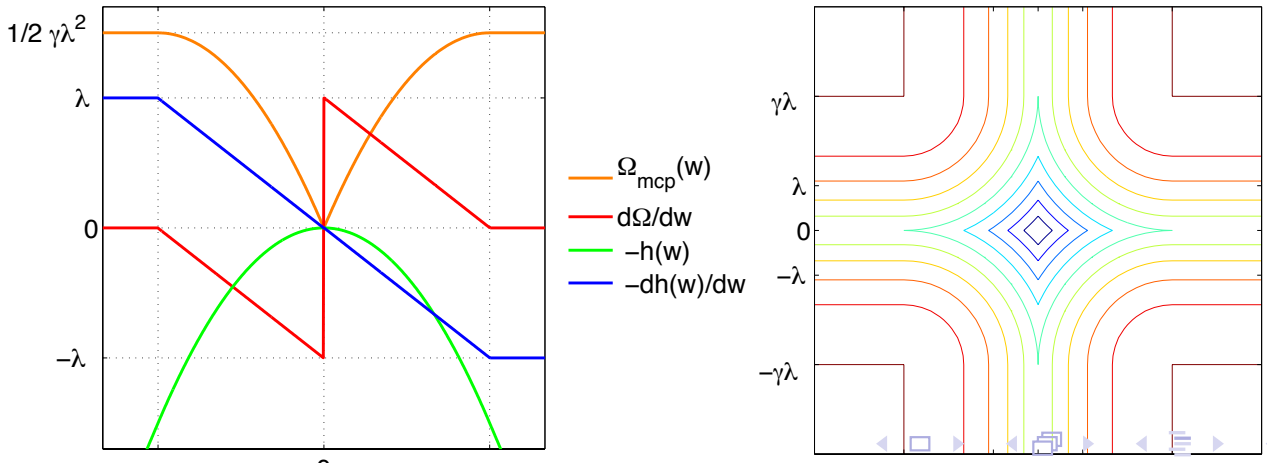
$$\min_{\mathbf{w}\in\mathbb{R}^p} \sum_{i=1}^{n}\Big(-Y_i(\mathbf{X}\mathbf{w})_i + \log\big(1+\exp(\mathbf{X}\mathbf{w})_i\big)\Big) + \sum_{i=1}^{p}\Omega_{\lambda,\gamma}(|w_i|), \quad (20)$$

décomposition de la pénalité MCP

$$\min_{\mathbf{w}\in\mathbb{R}^p} \underbrace{J_\ell(\mathbf{w}) - \lambda\sum_{j=1}^{p} h_{\lambda,\gamma}(|w_j|)}_{J_m(\mathbf{w})} + \lambda\|\mathbf{w}\|_1, \quad (21)$$

avec

$$h_{\lambda,\gamma}(t) = \left\{\frac{t^2}{2\gamma\lambda}\mathbb{I}_{\{t\le\gamma\lambda\}} + \left(t-\frac{\gamma\lambda}{2}\right)\mathbb{I}_{\{t>\gamma\lambda\}}\right\},$$

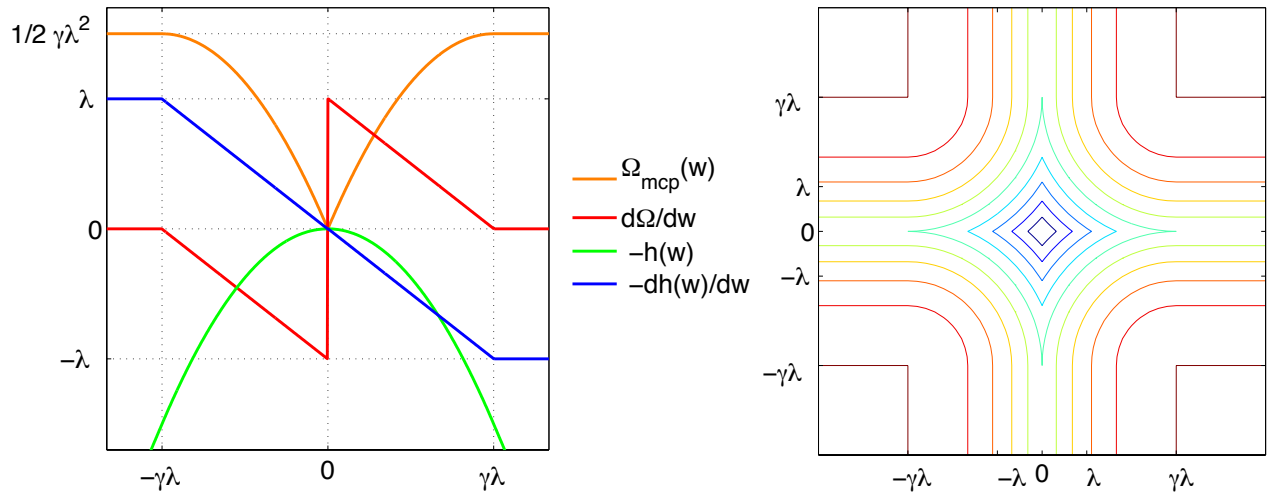


$J_m(\mathbf{w})$ est non convexe mais différentiable. Le gradient de J_m est

$$\nabla_\beta J_m(\mathbf{w}) = \mathbf{X}^\top(\mathbf{p} - Y) - \begin{cases} sign(w_j)\lambda & \text{si } |w_j| > \lambda\gamma \\ \dfrac{w_j}{\gamma} & \text{sinon.} \end{cases}$$

L'opérateur proximal de la norme ℓ_1 est

$$\text{prox}_{\ell_1}(u) \;=\; \begin{cases} 0 & \text{si } |u| \leq \lambda \\ sign(u)(|u| - \lambda) & \text{sinon.} \end{cases}$$



L'algorithme proximal ℓ_1 pour la régression logistique MCP

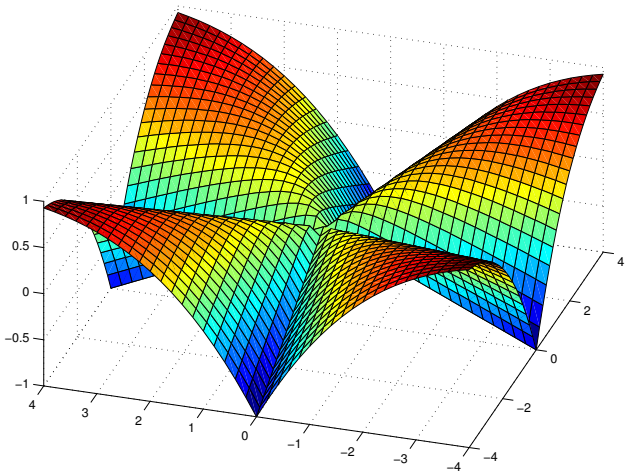
Algorithm 3 L'algorithme proximal ℓ_1 pour la régression logistique MCP

Entrées : $\mathbf{X}, Y$ données d'apprentissage.
Sortie : $\mathbf{w}$ paramètres du modèle
$\rho \leftarrow 1/\|\mathbf{X}^\top \mathbf{X}\|$ initialisation du pas
while on n'a pas convergé **do**
  $\mathbf{p} \leftarrow \dfrac{\exp^{\mathbf{Xw}}}{1+\exp^{\mathbf{Xw}}}$ division terme à terme
  $\mathbf{w} \leftarrow \mathbf{w} - \rho\big(\mathbf{X}^\top(\mathbf{p} - Y) - \mathit{sign}(\mathbf{w}) \min(\lambda, \frac{|\mathbf{w}|}{\gamma})\big)$
  $\mathbf{w} \leftarrow \mathit{sign}(\mathbf{w}) \max(0, |\mathbf{w}| - \rho\lambda)$ l'opérateur prox_{ℓ_1}
end while

Soulignons enfin que la décomposition (21) n'est pas unique et qu'il est donc possible de dériver d'autres algorithmes proximaux pour résoudre ce problème.

Plan

1. Exemples de problèmes d'optimisation en statistique
2. Optimisation différentiable, convexe et sans contraintes
3. Optimisation différentiable, convexe et avec contraintes
4. Optimisation sans contraintes non différentiable



Conclusion

- une certaine maturité de l'optimisation
- l'exemple du LASSO : la taille et la vitesse
  - LASSO avec CVX
  - le LASSO comme un QP
  - le LASSO à l'aide d'un algorithme proximal
- convexité et différentiabilité
- Le futur de l'optimisation
  - gros volumes de données : les méthodes stochastiques
  - non convexe, non différentiable

de 1996 à 2016

	facteur d'accélération
machine	$\times 2^{10} = 1000 - 1600$
solver	$\times 1000 - 3600$
formulation	???
global	$\times 1 - 5\ 10^6$

Bibliographie

[Ber99] Dimitri P Bertsekas. Nonlinear programming. 1999.

[Bot10] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010.

[BT09] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

[GB14] Michael Grant and Stephen Boyd. CVX: Matlab software for disciplined convex programming, version 2.1. `http://cvxr.com/cvx`, March 2014.

[HU02] Jean-Baptiste Hiriart-Urruty. L'optimisation: deux ou trois choses que je sais d'elle. *Note Technique- Centre national d'études spatiales*, 2002.

[N+07] Yurii Nesterov et al. Gradient methods for minimizing composite objective function. Technical report, UCL, 2007.

[NW06] Jorge Nocedal and Stephen Wright. *Numerical optimization*. Springer Science & Business Media, 2006.

[WNF09] Stephen J Wright, Robert D Nowak, and Mário AT Figueiredo. Sparse reconstruction by separable approximation. *Signal Processing, IEEE Transactions on*, 57(7):2479–2493, 2009.

[XCXZ12] Zongben Xu, Xiangyu Chang, Fengmin Xu, and Hai Zhang. $L_{1/2}$ regularization: a thresholding representation theory and a fast solver. *Neural Networks and Learning Systems, IEEE Transactions on*, 23(7):1013–1027, 2012.