Le Lasso

Stéphane Canu
stephane.canu@insa-rouen.fr

Advanced Machine Learning

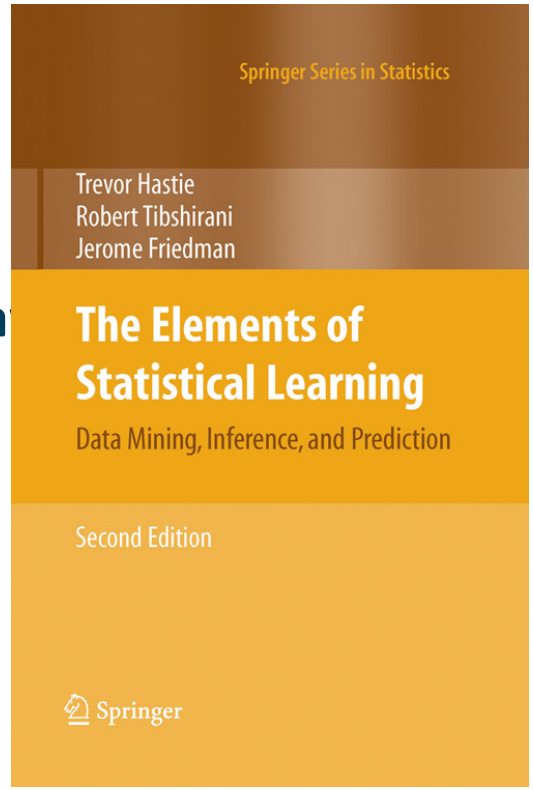
Winter 2023-2024

Outline

1. **Introduction**
   - Existence of a solution
2. **Gradient, subgradient and subdifferentials**
3. **Le Lasso comme un problème d'optimisation avec contraintes**
   - Constraints
     - Equality constraints
     - Inequality constraints
4. **Conclusions**

Chapter 3: Linear Methods for Regression, page 43

1. **Introduction**
   - Existence of a solution
2. **Gradient, subgradient and subdifferentials**
3. **Le Lasso comme un problème d'optimisation a**
   - Constraints
     - Equality constraints
     - Inequality constraints
4. **Conclusions**



Le Lasso

$$\min_{\beta \in \mathbb{R}^p} J_\lambda(\beta) \quad \text{avec} \quad J_\lambda(\beta) = \tfrac{1}{2}\|X\beta - y\|^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

Outline

1. **Introduction**
   - Existence of a solution
2. **Gradient, subgradient and subdifferentials**
3. **Le Lasso comme un problème d'optimisation avec contraintes**
   - Constraints
     - Equality constraints
     - Inequality constraints
4. **Conclusions**

Existence of a solution

Definition (local minima)

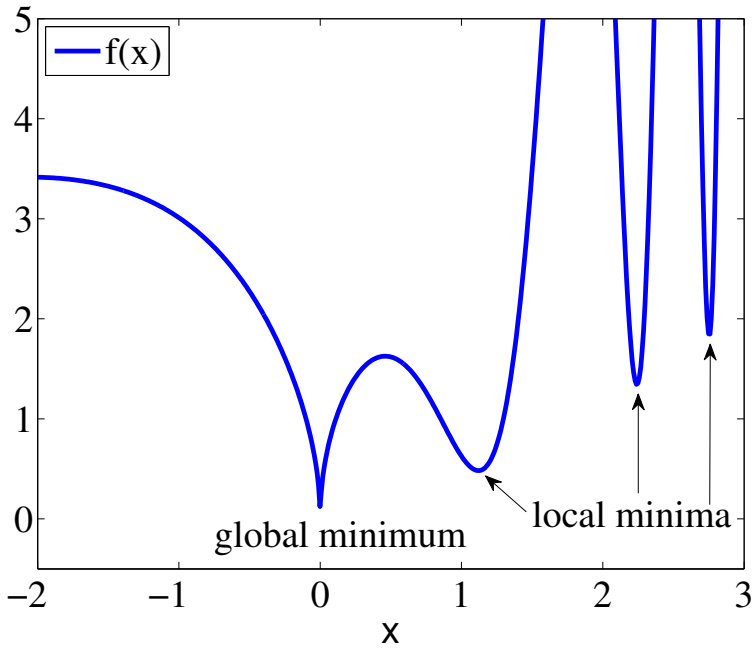
$\mathrm{x}^\star$ is a local minima for J if there exists $\varepsilon > 0$ such that

$$J(\mathrm{x}^\star) \leq J(\mathrm{x}) \qquad \forall \mathrm{x} \text{ with } \|\mathrm{x} - \mathrm{x}^\star\| \leq \varepsilon$$

Definition (global minimum)

$\mathrm{x}^\star$ is a global minimum for f if

$$J(\mathrm{x}^\star) \leq J(\mathrm{x}) \qquad \forall \mathrm{x} \in \mathbb{R}^p$$



The question of the existence of a global minima to this unconstrained minimization problem require some definition.

Definition (l.s.c.)

a real valued function f is lower semi-continuous (l.s.c.) at some point x if for every sequence $\{\mathrm{x}_k, k \in \mathbb{N}\}$ that converges to x

$$J(\mathrm{x}) \leq \liminf_{k \to \infty} J(\mathrm{x}_k).$$

The 0/1 loss function, the counting function and the rank function are l.s.c.

Definition (coercive)

a real valued function J is coercive if for every sequence $\{\mathrm{x}_k, k \in \mathbb{N}\}$ such that $\lim_{k \to \infty} \|\mathrm{x}_k\| = \infty$

$$\lim_{k \to \infty} J(\mathrm{x}_k) = \infty.$$

Proposition

Weierstrass' theorem: existence of a solution (sufficient condition). If J is l.s.c. and coercive or admits a non empty bounded level set, then there exists a global minimizer for J.

Outline

1. Introduction
   - Existence of a solution
2. **Gradient, subgradient and subdifferentials**
3. Le Lasso comme un problème d'optimisation avec contraintes
   - Constraints
     - Equality constraints
     - Inequality constraints
4. Conclusions

Gradient

The gradient is a generalization of the usual concept of derivative of a function in one dimension to a function in several dimensions.

$$\begin{array}{rccc} J: & \mathbb{R}^p & \longrightarrow & \mathbb{R} \\ & \mathrm{x} & \longmapsto & J(\mathrm{x}) \end{array}$$

Assume J is differentiable in the sense that all its partial derivatives $\frac{\partial J}{\partial w_i}$ exists.

Definition (gradient)

The gradient $\nabla J(\mathrm{x})$ of a function J at point x is the vector whose components are the partial derivatives of J

Example

The least square

$$J_1(\mathrm{x}) = \|A\mathrm{x} - y\|^2 \qquad \nabla J_1(x) = 2A^t(A\mathrm{x} - y)$$

Comment calculer un gradient ?

Connaitre ses formules de dérivation
Passer par le calcul de la dérivée directionnelle

Definition (dérivée directionnelle)

On appelle dérivée directionnelle de J au point x et dans la direction $\mathrm{d} \in \mathbb{R}^n$ la limite :

$$D_{\mathrm{x}} J(\mathrm{x}, \mathrm{d}) = \lim_{\varepsilon \to 0} \frac{J(\mathrm{x} + \varepsilon \mathrm{d}) - J(\mathrm{x})}{\varepsilon}$$

si elle existe

une manière commode de calculer cette dérivée est d'utiliser la définition suivante : $\varphi(\varepsilon) = J(\mathrm{x} + \varepsilon \mathrm{d})$

Theorem (Calcul pratique de la dérivée)

soit J une fonction de $\mathbb{R}^n$ à valeur dans $\mathbb{R}$:

$$D_{\mathrm{x}} J(\mathrm{x}, \mathrm{d}) = \left. \frac{dJ(\mathrm{x} + \varepsilon \mathrm{d})}{d\varepsilon} \right|_{\varepsilon=0} = \varphi'(0)$$

Démonstration :

$$\begin{aligned} \varphi'(0) &= \lim_{\epsilon \to 0} \frac{\varphi(\epsilon) - \varphi(0)}{\epsilon} \\ &= \lim_{\epsilon \to 0} \frac{J(\mathrm{x} + \epsilon \mathrm{d}) - J(\mathrm{x})}{\epsilon} \quad = D_{\mathrm{x}} J(\mathrm{x}, \mathrm{d}) \end{aligned}$$

recette :

1. calculer $J(\mathrm{x} + \varepsilon \mathrm{d})$
2. calculer la dérivée par rapport à ε
3. prendre la valeur de la dérivée en $\varepsilon = 0$

Exemples de dérivées directionnelles

Exemples

$$J_4(\mathrm{x}) = \|\mathrm{x}\|^2 \qquad D_{\mathrm{x}} J_4(\mathrm{x}, \mathrm{d}) = 2\mathrm{x}^\top \mathrm{d}$$

$$J_5(\mathrm{x}) = \|A\mathrm{x} - b\|^2 \qquad D_{\mathrm{x}} J_5(\mathrm{x}, \mathrm{d}) = 2(A\mathrm{x} - b)^\top A\mathrm{d}$$

$$J_6(\mathrm{x}) = \|\mathrm{x}\|_2 \qquad D_{\mathrm{x}} J_6(\mathrm{x}, \mathrm{d}) = \frac{\mathrm{x}^\top \mathrm{d}}{\|\mathrm{x}\|}$$

$$\begin{aligned} \varphi_4(\varepsilon) &= \|\mathrm{x} + \varepsilon \mathrm{d}\|^2 \\ &= \|\mathrm{x}\|^2 + 2\varepsilon \mathrm{x}^\top \mathrm{d} + \|\varepsilon \mathrm{d}\|^2 \\ &= \underbrace{a}_{\|\mathrm{x}\|^2} + \underbrace{b}_{2\mathrm{x}^\top \mathrm{d}} \varepsilon + \underbrace{c}_{\|\mathrm{d}\|^2} \varepsilon^2 \end{aligned}$$

$$\varphi_4'(\varepsilon) = b + 2c\varepsilon \quad \Rightarrow \quad \varphi_4'(0) = b \quad \Leftrightarrow \quad D_{\mathrm{x}} J_4(\mathrm{x}, \mathrm{d}) = 2\mathrm{x}^\top \mathrm{d}$$

Calculez

$D_{\mathrm{x}} J_5(\mathrm{x}, \mathrm{d})$ (et $D_{\mathrm{x}} J_5(\mathrm{x}, \mathrm{d})$ chez vous)

Dérivée directionnelle et gradient

Theorem (p=1)

Si, pour x *fixé, l'application* $\mathrm{d} \to D_{\mathrm{x}}J(\mathrm{x},\mathrm{d})$ *est linéaire continue,*

$$D_{\mathrm{x}}J(\mathrm{x},\mathrm{d}) = \nabla J_{\mathrm{x}}(\mathrm{x})^{\top}\mathrm{d}$$

Calcul pratique du gradient :

1. calculer $J(\mathrm{x}+\varepsilon\mathrm{d})$
2. calculer la dérivée par rapport à ε
3. prendre la valeur de la dérivée en $\varepsilon = 0$
4. écrire $\varphi'(0)$ sous la forme $\nabla J_{\mathrm{x}}(\mathrm{x})^{\top}\mathrm{d}$ et identifier le gradient

Exemple de calcul de dérivée directionnelle

$$
\begin{aligned}
J(\mathrm{x}) &= \sum_{i=1}^{n} y_i A_i^\top \mathrm{x} - \sum_{i=1}^{n} \log\left(1 + \exp^{A_i^\top \mathrm{x}}\right) \\
&= \mathrm{y}^\top A\mathrm{x} - \mathrm{e}^\top \log\left(\mathbb{1} + \exp^{A\mathrm{x}}\right) \\
J(\mathrm{x} + \varepsilon \mathrm{d}) &= \mathrm{y}^\top A(\mathrm{x} + \varepsilon \mathrm{d}) - \mathrm{e}^\top \log\left(\mathbb{1} + \exp^{A(\mathrm{x}+\varepsilon \mathrm{d})}\right) \\
&= \mathrm{y}^\top A\mathrm{x} + \varepsilon \mathrm{y}^\top A\mathrm{d} - \mathrm{e}^\top \log\left(\mathbb{1} + \exp^{A\mathrm{x}} \exp^{\varepsilon A\mathrm{d}}\right)
\end{aligned}
$$

pour x et d fixés :

$$
\begin{aligned}
\varphi(\varepsilon) &= \mathrm{y}^\top A\mathrm{x} + \varepsilon \mathrm{y}^\top A\mathrm{d} - \mathrm{e}^\top \log\left(\mathbb{1} + \exp^{A\mathrm{x}} \exp^{\varepsilon A\mathrm{d}}\right) \\
&= \mathrm{y}^\top A\mathrm{x} + \varepsilon \mathrm{y}^\top A\mathrm{d} - \sum_{i=1}^{n} \log\left(1 + \exp^{A_i^\top \mathrm{x}} \exp^{\varepsilon A_i^\top \mathrm{d}}\right) \\
\varphi'(\varepsilon) &= \mathrm{y}^\top A\mathrm{d} - \sum_{i=1}^{n} \frac{\exp^{A_i^\top \mathrm{x}} \exp^{\varepsilon A_i^\top \mathrm{d}} A_i^\top \mathrm{d}}{1 + \exp^{A_i^\top \mathrm{x}} \exp^{\varepsilon A_i^\top \mathrm{d}}} \\
\varphi'(0) &= \mathrm{y}^\top A\mathrm{d} - \sum_{i=1}^{n} \frac{\exp^{A_i^\top \mathrm{x}} A_i^\top \mathrm{d}}{1 + \exp^{A_i^\top \mathrm{x}}} \\
&= \mathrm{y}^\top A\mathrm{d} - \frac{\exp^{A\mathrm{x}}}{\mathbb{1} + \exp^{A\mathrm{x}}}^\top A\mathrm{d} \quad = \underbrace{\left(A^\top \mathrm{y} - A^\top \frac{\exp^{A\mathrm{x}}}{\mathbb{1} + \exp^{A\mathrm{x}}}\right)^\top \mathrm{d}}_{\nabla_{\mathrm{x}} J(\mathrm{x})} \\
\nabla_{\mathrm{x}} J(\mathrm{x}) &= A^\top(\mathrm{y} - \mathbf{p}) \qquad \text{avec} \quad p_i = \frac{\exp^{A_i^\top \mathrm{x}}}{1 + \exp^{A_i^\top \mathrm{x}}}
\end{aligned}
$$

Règles de calcul pour la dérivée

c'est un opérateur linéaire :

$$\nabla(J_1 + \alpha J_2)_{\mathrm{x}}(\mathrm{x}) = \nabla_{\mathrm{x}} J_1(\mathrm{x}) + \alpha \nabla_{\mathrm{x}} J_2(\mathrm{x})$$

Combinaison d'applications (chain rule) :

exemples

$$J_\alpha(\mathrm{x}) = \tfrac{1}{2}\mathrm{x}^\top A\mathrm{x} - \mathrm{x}^\top \mathrm{b}$$

$$J_\beta(\mathrm{x}) = \mathbb{1}^\top \log(A\mathrm{x})$$

$$J_\gamma(\mathrm{x}) = f(\mathrm{a}^\top \mathrm{x})$$

$$\nabla_{\mathrm{x}} J_\gamma(\mathrm{x}) = f'(\mathrm{a}^\top \mathrm{x})\mathrm{a}$$

Développement limité au premier ordre

Propriété : développement au premier ordre

si $\|\mathrm{d}\| = 1$ et $\varepsilon > 0$,

$$J(\mathrm{x} + \varepsilon \mathrm{d}) = J(\mathrm{x}) + \varepsilon \underbrace{\nabla_x J(x)^\top \mathrm{d}}_{DJ_{\mathrm{x}}(\mathrm{x},\mathrm{d})} + o(\varepsilon)$$

Application dans le cas p=1 :

si on cherche une direction de descente d (qui fait diminuer J)

$$J(\mathrm{x} + \varepsilon \mathrm{d}) < J(\mathrm{x})$$

On peut choisir $\mathrm{d} = -\nabla_x J(x)$ (ce n'est pas la seule solution)

Démonstration

$$J(\mathrm{x} + \varepsilon \mathrm{d}) = J(\mathrm{x} - \varepsilon \nabla_x J(\mathrm{x})) = J(\mathrm{x}) - \varepsilon \nabla_x J(x)^\top \nabla_x J(x) + o(\varepsilon) = J(\mathrm{x}) - \underbrace{\varepsilon \|\nabla_x J(x)\|^2}_{>0} + o(\varepsilon)$$

Dérivée seconde

Definition (Matrice Hessienne)

C'est la matrice des dérivée seconde de l fonctionnelle J, si elle existe :

$$H_J(\mathrm{x}) = \begin{pmatrix} \frac{\partial^2 J(\mathrm{x})}{\partial x_1^2} & \frac{\partial^2 J(\mathrm{x})}{\partial x_2 \partial x_1} & \cdots & \frac{\partial^2 J(\mathrm{x})}{\partial x_j \partial x_1} & \cdots & \frac{\partial^2 J(\mathrm{x})}{\partial x_n \partial x_1} \\ \frac{\partial J(\mathrm{x})}{\partial x_2 \partial x_1} & \frac{\partial^2 J(\mathrm{x})}{\partial x_2^2} & \cdots & \frac{\partial^2 J(\mathrm{x})}{\partial x_j \partial x_2} & \cdots & \frac{\partial^2 J(\mathrm{x})}{\partial x_n \partial x_2} \\ \vdots & \ddots & & & & \\ \frac{\partial^2 J(\mathrm{x})}{\partial x_i \partial x_1} & \cdots & \cdots & \frac{\partial^2 J(\mathrm{x})}{\partial x_j \partial x_i} & \cdots & \frac{\partial^2 J(\mathrm{x})}{\partial x_n \partial x_i} \\ \vdots & \ddots & & & & \\ \frac{\partial^2 J(\mathrm{x})}{\partial x_n \partial x_1} & \cdots & \cdots & \frac{\partial^2 J(\mathrm{x})}{\partial x_j \partial x_n} & \cdots & \frac{\partial^2 J(\mathrm{x})}{\partial x_n \partial x_n} \end{pmatrix}$$

Exemples de dérivée seconde

Exemples

$J_2(\mathrm{x}) = A\mathrm{x} - b$	$\nabla_{\mathrm{x}} J(\mathrm{x}) = A$	$H_{\mathrm{x}}(\mathrm{x}) = 0$
$J_4(\mathrm{x}) = \|\mathrm{x}\|^2$	$\nabla_{\mathrm{x}} J(\mathrm{x}) = 2\mathrm{x}$	$H_{\mathrm{x}}(\mathrm{x}) = 2I$
$J_5(\mathrm{x}) = \|A\mathrm{x} - b\|^2$	$\nabla_{\mathrm{x}} J(\mathrm{x}) = 2A^\top(A\mathrm{x} - b)$	$H_{\mathrm{x}}(\mathrm{x}) = 2A^\top A$
$J_6(\mathrm{x}) = \|\mathrm{x}\|_2$	$\nabla_{\mathrm{x}} J(\mathrm{x}) = \frac{\mathrm{x}}{\|\mathrm{x}\|}$	$H_{\mathrm{x}}(\mathrm{x}) = \frac{1}{\|\mathrm{x}\|^2} M(\mathrm{x})$
$J_8(f) = \int_0^1 f(t)^2 dt$	$DJ_f(f) = f$	$H_f(f) = I$

Développement au second ordre

Soit $J : \mathbb{R}^n \longrightarrow \mathbb{R}$

Propriété : développement de Taylor au second ordre

$$J(\mathrm{x}+\mathrm{d}) = J(\mathrm{x}) + \nabla_{\mathrm{x}} J(\mathrm{x})\mathrm{d} + \tfrac{1}{2}\mathrm{d}^\top H_{\mathrm{x}}(\mathrm{x})\mathrm{d} + o(\|\mathrm{d}\|^2)$$

Application : si au point x on a $\nabla_{\mathrm{x}} J(\mathrm{x}) = 0$ et $H_{\mathrm{x}} J(\mathrm{x})$ définie positive, alors x est un minimum local de J.

ce résultat peut être généralisé

Exemple

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

$$\nabla_{\mathrm{x}} J(\mathrm{x}) = A\mathrm{x} - \mathrm{b} \qquad\qquad H_{\mathrm{x}}(\mathrm{x}) = A$$

Si A est définie positive la solution du problème d'optimisation est $x^\star = A^{-1}\mathrm{b}$ c'est un minimum global

Si A n'est pas définie positive...

Fonction convexe

Definition (fonction convexe)

une fonction J est dite convexe si, pour tout $\mathrm{x}, \mathrm{y} \in \Omega$ on a

$$\forall \alpha \in]0,1[, \quad J(\alpha \mathrm{x} + (1-\alpha)\mathrm{y}) \leq \alpha J(\mathrm{x}) + (1-\alpha) J(\mathrm{y})$$

si J est une fonction convexe, la fonction $-J$ est dite concave.

exemple

$$J_1(\mathrm{x}) = \mathrm{a}^\top \mathrm{x} \qquad J_2(\mathrm{x}) = \tfrac{1}{2}\mathrm{x}^\top A \mathrm{x}$$

$$J_3(\mathrm{x}) = \sum_{i=1}^{d} \exp^{-x_i} \qquad J_4(\mathrm{x}) = -\log \mathrm{a}^\top \mathrm{x}$$

On vérifiera que toute somme de fonctions convexe est convexe.

Theorem

si x_0 est une solution locale d'un problème d'optimisation convexe, alors c'est aussi la solution globale.

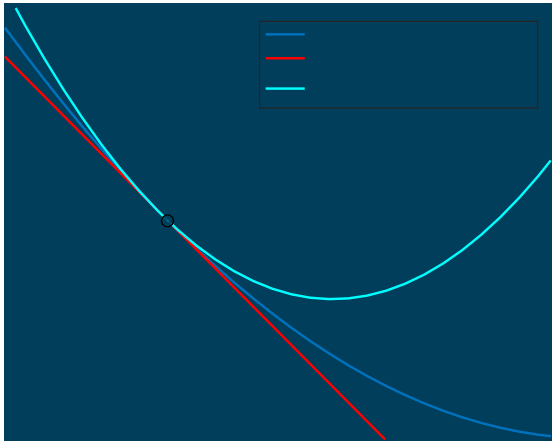
Caractérisation d'une fonction convexe

Si J est convexe, sa matrice hessiènne $H(x) \succeq 0$ est définie positive

Exemple

$$J(\mathrm{x}) = \tfrac{1}{2}\|A\mathrm{x} - b\|^2 \qquad\qquad H = A^\top A$$

Plus fort que convexe : J convexe et Gradient Lipschitz : $LId - H \succeq 0$ est positive



$$\nabla_x J(\mathrm{x})^\top(\mathrm{y}-\mathrm{x}) \;\leq\; J(\mathrm{y}) - J(\mathrm{x}) \;\leq\; \nabla_x J(\mathrm{x})^\top(\mathrm{y}-\mathrm{x}) + \tfrac{L}{2}\|\mathrm{y}-\mathrm{x}\|^2$$

Preuve

si J est convexe

$$\begin{aligned} J\Big(\alpha y + (1-\alpha)x\Big) &\leq \alpha J(y) + (1-\alpha)J(x) \\ J\Big(x + \alpha(y-x)\Big) &\leq \alpha\Big(J(y) - J(x)\Big) + J(x) \end{aligned}$$

et donc pour tout $\alpha > 0$

$$\frac{J\Big(x + \alpha(y-x)\Big) - J(x)}{\alpha} \leq J(y) - J(x)$$

soit en passant à la limite, par définition de la dérivée directionnelle et du gradient associé quant il existe, on a :

$$\nabla_x J(\mathrm{x})^\top (\mathrm{y} - \mathrm{x}) \leq J(y) - J(x)$$

si $LId - H \succeq 0$ est positive, alors la fonction

$$K(x) = \tfrac{L}{2} x^\top x - J(x)$$

est convexe. Donc

$$\begin{aligned} \nabla_x K(\mathrm{x})^\top (\mathrm{y} - \mathrm{x}) &\leq K(y) - K(x) \\ \Big(Lx - \nabla_x J(\mathrm{x})\Big)^\top (\mathrm{y} - \mathrm{x}) &\leq \tfrac{L}{2} y^\top y - \tfrac{L}{2} x^\top x - J(y) + J(x) \end{aligned}$$

d'où

$$\nabla_x J(\mathrm{x})^\top (\mathrm{y} - \mathrm{x}) - Lx^\top (\mathrm{y} - \mathrm{x}) + \tfrac{L}{2} y^\top y - \tfrac{L}{2} x^\top x \geq J(y) - J(x)$$

Sous-gradient

Lié aux fonctions convexes

Definition (Sous-gradient)

On appelle sous gradient de J au point x_0 tout vecteur $\mathbf{g} \in \mathbb{R}^n$ vérifiant

$$\forall \mathrm{x} \in \mathcal{V}(x_0), \qquad J(\mathrm{x}) \geq J(\mathrm{x}_0) + \mathbf{g}^\top (\mathrm{x} - \mathrm{x}_0)$$

Definition (sous-différentielle)

On appelle sous différentielle de J au point x_0 l'ensemble de tous les sous-gradient de J au point x_0. On le note $\partial J(\mathrm{x}_0)$

exemple

$$J_5(\mathrm{x}) = \sum_{i=1}^{d} |x_i| \qquad J_6(\mathrm{x}) = \sum_{i=1}^{d} x_i \mathbb{1}_{x_i<0}$$

Exemple de sous différentielle

$$J(x) = |x|$$

un sous gradient est un réel g tel que

$$\forall \mathrm{x} \in \mathcal{V}(x_0), \qquad J(\mathrm{x}) \geq J(\mathrm{x}_0) + \mathbf{g}^\top(\mathrm{x} - \mathrm{x}_0)$$

soit :

si $x_0 > 0$, alors $J(x) = x = \mathrm{x}_0 + (\mathrm{x} - \mathrm{x}_0)$ et g $= 1$

si $x_0 < 0$, alors $J(x) = -x = -x_0 + -(\mathrm{x} - \mathrm{x}_0)$ et g $=$ -1

si $x_0 = 0$ alors :

pour $x > 0$, on a $J(x) = x \geq 0 + g(x - 0)$ pour tout $g \leq 1$

pour $x < 0$, on a $J(x) = -x \geq 0 + g(x - 0)$ pour tout $g \geq -1$

tout $g \in [-1, 1]$ est un sous gradient

Et donc la sous différentielle de J en x_0 est :

$$\partial J(\mathrm{x}_0) = \begin{cases} 1 & \text{si } x_0 > 0 \\ [-1, 1] & \text{si } x_0 = 0 \\ -1 & \text{si } x_0 < 0 \end{cases}$$

La sous différentielle du Lasso

$$\min_{\beta \in \mathbb{R}^p} J_\lambda(\beta) \quad \text{avec} \quad J_\lambda(\beta) = \frac{1}{2}\|X\beta - y\|^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

$$g(\beta) = X^\top (X\beta - y)$$

$$\partial J(\beta) = g(\beta) + \lambda s(\beta) \qquad \text{avec } s_j(\beta) = \begin{cases} 1 & \text{si } \beta_j > 0 \\ [-1, 1] & \text{si } \beta_j = 0 \\ -1 & \text{si } \beta_j < 0 \end{cases}$$

β optimal si $0 \in \partial J(\beta)$ c'est-à-dire si

$$\begin{cases} g_j(\beta) + \lambda = 0 & \text{si } \beta_j > 0 \\ -\lambda \leq g \leq \lambda & \text{si } \beta_j = 0 \\ g_j(\beta) - \lambda = 0 & \text{si } \beta_j < 0 \end{cases}$$

en effet si $-\lambda < g < \lambda$ cela signifie que $g - \lambda < 0$ et $0 < g + \lambda$ et donc il existe bien un α entre -1 et 1 tel que $g + \alpha\lambda = 0$

Unconstrained optimization (Fermat's rule)

For J convex & differentiable, global minima

$$\mathrm{x}^\star = \operatorname*{argmin}_{\mathrm{x}} J(\mathrm{x}) \quad \Longleftrightarrow \quad \nabla J(\mathrm{x}^\star) = 0$$

For J convex & nondifferentiable, global minima

$$\mathrm{x}^\star = \operatorname*{argmin}_{\mathrm{x}} J(\mathrm{x}) \quad \Longleftrightarrow \quad 0 \in \partial J(\mathrm{x}^\star)$$

For J non convex & nondifferentiable, local minima, necessary condition

$$\mathrm{x}^\star = \operatorname*{argmin}_{\mathrm{x} \in V(\mathrm{x}^\star)} J(\mathrm{x}) \quad \Longrightarrow \quad 0 \in \partial_c J(\mathrm{x}^\star)$$

or equivalently (sometimes referred as Oresme's rule)

$$\mathrm{x}^\star = \operatorname*{argmin}_{\mathrm{x} \in V(\mathrm{x}^\star)} J(\mathrm{x}) \quad \Longrightarrow \quad D_c J(\mathrm{x}^\star, \mathrm{x}) \geq 0, \quad \forall \mathrm{x} \in V(\mathrm{x}^\star)$$

Fermat's rule and critical points

Table: Fermat's rule and critical (or stationary) points

	differentiable	non differentiable
cvx	$\mathrm{x}^\star = \underset{\mathrm{x}}{\operatorname{argmin}}\, J(\mathrm{x}) \Leftrightarrow \nabla J(\mathrm{x}^\star) = 0$	$\mathrm{x}^\star = \underset{\mathrm{x}}{\operatorname{argmin}}\, J(\mathrm{x}) \Leftrightarrow 0 \in \partial J(\mathrm{x}^\star)$
non cvx	$\mathrm{x}^\star = \underset{\mathrm{x}\in V(\mathrm{x}_0)}{\operatorname{argmin}}\, J(\mathrm{x}) \Rightarrow \nabla J(\mathrm{x}^\star) = 0$	$\mathrm{x}^\star = \underset{\mathrm{x}\in V(\mathrm{x}_0)}{\operatorname{argmin}}\, J(\mathrm{x}) \Rightarrow 0 \in \partial_c J(\mathrm{x}^\star)$

Definition (Clarke critical (or stationary) point)

Let f be a locally Lipschitz continuous function at x. A point $\mathrm{x}^\star$ is called a (Clarke) critical point if

$$0 \in \partial_c J(\mathrm{x})$$

Local minima are critical points but the converse is not always true.

Outline

1. Introduction
   - Existence of a solution
2. Gradient, subgradient and subdifferentials
3. **Le Lasso comme un problème d'optimisation avec contraintes**
   - Constraints
     - Equality constraints
     - Inequality constraints
4. Conclusions

Le problème

$$\begin{cases} \min\limits_{\beta \in \mathbb{R}^p} & \frac{1}{2}\|X\beta - y\|^2 \\ \text{avec} & \sum\limits_{j=1}^{p} |\beta_j| \leq t \end{cases} \tag{1}$$

La lagangien

$$\mathcal{L}(\lambda, \beta) = \frac{1}{2}\|X\beta - y\|^2 + \lambda \left(\sum_{j=1}^{p} |\beta_j| - t \right)$$

Remarque : quand λ est fixé , $J_\lambda(\beta) = \mathcal{L}(\lambda, \beta)$

les KKT

$$\begin{aligned} \text{stationarity} \quad & X^\top (X\beta - y) = 0 \\ \text{primal admissibility} \quad & \sum_{j=1}^{p} |\beta_j| \leq t \\ \text{dual admissibility} \quad & \lambda \geq 0 \\ \text{complementarity} \quad & \lambda(\textstyle\sum_{j=1}^{p} |\beta_j| - t) = 0 \end{aligned}$$

Outline

1 Introduction
- Existence of a solution

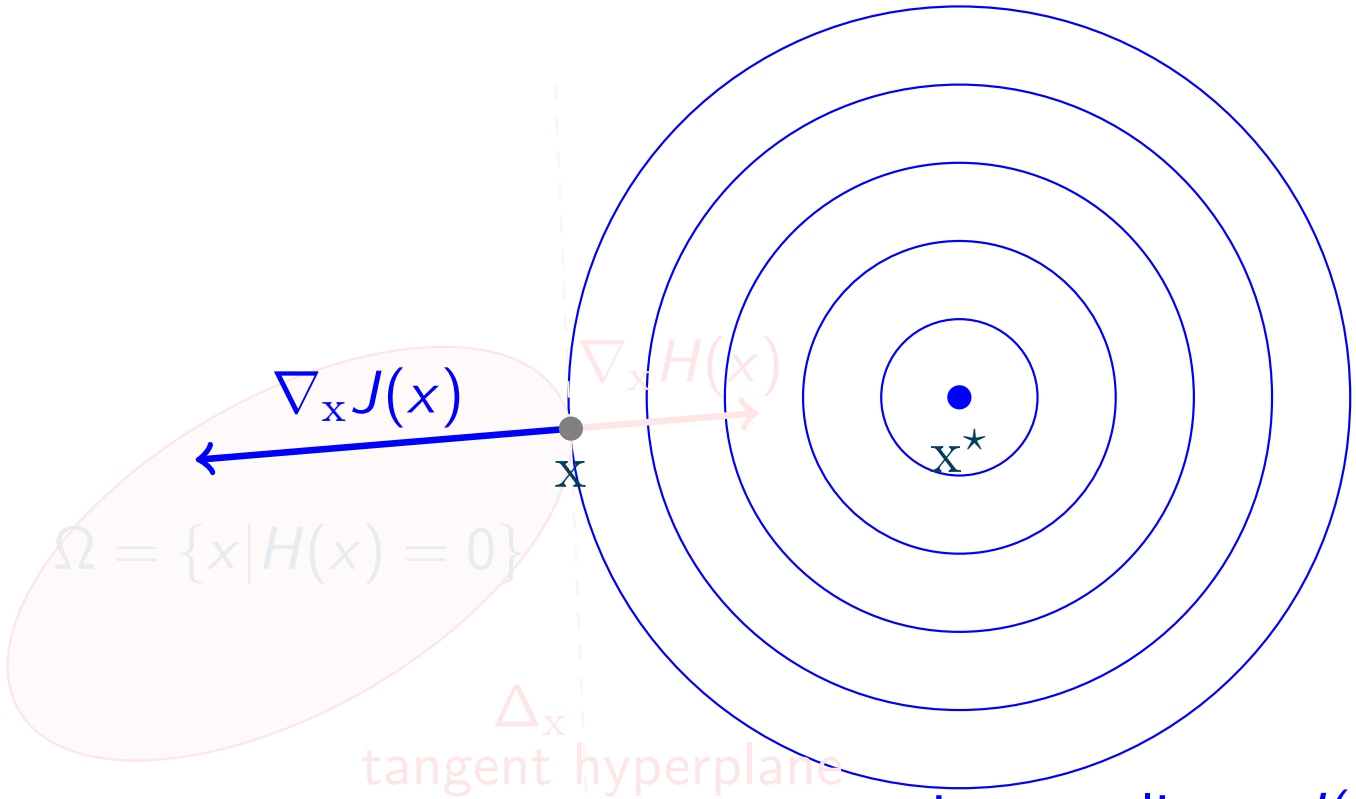
2 Gradient, subgradient and subdifferentials

3 **Le Lasso comme un problème d'optimisation avec contraintes**
- Constraints
  - Equality constraints
  - Inequality constraints

4 Conclusions

A simple example (to begin with)

$$\begin{cases} \min\limits_{x_1,x_2} & J(x) = (x_1 - a)^2 + (x_2 - b)^2 \\ \text{with} & H(x) = \alpha(x_1 - c)^2 + \beta(x_2 - d)^2 + \gamma x_1 x_2 - 1 \end{cases}$$

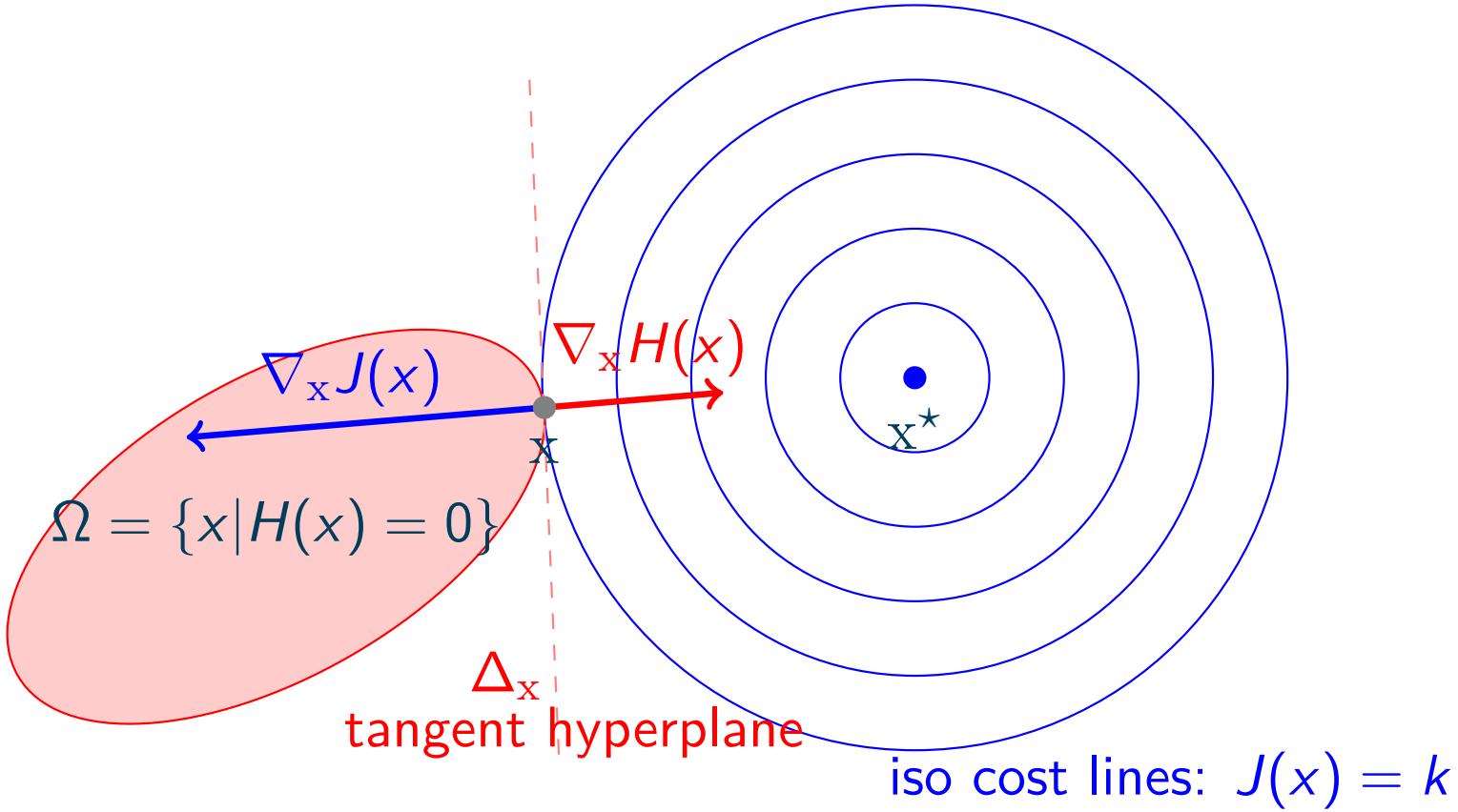


iso cost lines: $J(x) = k$

$$\nabla_x H(x) = \lambda \nabla_x J(x)$$

A simple example (to begin with)

$$\begin{cases} \min\limits_{x_1,x_2} & J(\mathrm{x}) = (x_1 - a)^2 + (x_2 - b)^2 \\ \text{with} & H(\mathrm{x}) = \alpha(x_1 - c)^2 + \beta(x_2 - d)^2 + \gamma x_1 x_2 - 1 \end{cases}$$



$$\nabla_x H(x) = \lambda \, \nabla_x J(x)$$

The only one equality constraint case

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

Loss J : d is a descent direction if it exists $\varepsilon_0 \in \mathbb{R}$ such that $\forall \varepsilon \in \mathbb{R},\ 0 < \varepsilon \leq \varepsilon_0$

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

constraint H : d is a feasible descent direction if it exists $\varepsilon_0 \in \mathbb{R}$ such that $\forall \varepsilon \in \mathbb{R},\ 0 < \varepsilon \leq \varepsilon_0$

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

If at $x^\star$, vectors $\nabla_{\mathrm{x}} J(\mathrm{x}^\star)$ and $\nabla_{\mathrm{x}} H(\mathrm{x}^\star)$ are collinear there is no feasible descent direction d. Therefore, $x^\star$ is a local solution of the problem.

Lagrange multipliers

Assume J and functions H_i are continuously differentials (and independent)

$$\mathcal{P} = \begin{cases} \min\limits_{\mathrm{x}\in\mathbb{R}^n} & J(\mathrm{x}) & \\ \text{with} & H_1(\mathrm{x}) = 0 & \lambda_1 \\ \text{and} & H_2(\mathrm{x}) = 0 & \lambda_2 \\ & \dots & \\ & H_p(\mathrm{x}) = 0 & \lambda_p \end{cases}$$

each constraint is associated with λ_i : the Lagrange multiplier.

Theorem (First order optimality conditions)

for $\mathrm{x}^\star$ being a local minima of $\mathcal{P}$, it is necessary that:

$$\nabla_x J(\mathrm{x}^\star) + \sum_{i=1}^{p} \lambda_i \nabla_x H_i(\mathrm{x}^\star) = 0 \qquad \textit{and} \qquad H_i(\mathrm{x}^\star) = 0, \quad i = 1, p$$

Lagrange multipliers

Assume J and functions H_i are continuously differentials (and independent)

$$\mathcal{P} = \begin{cases} \min\limits_{\mathrm{x}\in\mathbb{R}^n} & J(\mathrm{x}) & \\ \text{with} & H_1(\mathrm{x}) = 0 & \lambda_1 \\ \text{and} & H_2(\mathrm{x}) = 0 & \lambda_2 \\ & \dots & \\ & H_p(\mathrm{x}) = 0 & \lambda_p \end{cases}$$

each constraint is associated with λ_i : the Lagrange multiplier.

Theorem (First order optimality conditions)

for $\mathrm{x}^\star$ being a local minima of $\mathcal{P}$, it is necessary that:

$$\nabla_x J(\mathrm{x}^\star) + \sum_{i=1}^{p} \lambda_i \nabla_x H_i(\mathrm{x}^\star) = 0 \qquad \textit{and} \qquad H_i(\mathrm{x}^\star) = 0, \quad i = 1, p$$

Lagrange multipliers

Assume J and functions H_i are continuously differentials (and independent)

$$\mathcal{P} = \begin{cases} \min\limits_{\mathrm{x}\in\mathbb{R}^n} & J(\mathrm{x}) & \\ \text{with} & H_1(\mathrm{x}) = 0 & \lambda_1 \\ \text{and} & H_2(\mathrm{x}) = 0 & \lambda_2 \\ & \dots & \\ & H_p(\mathrm{x}) = 0 & \lambda_p \end{cases}$$

each constraint is associated with λ_i : the Lagrange multiplier.

Theorem (First order optimality conditions)

for $\mathrm{x}^\star$ being a local minima of $\mathcal{P}$, it is necessary that:

$$\nabla_x J(\mathrm{x}^\star) + \sum_{i=1}^{p} \lambda_i \nabla_x H_i(\mathrm{x}^\star) = 0 \qquad \textit{and} \qquad H_i(\mathrm{x}^\star) = 0, \;\; i = 1, p$$

Un exemple où ca marche

$$\begin{cases} \min\limits_{\mathrm{x}\in\mathbb{R}^3} & J(\mathrm{x}) = -x_1x_2 - x_1x_3 - x_2x_3 \\ \text{avec} & H(\mathrm{x}) = x_1 + x_2 + x_3 - 3 = 0 \end{cases}$$

$$\nabla_x J(x) = -\begin{pmatrix} x_2 + x_3 \\ x_1 + x_3 \\ x_1 + x_2 \end{pmatrix} \qquad \nabla_x H(x) = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

Les conditions d'optimalité sont

$$\begin{cases} -x_2 & -x_3 & +\lambda & = 0 \\ -x_1 & & -x_3 & +\lambda & = 0 \\ -x_1 & -x_2 & & +\lambda & = 0 \\ x_1 & +x_2 & +x_3 & & = 3 \end{cases}$$

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

$$x_1 = x_2 = x_3 = 1 \text{ et } \lambda = 2$$

Un exemple où ca marche

$$\begin{cases} \min\limits_{\mathrm{x}\in\mathbb{R}^3} & J(\mathrm{x}) = -x_1x_2 - x_1x_3 - x_2x_3 \\ \text{avec} & H(\mathrm{x}) = x_1 + x_2 + x_3 - 3 = 0 \end{cases}$$

$$\nabla_x J(x) = -\begin{pmatrix} x_2 + x_3 \\ x_1 + x_3 \\ x_1 + x_2 \end{pmatrix} \qquad \nabla_x H(x) = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

Les conditions d'optimalité sont

$$\begin{cases} & -x_2 & -x_3 & +\lambda & = 0 \\ -x_1 & & -x_3 & +\lambda & = 0 \\ -x_1 & -x_2 & & +\lambda & = 0 \\ x_1 & +x_2 & +x_3 & & = 3 \end{cases}$$

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

$$x_1 = x_2 = x_3 = 1 \text{ et } \lambda = 2$$

Un autre exemple où ca marche

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

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

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

...et on résout le système linéaire.

Attention si en plus on cherche $\mathrm{x} \geq 0$ ca devient plus compliqué

Un exemple où ca ne marche pas

$$\begin{cases} \min\limits_{x_1,x_2} & x_1 + x_2 \\ \text{avec} & (x_1+1)^2 + x_2^2 = 1 \\ \text{et} & (x_1-2)^2 + x_2^2 = 4 \end{cases}$$

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 :

$$L(\mathrm{x}, \lambda) = J(x) + \sum_{i=1}^{p} \lambda_i H_i(\mathrm{x})$$

Grâce au lagrangien on retrouve les conditions d'optimalité :

$$\begin{cases} \nabla_x L(\mathrm{x}, \lambda) = 0 & \Rightarrow \quad \nabla_x J(\mathrm{x}) + \sum_{i=1}^{p} \lambda_i \nabla_x H_i(\mathrm{x}) = 0 \\ \nabla_{\lambda_i} L(\mathrm{x}, \lambda) = 0 & \Rightarrow \quad H_i(\mathrm{x}) = 0 \end{cases}$$

Interprétation graphique : optimisation multicritères.

The only one inequality constraint case

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

cost J : d is a descent direction if it exists $\varepsilon_0 \in \mathbb{R}$ such that
$\forall \varepsilon \in \mathbb{R},\ 0 < \varepsilon \leq \varepsilon_0$

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

constraint G : d is a feasible descent direction if it exists $\varepsilon_0 \in \mathbb{R}$ such that
$\forall \varepsilon \in \mathbb{R},\ 0 < \varepsilon \leq \varepsilon_0$

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

Two possibilities

If $x^\star$ lies at the limit of the feasible domain ($G(\mathrm{x}^\star) = 0$) and if vectors $\nabla_{\mathrm{x}} J(\mathrm{x}^\star)$ and $\nabla_{\mathrm{x}} G(\mathrm{x}^\star)$ are collinear and in opposite directions, there is no feasible descent direction d at that point. Therefore, $x^\star$ is a local solution of the problem... Or if $\nabla_{\mathrm{x}} J(\mathrm{x}^\star) = 0$

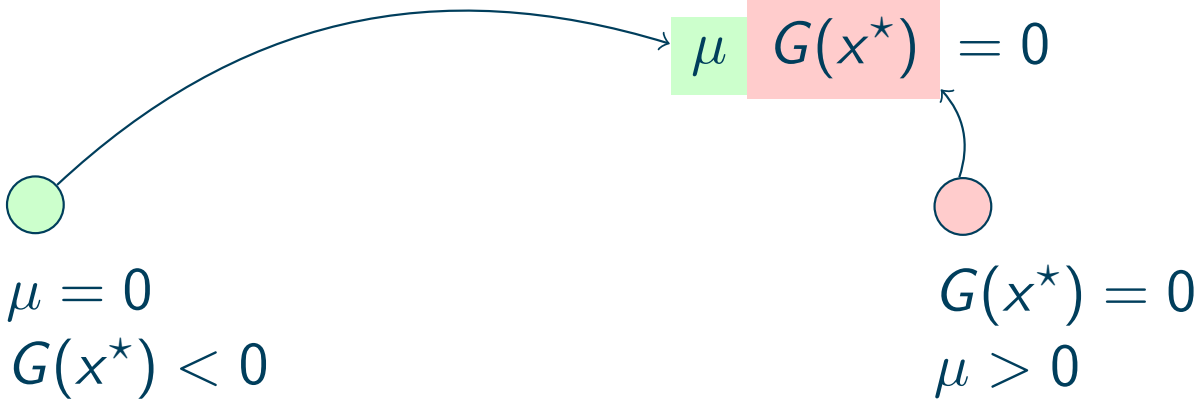
Two possibilities for optimality

$$\nabla_{\mathrm{x}} J(\mathrm{x}^\star) = -\mu \, \nabla_{\mathrm{x}} G(\mathrm{x}^\star) \quad \text{and} \quad \mu > 0; G(\mathrm{x}^\star) = 0$$

or

$$\nabla_{\mathrm{x}} J(\mathrm{x}^\star) = 0 \quad \text{and} \quad \mu = 0; G(\mathrm{x}^\star) < 0$$

This alternative is summarized in the so called complementarity condition:



First order optimality condition (1)

Le Lasso comme un problème d'optimisati avec contraintes

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

Definition: Karush, Kuhn and Tucker (KKT) conditions

$$\begin{aligned} \text{stationarity} \quad & \nabla J(x^\star) + \sum_{j=1}^{p} \lambda_j \nabla h_j(x^\star) + \sum_{i=1}^{q} \mu_i \nabla g_i(x^\star) = 0 \\ \text{primal admissibility} \quad & h_j(x^\star) = 0 \qquad j = 1, \ldots, p \\ & g_i(x^\star) \leq 0 \qquad i = 1, \ldots, q \\ \text{dual admissibility} \quad & \mu_i \geq 0 \qquad i = 1, \ldots, q \\ \text{complementarity} \quad & \mu_i g_i(x^\star) = 0 \qquad i = 1, \ldots, q \end{aligned}$$

λ_j and μ_i are called the Lagrange multipliers of problem $\mathcal{P}$

First order optimality condition (2)

Theorem (12.1 Nocedal & Wright pp 321)

If a vector $x^\star$ is a stationary point of problem $\mathcal{P}$
Then there exists[a] Lagrange multipliers such that $\left(x^\star, \{\lambda_j\}_{j=1:p}, \{\mu_i\}_{i=1:q}\right)$ fulfill KKT conditions

[a] under some conditions *e.g.* linear independence constraint qualification

If the problem is **convex**, then a stationary point is the solution of the problem

A quadratic program (QP) is convex when...

$$(QP) \quad \begin{cases} \min\limits_{\mathrm{z}} & \frac{1}{2}\mathrm{z}^\top A\mathrm{z} - \mathrm{d}^\top \mathrm{z} \\ \text{with} & B\mathrm{z} \leq \mathrm{e} \end{cases}$$

...when matrix A is positive definite

Exemple

$$\mathcal{P} = \begin{cases} \min\limits_{x \in \mathbb{R}^2} & J(x) = x_1^2 + x_2^2 \\ \text{avec} & 2x_1 + x_2 \leq -4 \end{cases}$$

$$\begin{aligned} \text{stationarité} \quad & 2x_1 + 2\mu = 0 \\ & 2x_2 + \mu = 0 \\ \text{admissibilité primal} \quad & 2x_1 + x_2 + 4 \leq 0 \\ \text{admissibilité duale} \quad & \mu \geq 0 \\ \text{complementarité} \quad & \mu(2x_1 + x_2 + 4) = 0 \end{aligned}$$

$$x_1 = -\frac{8}{5},\ x_2 = -\frac{4}{5},\ \mu = \frac{8}{5},$$

Exemple

$$\mathcal{P} = \begin{cases} \min\limits_{\mathrm{x}\in\mathbb{R}^2} & J(\mathrm{x}) = x_1^2 + x_2^2 \\ \text{avec} & 2x_1 + x_2 \leq -4 \end{cases}$$

$$\begin{aligned} \text{stationarité} \quad & 2x_1 + 2\mu = 0 \\ & 2x_2 + \mu = 0 \\ \text{admissibilité primal} \quad & 2x_1 + x_2 + 4 \leq 0 \\ \text{admissibilité duale} \quad & \mu \geq 0 \\ \text{complementarité} \quad & \mu(2x_1 + x_2 + 4) = 0 \end{aligned}$$

$$x_1 = -\frac{8}{5},\ x_2 = -\frac{4}{5},\ \mu = \frac{8}{5},$$

KKT condition - Lagrangian (3)

Le Lasso comme un problème d'optimisati avec contraintes

$$\text{problem } \mathcal{P} = \begin{cases} \min\limits_{x \in \mathbb{R}^n} & J(x) \\ \text{with} & h_j(x) = 0 \quad j = 1, \dots, p \\ \text{and} & g_i(x) \leq 0 \; i = 1, \dots, q \end{cases}$$

Definition: Lagrangian

The lagrangian of problem $\mathcal{P}$ is the following function:

$$\mathcal{L}(x, \lambda, \mu) = J(x) + \sum_{j=1}^{p} \lambda_j h_j(x) + \sum_{i=1}^{q} \mu_i g_i(x)$$

The importance of being a lagrangian

the stationarity condition can be written: $\nabla \mathcal{L}(x^\star, \lambda, \mu) = 0$

the lagrangian saddle point $\max\limits_{\lambda, \mu} \min\limits_{x} \mathcal{L}(x, \lambda, \mu)$

Primal variables: x and **dual** variables λ, μ (the Lagrange multipliers)

Duality – definitions (1)

Primal and (Lagrange) dual problems

$$\mathcal{P} = \begin{cases} \min\limits_{x \in \mathbb{R}^n} & J(x) \\ \text{with} & h_j(x) = 0 \quad j = 1, p \\ \text{and} & g_i(x) \leq 0 \quad i = 1, q \end{cases} \qquad \mathcal{D} = \begin{cases} \max\limits_{\lambda \in \mathbb{R}^p, \mu \in \mathbb{R}^q} & Q(\lambda, \mu) \\ \text{with} & \mu_j \geq 0 \quad j = 1, q \end{cases}$$

Dual objective function:

$$\begin{aligned} Q(\lambda, \mu) &= \inf_x \quad \mathcal{L}(x, \lambda, \mu) \\ &= \inf_x \quad J(x) + \sum_{j=1}^{p} \lambda_j h_j(x) + \sum_{i=1}^{q} \mu_i g_i(x) \end{aligned}$$

Wolf dual problem

$$\mathcal{W} = \begin{cases} \max\limits_{x, \lambda \in \mathbb{R}^p, \mu \in \mathbb{R}^q} & \mathcal{L}(x, \lambda, \mu) \\ \text{with} & \mu_j \geq 0 \quad j = 1, q \\ \text{and} & \nabla J(x) + \sum\limits_{j=1}^{p} \lambda_j \nabla h_j(x) + \sum\limits_{i=1}^{q} \mu_i \nabla g_i(x) = 0 \end{cases}$$

Duality – theorems (2)

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

If f, g and h are convex and continuously differentiable[a], then the solution of the dual problem is the same as the solution of the primal

[a] under some conditions *e.g.* linear independence constraint qualification

$$
\begin{aligned}
(\lambda^\star, \mu^\star) &= \text{solution of problem } \mathcal{D} \\
\mathrm{x}^\star &= \operatorname*{argmin}_{\mathrm{x}} \mathcal{L}(\mathrm{x}, \lambda^\star, \mu^\star) \\
Q(\lambda^\star, \mu^\star) = \operatorname*{argmin}_{\mathrm{x}} \mathcal{L}(\mathrm{x}, \lambda^\star, \mu^\star) &= \mathcal{L}(\mathrm{x}^\star, \lambda^\star, \mu^\star) \\
&= J(\mathrm{x}^\star) + \lambda^\star H(\mathrm{x}^\star) + \mu^\star G(\mathrm{x}^\star) = J(\mathrm{x}^\star)
\end{aligned}
$$

and for any feasible point x

$$
Q(\lambda, \mu) \leq J(\mathrm{x}) \qquad \rightarrow \qquad 0 \leq J(\mathrm{x}) - Q(\lambda, \mu)
$$

The duality gap is the difference between the primal and dual cost functions

Problème du lasso

A cause de la valeur absolue, le lagrangien n'est pas différentiable
Idée : réécrire les contraintes SANS valeur absolues

Posons $\alpha_j^+ - \alpha_j^- = \beta_j$. On a $\alpha_j^+ + \alpha_j^- = |\beta_j|$

$$\begin{cases} \min\limits_{\alpha_j^+, \alpha_j^- \in \mathbb{R}^p} & \frac{1}{2}\|X(\alpha^+ - \alpha^-) - y\|^2 \\ \text{avec} & \sum\limits_{j=1}^{p}(\alpha_j^+ + \alpha_j^-) \leq t \\ \text{et} & 0 \leq \alpha_j^+, \alpha_j^-, \qquad j = 1, \ldots, p \end{cases} \tag{2}$$

C'est un QP de la forme

$$\begin{cases} \min\limits_{z \in \mathbb{R}^q} & \frac{1}{2} z^\top H z + z^\top c \\ \text{avec} & Az \leq b \\ \text{et} & 0 \leq z \end{cases} \tag{3}$$

Le Dual du Lasso

Le Lasso comme un problème d'optimisati avec contraintes

$$\begin{cases} \min\limits_{\alpha_j^+,\alpha_j^- \in \mathbb{R}^p} & \frac{1}{2}\|X(\alpha^+ - \alpha^-) - y\|^2 \\ \text{avec} & \sum\limits_{j=1}^{p}(\alpha_j^+ + \alpha_j^-) \leq t \\ \text{et} & 0 \leq \alpha_j^+, \alpha_j^-, \qquad j = 1, \ldots, p \end{cases} \tag{4}$$

$$\mathcal{W} = \begin{cases} \max\limits_{x,\lambda \in \mathbb{R}^p, \mu \in \mathbb{R}^q} & \mathcal{L}(x, \lambda, \mu) \\ \text{with} & \mu_j \geq 0 \quad j = 1, q \\ \text{and} & \nabla J(x) + \sum\limits_{j=1}^{p} \lambda_j \nabla h_j(x) + \sum\limits_{i=1}^{q} \mu_i \nabla g_i(x) = 0 \end{cases}$$

$$\mathcal{L}(\alpha_j^+, \alpha_j^-, \lambda, \gamma_j^+, \gamma_j^-) =$$

$$\frac{1}{2}\|X(\alpha^+ - \alpha^-) - y\|^2 + \lambda\left(\sum_{j=1}^{p}(\alpha_j^+ + \alpha_j^-) - t\right) - \sum_{j=1}^{p}\gamma^+\alpha_j^+ - \sum_{j=1}^{p}\gamma^-\alpha_j^-$$

$$\begin{cases} \nabla_{\alpha^+}\mathcal{L}(\alpha_j^+, \alpha_j^-, \lambda, \gamma_j^+, \gamma_j^-) = \quad X^\top(X(\alpha^+ - \alpha^-) - y) + \lambda - \gamma^+ \\ \nabla_{\alpha^-}\mathcal{L}(\alpha_j^+, \alpha_j^-, \lambda, \gamma_j^+, \gamma_j^-) = -X^\top(X(\alpha^+ - \alpha^-) - y) + \lambda - \gamma^- \end{cases}$$

Dual du Lasso

$$\begin{array}{ll} \max\limits_{\alpha_j^+,\alpha_j^-,\lambda,\gamma_j^+,\gamma_j^-} & \frac{1}{2}\|X(\alpha^+-\alpha^-)-y\|^2+\lambda\left(\sum_{j=1}^{p}(\alpha_j^++\alpha_j^-)-t\right)-\sum_{j=1}^{p}\gamma^+\alpha_j^+-\sum_{j=1}^{p} \\ \text{with} & \lambda\geq 0,\ \gamma_j^+\geq 0,\ \gamma_j^-\geq 0 \quad j=1,p \\ \text{and} & X^\top(X(\alpha^+-\alpha^-)-y)+\lambda-\gamma^+=0 \\ & -X^\top(X(\alpha^+-\alpha^-)-y)+\lambda-\gamma^-=0 \end{array}$$

$$\begin{cases} \max\limits_{\beta} & -\frac{1}{2}\|X\beta\|^2 \\ s.t. & \|X^\top(X\beta-y)\|_\infty\leq\lambda \end{cases}$$

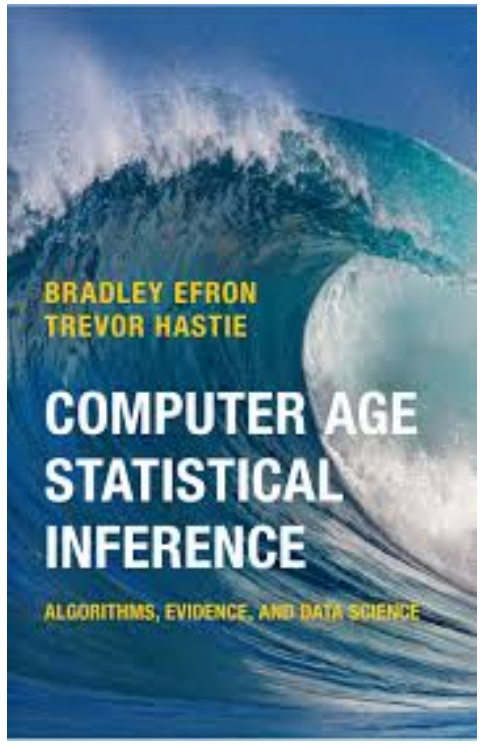
Conclusions

Le Lasso comme un problème d'optimisati avec contraintes

Outline

1. Introduction
   - Existence of a solution
2. Gradient, subgradient and subdifferentials
3. Le Lasso comme un problème d'optimisation avec contraintes
   - Constraints
     - Equality constraints
     - Inequality constraints
4. Conclusions

Conclusions



- key issue: Statistics and optimization
- beware: hyperparameter tuning
- focus: details (preprocessing, normalization, intercept. . .)
- my choice: I use the adaptive lasso after variable screening
- my research: non convex and L_0 penalty (MIP optimization)