

Le Lasso

Stéphane Canu
stephane.canu@insa-rouen.fr

Advanced Machine Learning

Winter 2024-2025

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

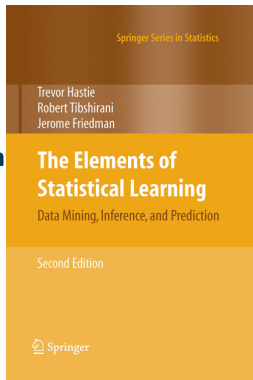
- Existence of a solution

2 Gradient, subgradient and subdifferentials

3 Le Lasso comme un problème d'optimisation a

- Constraints
 - Equality constraints
 - Inequality constraints

4 Conclusions



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

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)

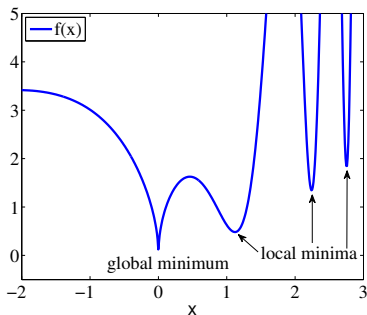
x^* is a local minima for J if there exists $\varepsilon > 0$ such that

$$J(x^*) \leq J(x) \quad \forall x \text{ with } \|x - x^*\| \leq \varepsilon$$

Definition (global minimum)

x^* is a global minimum for f if

$$J(x^*) \leq J(x) \quad \forall x \in \mathbb{R}^p$$



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 $\{x_k, k \in \mathbb{N}\}$ that converges to x

$$J(x) \leq \liminf_{k \rightarrow \infty} J(x_k).$$

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 $\{x_k, k \in \mathbb{N}\}$ such that $\lim_{k \rightarrow \infty} \|x_k\| = \infty$

$$\lim_{k \rightarrow \infty} J(x_k) = \infty.$$

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

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

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

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

Definition (gradient)

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

Example

The least square

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

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 $d \in \mathbb{R}^n$ la limite :

$$D_x J(x, d) = \lim_{\varepsilon \rightarrow 0} \frac{J(x + \varepsilon d) - J(x)}{\varepsilon}$$

si elle existe

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

Theorem (Calcul pratique de la dérivée)

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

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

recette :

Démonstration :

$$\begin{aligned} \varphi'(0) &= \lim_{\varepsilon \rightarrow 0} \frac{\varphi(\varepsilon) - \varphi(0)}{\varepsilon} \\ &= \lim_{\varepsilon \rightarrow 0} \frac{J(\mathbf{x} + \varepsilon \mathbf{d}) - J(\mathbf{x})}{\varepsilon} = D_{\mathbf{x}}J(\mathbf{x}, \mathbf{d}) \end{aligned}$$

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

Exemples de dérivées directionnelles

Gradient, subgradient
and subdifferentials

Exemples

$$J_4(x) = \|x\|^2 \quad D_x J_4(x, d) = 2x^\top d$$

$$J_5(x) = \|Ax - b\|^2 \quad D_x J_5(x, d) = 2(Ax - b)^\top Ad$$

$$J_6(x) = \|x\|_2 \quad D_x J_6(x, d) = \frac{x^\top d}{\|x\|}$$

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

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

Calculez

$D_x J_5(x, d)$ (et $D_x J_6(x, d)$ chez vous)

Theorem ($p=1$)

Si, pour x fixé, l'application $d \rightarrow D_x J(x, d)$ est linéaire continue,

$$D_x J(x, d) = \nabla J_x(x)^\top d$$

Calcul pratique du gradient :

- 1 calculer $J(x + \varepsilon d)$
- 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 $\nabla J_x(x)^\top d$ et identifier le gradient

Exemple de calcul de dérivée directionnelle

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

pour x et d fixés :

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

Règles de calcul pour la dérivée

c'est un opérateur linéaire :

$$\nabla(J_1 + \alpha J_2)_x(x) = \nabla_x J_1(x) + \alpha \nabla_x J_2(x)$$

Combinaison d'applications (chain rule) :

exemples

$$J_\alpha(x) = \frac{1}{2}x^\top Ax - x^\top b$$

$$J_\beta(x) = \mathbf{1}^\top \log(Ax)$$

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

$$\nabla_x J_\gamma(x) = f'(a^\top x)a$$

Propriété : développement au premier ordre

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

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

Application dans le cas $p=1$:

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

$$J(x + \varepsilon d) < J(x)$$

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

Démonstration

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

Definition (Matrice Hessienne)

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

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

Exemples de dérivée seconde

Gradient, subgradient
and subdifferentials

Exemples

$$J_2(x) = Ax - b$$

$$\nabla_x J(x) = A$$

$$H_x(x) = 0$$

$$J_4(x) = \|x\|^2$$

$$\nabla_x J(x) = 2x$$

$$H_x(x) = 2I$$

$$J_5(x) = \|Ax - b\|^2$$

$$\nabla_x J(x) = 2A^T(Ax - b)$$

$$H_x(x) = 2A^T A$$

$$J_6(x) = \|x\|_2$$

$$\nabla_x J(x) = \frac{x}{\|x\|}$$

$$H_x(x) = \frac{1}{\|x\|^2} M(x)$$

$$J_8(f) = \int_0^1 f(t)^2 dt$$

$$DJ_f(f) = f$$

$$H_f(f) = I$$

Développement au second ordre

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

Propriété : développement de Taylor au second ordre

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

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

ce résultat peut être généralisé

Exemple

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

$$\nabla_x J(x) = Ax - b$$

$$H_x(x) = A$$

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

Si A n'est pas définie positive...

Definition (fonction convexe)

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

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

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

exemple

$$\begin{aligned} J_1(x) &= a^\top x & J_2(x) &= \frac{1}{2}x^\top Ax \\ J_3(x) &= \sum_{i=1}^d \exp^{-x_i} & J_4(x) &= -\log a^\top x \end{aligned}$$

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.

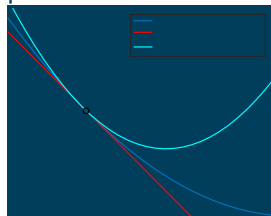
Caractérisation d'une fonction convexe

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

Exemple

$$J(x) = \frac{1}{2} \|Ax - b\|^2 \quad H = A^T A$$

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



$$\nabla_x J(x)^T (y - x) \leq J(y) - J(x) \leq \nabla_x J(x)^T (y - x) + \frac{L}{2} \|y - x\|^2$$

si J est convexe

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

et donc pour tout $\alpha > 0$

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

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

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

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

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

est convexe. Donc

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

d'où

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

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

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

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 $\partial J(x_0)$

exemple

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

Exemple de sous différentielle

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

un sous gradient est un réel g tel que

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

soit :

si $x_0 > 0$, alors $J(x) = x = x_0 + (x - x_0)$ et $g = 1$

si $x_0 < 0$, alors $J(x) = -x = -x_0 + -(x - x_0)$ 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 :

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

La sous différentielle du Lasso

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

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

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

$$x^* = \underset{x}{\operatorname{argmin}} J(x) \iff \nabla J(x^*) = 0$$

For J convex & nondifferentiable, global minima

$$x^* = \underset{x}{\operatorname{argmin}} J(x) \iff 0 \in \partial J(x^*)$$

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

$$x^* = \underset{x \in V(x^*)}{\operatorname{argmin}} J(x) \implies 0 \in \partial_c J(x^*)$$

or equivalently (sometimes referred as Oresme's rule)

$$x^* = \underset{x \in V(x^*)}{\operatorname{argmin}} J(x) \implies D_c J(x^*, x) \geq 0, \quad \forall x \in V(x^*)$$

Fermat's rule and critical points

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

	differentiable	non differentiable
cvx	$x^* = \underset{x}{\operatorname{argmin}} J(x) \Leftrightarrow \nabla J(x^*) = 0$	$x^* = \underset{x}{\operatorname{argmin}} J(x) \Leftrightarrow 0 \in \partial J(x^*)$
non cvx	$x^* = \underset{x \in V(x_0)}{\operatorname{argmin}} J(x) \Rightarrow \nabla J(x^*) = 0$	$x^* = \underset{x \in V(x_0)}{\operatorname{argmin}} J(x) \Rightarrow 0 \in \partial_c J(x^*)$

Definition (Clarke critical (or stationary) point)

Let f be a locally Lipschitz continuous function at x . A point x^* is called a (Clarke) critical point if

$$0 \in \partial_c J(x)$$

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

Le Lasso comme un
problème d'optimisation
avec contraintes

$$\left\{ \begin{array}{l} \min_{\beta \in \mathbb{R}^p} \quad \frac{1}{2} \|X\beta - y\|^2 \\ \text{avec} \quad \sum_{j=1}^p |\beta_j| \leq t \end{array} \right. \quad (1)$$

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

stationarity $X^\top (X\beta - y) + \lambda\alpha = 0$

primal admissibility $\sum_{j=1}^p |\beta_j| \leq t$

dual admissibility $\lambda \geq 0$

complementarity $\lambda(\sum_{j=1}^p |\beta_j| - t) = 0$

1 Introduction

- Existence of a solution

2 Gradient, subgradient and subdifferentials

3 Le Lasso comme un problème d'optimisation avec contraintes

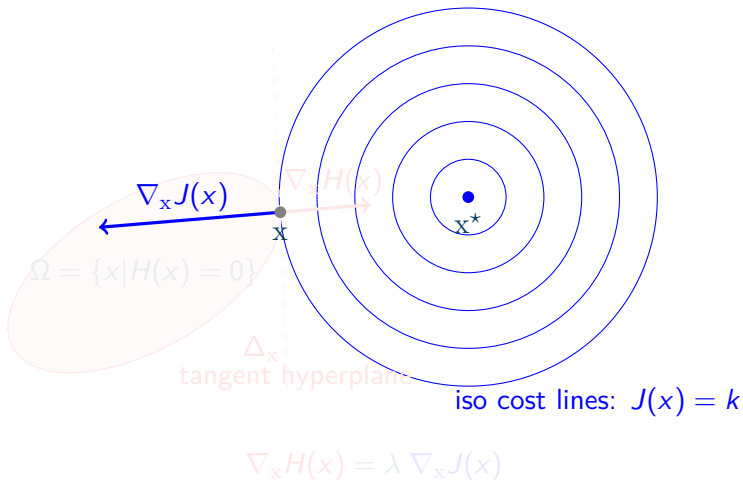
- Constraints
 - Equality constraints
 - Inequality constraints

4 Conclusions

A simple example (to begin with)

Le Lasso comme un
problème d'optimisation
avec contraintes

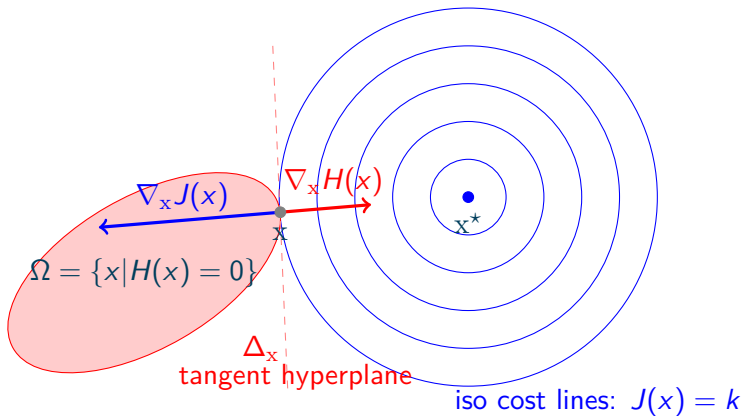
$$\begin{cases} \min_{x_1, x_2} & J(x) = (x_1 - a)^2 + (x_2 - b)^2 \\ \text{with} & H(x) = \alpha(x_1 - c)^2 + \beta(x_2 - d)^2 + \gamma x_1 x_2 - 1 \end{cases}$$



A simple example (to begin with)

Le Lasso comme un
problème d'optimisation
avec contraintes

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



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

The only one equality constraint case

Le Lasso comme un
problème d'optimisation
avec contraintes

$$\begin{cases} \min_x & J(x) & J(x + \varepsilon d) \approx J(x) + \varepsilon \nabla_x J(x)^\top d \\ \text{with} & H(x) = 0 & H(x + \varepsilon d) \approx H(x) + \varepsilon \nabla_x H(x)^\top d \end{cases}$$

Loss J : d is a descent direction if it exists $\varepsilon_0 \in \mathbb{R}$ such that

$$\forall \varepsilon \in \mathbb{R}, 0 < \varepsilon \leq \varepsilon_0$$

$$J(x + \varepsilon d) < J(x) \quad \Rightarrow \quad \nabla_x J(x)^\top d < 0$$

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

$$H(x + \varepsilon d) = 0 \quad \Rightarrow \quad \nabla_x H(x)^\top d = 0$$

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

Lagrange multipliers

Le Lasso comme un
problème d'optimisation
avec contraintes

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

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

each constraint is associated with λ_i : the Lagrange multiplier.

Theorem (First order optimality conditions)

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

$$\nabla_x J(x^*) + \sum_{i=1}^p \lambda_i \nabla_x H_i(x^*) = 0 \quad \text{and} \quad H_i(x^*) = 0, \quad i = 1, p$$

Lagrange multipliers

Le Lasso comme un
problème d'optimisation
avec contraintes

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

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

each constraint is associated with λ_i : the Lagrange multiplier.

Theorem (First order optimality conditions)

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

$$\nabla_x J(x^*) + \sum_{i=1}^p \lambda_i \nabla_x H_i(x^*) = 0 \quad \text{and} \quad H_i(x^*) = 0, \quad i = 1, p$$

Lagrange multipliers

Le Lasso comme un
problème d'optimisation
avec contraintes

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

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

each constraint is associated with λ_i : the Lagrange multiplier.

Theorem (First order optimality conditions)

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

$$\nabla_x J(x^*) + \sum_{i=1}^p \lambda_i \nabla_x H_i(x^*) = 0 \quad \text{and} \quad H_i(x^*) = 0, \quad i = 1, p$$

Un exemple où ça marche

Le Lasso comme un
problème d'optimisation
avec contraintes

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

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

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 $(A \setminus b)$ donne :

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

Un exemple où ça marche

Le Lasso comme un
problème d'optimisation
avec contraintes

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

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

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 $(A \setminus b)$ donne :

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

Un autre exemple où ça marche

Le Lasso comme un
problème d'optimisation
avec contraintes

$$\begin{cases} \min_{x \in \mathbb{R}^n} & J(x) = -\frac{1}{2} x^T A x - x^T b \\ \text{avec} & Cx = d \end{cases}$$

$$\begin{aligned} \nabla_x J &= Ax - b \\ \nabla_x H &= C^T \end{aligned}$$

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

...et on résout le système linéaire.

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

Un exemple où ça ne marche pas

Le Lasso comme un
problème d'optimisation
avec contraintes

$$\left\{ \begin{array}{l} \min_{x_1, x_2} \quad x_1 + x_2 \\ \text{avec} \quad (x_1 + 1)^2 + x_2^2 = 1 \\ \text{et} \quad (x_1 - 2)^2 + x_2^2 = 4 \end{array} \right.$$

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

Lagrangien

Le Lasso comme un
problème d'optimisation
avec contraintes

Une fonction bien pratique :

définition : lagrangien

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

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

Grâce au lagrangien on retrouve les conditions d'optimalité :

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

Interprétation graphique : optimisation multicritères.

The only one inequality constraint case

Le Lasso comme un problème d'optimisation avec contraintes

$$\begin{cases} \min_x & J(x) & J(x + \varepsilon d) \approx J(x) + \varepsilon \nabla_x J(x)^\top d \\ \text{with} & G(x) \leq 0 & G(x + \varepsilon d) \approx G(x) + \varepsilon \nabla_x G(x)^\top d \end{cases}$$

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

$$J(x + \varepsilon d) < J(x) \quad \Rightarrow \quad \nabla_x J(x)^\top d < 0$$

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

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

Two possibilities

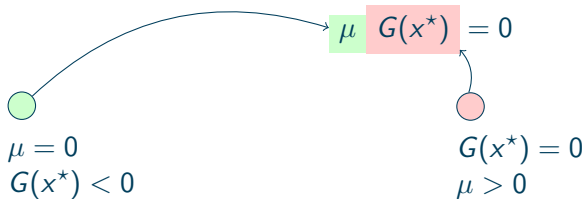
If x^* lies at the limit of the feasible domain ($G(x^*) = 0$) and if vectors $\nabla_x J(x^*)$ and $\nabla_x G(x^*)$ are collinear **and in opposite directions**, there is no feasible descent direction d at that point. Therefore, x^* is a local solution of the problem... Or if $\nabla_x J(x^*) = 0$

Two possibilities for optimality

Le Lasso comme un
problème d'optimisation
avec contraintes

$$\begin{array}{ll} \nabla_x J(x^*) = -\mu \nabla_x G(x^*) & \text{and } \mu > 0; G(x^*) = 0 \\ \text{or} & \\ \nabla_x J(x^*) = 0 & \text{and } \mu = 0; G(x^*) < 0 \end{array}$$

This alternative is summarized in the so called complementarity condition:



First order optimality condition (1)

Le Lasso comme un
problème d'optimisation
avec contraintes

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

Definition: Karush, Kuhn and Tucker (KKT) conditions

$$\text{stationarity} \quad \nabla J(x^*) + \sum_{j=1}^p \lambda_j \nabla h_j(x^*) + \sum_{i=1}^q \mu_i \nabla g_i(x^*) = 0$$

$$\text{primal admissibility} \quad \begin{array}{ll} h_j(x^*) = 0 & j = 1, \dots, p \\ g_i(x^*) \leq 0 & i = 1, \dots, q \end{array}$$

$$\text{dual admissibility} \quad \mu_i \geq 0 \quad i = 1, \dots, q$$

$$\text{complementarity} \quad \mu_i g_i(x^*) = 0 \quad i = 1, \dots, q$$

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

First order optimality condition (2)

Le Lasso comme un
problème d'optimisation
avec contraintes

Theorem (12.1 Nocedal & Wright pp 321)

If a vector x^* is a stationary point of problem \mathcal{P}

Then there exists^a Lagrange multipliers such that $(x^*, \{\lambda_j\}_{j=1:p}, \{\mu_i\}_{i=1:q})$
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...

$$(QP) \quad \begin{cases} \min_z & \frac{1}{2}z^T A z - d^T z \\ \text{with} & Bz \leq e \end{cases}$$

...when matrix A is positive definite

Exemple

Le Lasso comme un
problème d'optimisation
avec contraintes

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

stationarité $2x_1 + 2\mu = 0$
 $2x_2 + \mu = 0$

admissibilité primal $2x_1 + x_2 + 4 \leq 0$

admissibilité duale $\mu \geq 0$

complémentarité $\mu(2x_1 + x_2 + 4) = 0$

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

Exemple

Le Lasso comme un
problème d'optimisation
avec contraintes

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

stationarité $2x_1 + 2\mu = 0$
 $2x_2 + \mu = 0$

admissibilité primal $2x_1 + x_2 + 4 \leq 0$

admissibilité duale $\mu \geq 0$

complémentarité $\mu(2x_1 + x_2 + 4) = 0$

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

KKT condition - Lagrangian (3)

Le Lasso comme un
problème d'optimisation
avec contraintes

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

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: $\nabla \mathcal{L}(x^*, \lambda, \mu) = 0$

the lagrangian saddle point $\max_{\lambda, \mu} \min_x \mathcal{L}(x, \lambda, \mu)$

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

Duality – definitions (1)

Le Lasso comme un
problème d'optimisation
avec contraintes

Primal and (Lagrange) dual problems

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

Dual objective function:

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

Wolf dual problem

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

Duality – theorems (2)

Le Lasso comme un
problème d'optimisation
avec contraintes

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

$$\begin{aligned}(\lambda^*, \mu^*) &= \text{solution of problem } \mathcal{D} \\ x^* &= \underset{x}{\operatorname{argmin}} \mathcal{L}(x, \lambda^*, \mu^*)\end{aligned}$$

$$\begin{aligned}Q(\lambda^*, \mu^*) &= \underset{x}{\operatorname{argmin}} \mathcal{L}(x, \lambda^*, \mu^*) = \mathcal{L}(x^*, \lambda^*, \mu^*) \\ &= J(x^*) + \lambda^* H(x^*) + \mu^* G(x^*) = J(x^*)\end{aligned}$$

and for any feasible point x

$$Q(\lambda, \mu) \leq J(x) \quad \rightarrow \quad 0 \leq J(x) - Q(\lambda, \mu)$$

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

Problème du lasso

Le Lasso comme un
problème d'optimisation
avec contraintes

A cause de la valeur absolue, le lagrangien n'est pas différentiable

Idee : réécrire les contraintes SANS valeur absolues

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

$$\left\{ \begin{array}{ll} \min_{\alpha_j^+, \alpha_j^- \in \mathbb{R}^p} & \frac{1}{2} \|X(\alpha^+ - \alpha^-) - y\|^2 \\ \text{avec} & \sum_{j=1}^p (\alpha_j^+ + \alpha_j^-) \leq t \\ \text{et} & 0 \leq \alpha_j^+, \alpha_j^-, \quad j = 1, \dots, p \end{array} \right. \quad (2)$$

C'est un QP de la forme

$$\left\{ \begin{array}{ll} \min_{z \in \mathbb{R}^q} & \frac{1}{2} z^T H z + z^T c \\ \text{avec} & A z \leq b \\ \text{et} & 0 \leq z \end{array} \right. \quad (3)$$

Le Dual du Lasso

Le Lasso comme un
problème d'optimisation
avec contraintes

$$\left\{ \begin{array}{ll} \min_{\alpha_j^+, \alpha_j^- \in \mathbb{R}^p} & \frac{1}{2} \|X(\alpha^+ - \alpha^-) - y\|^2 \\ \text{avec} & \sum_{j=1}^p (\alpha_j^+ + \alpha_j^-) \leq t \\ \text{et} & 0 \leq \alpha_j^+, \alpha_j^-, \quad j = 1, \dots, p \end{array} \right. \quad (4)$$

$$\mathcal{W} = \left\{ \begin{array}{ll} \max_{x, \lambda \in \mathbb{R}^p, \mu \in \mathbb{R}^q} & \mathcal{L}(x, \lambda, \mu) \\ \text{with} & \mu_j \geq 0 \quad j = 1, q \\ \text{and} & \nabla J(x) + \sum_{j=1}^p \lambda_j \nabla h_j(x) + \sum_{i=1}^q \mu_i \nabla g_i(x) = 0 \end{array} \right.$$

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

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

$$\left\{ \begin{array}{l} \nabla_{\alpha^+} \mathcal{L}(\alpha_j^+, \alpha_j^-, \lambda, \gamma_j^+, \gamma_j^-) = X^T (X(\alpha^+ - \alpha^-) - y) + \lambda - \gamma^+ \\ \nabla_{\alpha^-} \mathcal{L}(\alpha_j^+, \alpha_j^-, \lambda, \gamma_j^+, \gamma_j^-) = -X^T (X(\alpha^+ - \alpha^-) - y) + \lambda - \gamma^- \end{array} \right.$$

Dual du Lasso

Le Lasso comme un
problème d'optimisation
avec contraintes

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

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

Conclusions

Le Lasso comme un
problème d'optimisation
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)

Conclusions

