

Le Lasso

Le gradient proximal

S.Canu
AML ITI /INSA de Rouen Normandie

13 octobre 2025

Le Lasso

- Données
 - ▶ $X \in \mathbb{R}^{n \times p}$ les p variables observées n fois
 - ▶ $y \in \mathbb{R}^n$ la variable réponse
- Inconnues
 - ▶ $\beta \in \mathbb{R}^p$
- hypothèses : pas de biais (pas de terme constant)
 - ▶ $\text{mean}(X) = 0$
 - ▶ $\|X_j\|^2 = 1$
 - ▶ $\text{mean}(y) = 0$

le cout pénalisé

$$J_\lambda(\beta) = \frac{1}{2} \|X\beta - y\|^2 + \lambda \|\beta\|_1 ,$$

Les conditions d'optimalité

$$0 \in \partial J_\lambda(\beta) = X^\top (X\beta - y) + \lambda \partial(\|\beta\|_1) ,$$

L'algorithme du gradient proximal

Le cout du lasso : $f(\beta) = J_\lambda(\beta) = \underbrace{g(\beta)} + \underbrace{h(\beta)}$.

- $g(\beta)$ convexe et différentiable
- $h(\beta)$ convexe et sous-différentiable

Algorithme du gradient proximal

Initialisation : par exemple $\beta^0 = 0$
choisir un pas t ,
Tant qu'on n'a pas convergé....

$$\beta^{k+1} = \text{Prox}_{th(\beta)}(\beta^k - t\nabla_\beta g(\beta^k))$$

L'algorithme du gradient proximal pour le Lasso

Le cout du lasso : $f(\beta) = J_\lambda(\beta) = \underbrace{g(\beta)}_{\frac{1}{2}\|X\beta - y\|^2} + \underbrace{h(\beta)}_{\lambda\|\beta\|_1} .$

- $g(\beta) = \frac{1}{2}\|X\beta - y\|^2$ convexe et différentiable
- $h(\beta) = \lambda\|\beta\|_1$ convexe et sous-différentiable

Algorithme du gradient proximal

Initialisation : par exemple $\beta^0 = 0$

choisir un pas t , par exemple $t = 2/\|X^T X\|$

Tant qu'on n'a pas convergé....

$$\begin{aligned}\beta^{k+1} &= \text{Prox}_{th(\beta)}(\beta^k - t\nabla_\beta g(\beta^k)) \\ \beta^{k+1} &= \text{Prox}_{t\lambda\|\cdot\|_1}(\beta^k - tX^t(X\beta^k - y))\end{aligned}$$

idée 1 : le proximal L1

Le proximal de la fonction $h(x)$ au point w est

$$\text{prox}_h(w) = \underset{u}{\operatorname{argmin}} \quad h(u) + \frac{1}{2} \|u - w\|^2$$

Le proximal du cout L1 $h(x) = \|u\|_1$ au point w est

$$\text{prox}_{l_1}(w) = \underset{u}{\operatorname{argmin}} \quad \|u\|_1 + \frac{1}{2} \|u - w\|^2$$

$$\text{prox}_{l_1}(w) = \operatorname{sign}(w) \max(|w| - 1, 0)$$

Le proximal de la fonction $h(x) = \lambda \|u\|_1$ au point w est

$$\text{prox}_{l_1}(w, \lambda) = \underset{u}{\operatorname{argmin}} \quad \lambda \|u\|_1 + \frac{1}{2} \|u - w\|^2$$

$$\text{prox}_{l_1}(w, \lambda) = \operatorname{sign}(w) \max(|w| - \lambda, 0)$$

idée 2 : la linéarisation

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

Développement de Taylor au premier ordre

$$\begin{aligned} J(\beta) &= J(\beta^0) + (\beta - \beta^0)^t \nabla_{\beta} J(\beta^0) + o(\|\beta - \beta^0\|) \\ \frac{1}{2} \|X\beta - y\|^2 &= \frac{1}{2} \|X\beta^0 - y\|^2 + (\beta - \beta^0)^t X^t (X\beta^0 - y) + o(\|\beta - \beta^0\|) \end{aligned}$$

ou, avec $v = (\beta - \beta^0)$

$$J(\beta) = J(\beta^0) + v^t \nabla_{\beta} J(\beta^0) + \int_0^1 (\nabla J(\beta^0 + tv) - \nabla J(\beta^0))^T v \, dt$$

Développement de Taylor au second ordre

$$J(\beta) = J(\beta^0) + (\beta - \beta^0)^t \nabla_{\beta} J(\beta^0) + \frac{1}{2} (\beta - \beta^0)^t H(\beta - \beta^0) + o(\|\beta - \beta^0\|^2)$$

idée 3 : la méthode de la région de confiance

Pour un β^0 donné

$$\begin{aligned} \min_{\beta} \quad & \frac{1}{2} \|X\beta - y\|^2 + \lambda \|\beta\|_1, \\ \text{avec} \quad & \|\beta - \beta^0\|^2 \leq \varepsilon \end{aligned}$$

$$\min_{\beta} \frac{1}{2} \|X\beta - y\|^2 + \lambda \|\beta\|_1 + \frac{1}{2t} (\|\beta - \beta^0\|^2 - \varepsilon)$$

idée 2+idée 3

Pour un β^0 donné

$$\min_{\beta} \frac{1}{2} \|X\beta - y\|^2 + \lambda \|\beta\|_1 + \frac{1}{2t} (\|\beta - \beta^0\|^2 - \varepsilon)$$

$$\begin{aligned} J(\beta) = \frac{1}{2} \|X\beta - y\|^2 &\approx J(\beta^0) + (\beta - \beta^0)^t \nabla_{\beta} J(\beta^0) \\ &\approx \frac{1}{2} \|X\beta^0 - y\|^2 + (\beta - \beta^0)^t X^t (X\beta^0 - y) \end{aligned}$$

$$\min_{\beta} (\beta - \beta^0)^t X^t (X\beta^0 - y) + \lambda \|\beta\|_1 + \frac{1}{2t} \|\beta - \beta^0\|^2$$

$$\min_{\beta} \lambda \|\beta\|_1 + \frac{1}{2t} \|\beta - \beta^0 + tX^t(X\beta^0 - y)\|^2$$

idée 2+idée 3+idée 1

Pour un β^0 donné

$$\min_{\beta} t\lambda\|\beta\|_1 + \frac{1}{2}\|\beta - \beta^0 + tX^t(X\beta^0 - y)\|^2$$

$$\begin{aligned} \min_{\beta} \quad & t\lambda\|\beta\|_1 + \frac{1}{2}\|\beta - v\|^2 \\ \text{avec} \quad & v = \beta^0 - tX^t(X\beta^0 - y) \end{aligned}$$

La solution est donnée par :

$$\beta = \text{Prox}_{t\lambda\|\cdot\|_1}(\beta^0 - tX^t(X\beta^0 - y))$$

L'algorithme du gradient proximal

Pour un β^0 donné

$$t = \frac{2}{\|X^t X\|}$$

Tant qu'on n'a pas convergé....

$$\begin{aligned} v &= \beta^k - tX^t(X\beta^k - y) \\ \beta^{k+1} &= \arg \min_{\beta} t\lambda\|\beta\|_1 + \frac{1}{2}\|\beta - v\|^2 \\ &= \text{sign}(v)\max(|v| - t\lambda, 0) \end{aligned}$$

```
def proxl1(w,lam):  
    return np.sign(w)*np.maximum((np.abs(w)-lam),0)
```

```
stepsize = 1/np.linalg.norm(X.T@X)  
for i in range(nb_itermax):  
    grad = -X.T@(y-X@w)  
    w = w - stepsize*grad  
    w = prox(w,stepsize*lambd)
```

Comment choisir t ?

J est gradient Lipschitz s'il existe une constante $0 < L$ telle que, $\forall \beta, \beta^0$

$$\|\nabla_{\beta} J(\beta) - \nabla_{\beta} J(\beta^0)\| \leq L \|\beta - \beta^0\|$$

Choix de t : argument de convergence

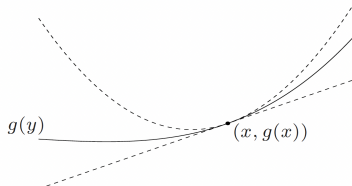
Si

$$t \leq \frac{2}{L}$$

L'algorithme du gradient proximal converge

```
stepsize = 2./np.linalg.norm(X.T@X)
```

Quadratic upper bound from Lipschitz property



- affine lower bound from convexity

$$g(y) \geq g(x) + \nabla g(x)^T (y - x) \quad \forall x, y$$

- quadratic upper bound from Lipschitz property

$$g(y) \leq g(x) + \nabla g(x)^T (y - x) + \frac{L}{2} \|y - x\|_2^2 \quad \forall x, y$$

The gradient Lipschitz Property

- g est gradient Lipschitz s'il existe L :

$$\|\nabla g(y) - \nabla g(x)\| \leq L \|y - x\|$$

Exemple $g(\beta) = \frac{1}{2}\|X\beta - y\|^2$, $\nabla g(\beta) = X^T(X\beta - y)$

$$\|\nabla g(y) - \nabla g(x)\| = \|X^T X(y - x)\| \leq \underbrace{\|X^T X\|}_{=L} \|y - x\|$$

- $v = y - x$

$$\begin{aligned} g(y) &= g(x) + \nabla g(x)^T v + \int_0^1 (\nabla g(x + tv) - \nabla g(x))^T v dt \\ &\leq g(x) + \nabla g(x)^T v + \int_0^1 \|\nabla g(x + tv) - \nabla g(x)\| \|v\| dt \\ &\leq g(x) + \nabla g(x)^T v + \int_0^1 tL \|v\|^2 dt \\ &\leq g(x) + \nabla g(x)^T v + \frac{L}{2} \|v\|^2 \end{aligned}$$

The Proximal Operator Property

Theorem (Proximal Operator Property)

Let $h : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{\infty\}$ be a proper, lower semicontinuous, convex function. For any $x \in \mathbb{R}^n$ and $t > 0$, let $z = \text{prox}_{th}(x - t\nabla g(x))$. Then for all $y \in \mathbb{R}^n$:

$$\langle z - (x - t\nabla g(x)), z - y \rangle + th(z) \leq th(y)$$

Démonstration.

- 1 The proximal operator : $\text{prox}_h(v) = \arg \min_u \{h(u) + \frac{1}{2}\|v - u\|^2\}$
- 2 Let $z = \text{prox}_h(v)$. Optimality condition $0 \in \partial h(z) + (z - v)$
- 3 This optimality condition can be rewritten as : $v - z \in \partial h(z)$
- 4 by definition of the subdifferential : $\partial h(z) = \{\delta : h(y) \geq h(z) + \langle \delta, y - z \rangle \text{ for all } y\}$
- 5 Applying this definition with $v - z \in \partial h(z)$: $h(y) \geq h(z) + \langle v - z, y - z \rangle \text{ for all } y$
- 6 Rearranging this inequality : $\langle z - v, z - y \rangle + h(z) \leq h(y) \text{ for all } y$
- 7 When $v = x - t\nabla g(x)$ and $z = \text{prox}_{th}(v)$ we have :

$$\langle z - (x - t\nabla g(x)), z - y \rangle + th(z) \leq th(y) \text{ for all } y$$

This last inequality is precisely the Proximal Operator Property, which completes the proof. $\square$

Le Lasso

Le cout du lasso : $J_\lambda(\beta) = \underbrace{g(\beta)}_{\frac{1}{2}\|X\beta - y\|^2} + \underbrace{h(\beta)}_{\lambda\|\beta\|_1} .$

- $g(\beta) = \frac{1}{2}\|X\beta - y\|^2$ is L-Lipshitz

$$\frac{1}{2}\|X\beta^{k+1} - y\|^2 \leq \frac{1}{2}\|X\beta^k - y\|^2 + (X\beta - y)^T X(\beta^{k+1} - \beta^k) + \frac{L}{2}\|\beta^{k+1} - \beta^k\|^2$$

- Proximal operator property

$$\begin{aligned} \langle \beta^{k+1} - (\beta^k - t\nabla g(\beta^k)), \beta^{k+1} - \beta^k \rangle + t\lambda\|\beta^{k+1}\|_1 &\leq t\lambda\|\beta^k\|_1 \\ \frac{1}{t}\|\beta^{k+1} - \beta^k\|^2 + (X\beta - y)^T X(\beta^{k+1} - \beta^k) + \lambda\|\beta^{k+1}\|_1 &\leq \lambda\|\beta^k\|_1 \end{aligned}$$

En combinant puis réarrangant les deux inégalités précédentes on obtient :

$$J_\lambda(\beta^{k+1}) \leq J_\lambda(\beta^k) + \left(\frac{L}{2} - \frac{1}{t} \right) \|\beta^{k+1} - \beta^k\|^2$$

Annex : Formulas for the Taylor Remainder Term

$$\phi(t) = \phi(0) + t\phi'(0) + \frac{1}{2} \int_0^t (t - \tau)\phi''(\tau)d\tau$$

Par intégration par parties

$$\begin{aligned} \int_0^t (t - \tau)\phi''(\tau)d\tau &= t \int_0^t \phi''(\tau)d\tau - \int_0^t \tau\phi''(\tau)d\tau \\ &= t \int_0^t \phi''(\tau)d\tau - [\tau\phi'(\tau)]_0^t + \int_0^t \phi'(\tau)d\tau \\ &= t(\phi'(t) - \phi'(0)) - t\phi'(t) + \int_0^t \phi'(\tau)d\tau \\ &= \int_0^t (\phi'(\tau) - \phi'(0))d\tau \end{aligned}$$

Thanks to the definition of the directional derivative in direction $v = \beta - \beta^0$

$$\phi(t) = J(\beta^0 + t(\beta - \beta^0)) \quad \phi'(t) = (\beta - \beta^0)^T \nabla J(\beta^0 + t(\beta - \beta^0))$$

Soit, pour $t = 1$ et $v = \beta - \beta^0$

$$J(\beta) = J(\beta^0) + (\beta - \beta^0)^T \nabla J(\beta^0) + \frac{1}{2} \int_0^1 (\beta - \beta^0)^T (\nabla J(\beta + tv) - \nabla J(\beta^0)) d\tau$$