

# **MS BGD**

## **MDI 720 : Statistiques**

**Joseph Salmon**

<http://josephsalmon.eu>

Télécom Paristech, Institut Mines-Télécom

# Plan

## Optimisation pour le Lasso

- Descente par coordonnée

- Alternatives

# Sommaire

Optimisation pour le Lasso

Descente par coordonnée

Alternatives

# Descente par coordonnée

Objectif : trouver une solution (approchée !) de  $\arg \min_{\theta \in \mathbb{R}^p} f(\theta)$

---

**Algorithme** : Descente par coordonnée

---

**Entrées** :  $f$ , nombre d'« époques »  $K$  (ou de « passes »)

Initialisation :  $k = 0$  et  $\theta^{(k)} = 0 \in \mathbb{R}^p$

# Descente par coordonnée

Objectif : trouver une solution (approchée !) de  $\arg \min_{\theta \in \mathbb{R}^p} f(\theta)$

---

**Algorithme** : Descente par coordonnée

---

**Entrées** :  $f$ , nombre d'« époques »  $K$  (ou de « passes »)

Initialisation :  $k = 0$  et  $\theta^{(k)} = 0 \in \mathbb{R}^p$

**pour**  $k = 1, \dots, K$  **faire**

# Descente par coordonnée

Objectif : trouver une solution (approchée !) de  $\arg \min_{\theta \in \mathbb{R}^p} f(\theta)$

---

**Algorithme** : Descente par coordonnée

---

**Entrées** :  $f$ , nombre d'« époques »  $K$  (ou de « passes »)

Initialisation :  $k = 0$  et  $\theta^{(k)} = 0 \in \mathbb{R}^p$

**pour**  $k = 1, \dots, K$  **faire**

$$\theta_1^{(k)} \leftarrow \arg \min_{\theta_1 \in \mathbb{R}} f(\theta_1, \theta_2^{(k-1)}, \theta_3^{(k-1)}, \dots, \theta_{p-1}^{(k-1)}, \theta_p^{(k-1)})$$

# Descente par coordonnée

Objectif : trouver une solution (approchée !) de  $\arg \min_{\theta \in \mathbb{R}^p} f(\theta)$

---

**Algorithme** : Descente par coordonnée

---

**Entrées** :  $f$ , nombre d'« époques »  $K$  (ou de « passes »)

Initialisation :  $k = 0$  et  $\theta^{(k)} = 0 \in \mathbb{R}^p$

**pour**  $k = 1, \dots, K$  **faire**

$$\theta_1^{(k)} \leftarrow \arg \min_{\theta_1 \in \mathbb{R}} f(\theta_1, \theta_2^{(k-1)}, \theta_3^{(k-1)}, \dots, \theta_{p-1}^{(k-1)}, \theta_p^{(k-1)})$$

$$\theta_2^{(k)} \leftarrow \arg \min_{\theta_2 \in \mathbb{R}} f(\theta_1^{(k)}, \theta_2, \theta_3^{(k-1)}, \dots, \theta_{p-1}^{(k-1)}, \theta_p^{(k-1)})$$

# Descente par coordonnée

Objectif : trouver une solution (approchée !) de  $\arg \min_{\theta \in \mathbb{R}^p} f(\theta)$

---

**Algorithme** : Descente par coordonnée

---

**Entrées** :  $f$ , nombre d'« époques »  $K$  (ou de « passes »)

Initialisation :  $k = 0$  et  $\theta^{(k)} = 0 \in \mathbb{R}^p$

**pour**  $k = 1, \dots, K$  **faire**

$$\theta_1^{(k)} \leftarrow \arg \min_{\theta_1 \in \mathbb{R}} f(\theta_1, \theta_2^{(k-1)}, \theta_3^{(k-1)}, \dots, \theta_{p-1}^{(k-1)}, \theta_p^{(k-1)})$$

$$\theta_2^{(k)} \leftarrow \arg \min_{\theta_2 \in \mathbb{R}} f(\theta_1^{(k)}, \theta_2, \theta_3^{(k-1)}, \dots, \theta_{p-1}^{(k-1)}, \theta_p^{(k-1)})$$

$$\theta_3^{(k)} \leftarrow \arg \min_{\theta_3 \in \mathbb{R}} f(\theta_1^{(k)}, \theta_2^{(k)}, \theta_3, \dots, \theta_{p-1}^{(k-1)}, \theta_p^{(k-1)})$$



# Descente par coordonnée

Objectif : trouver une solution (approchée !) de  $\arg \min_{\theta \in \mathbb{R}^p} f(\theta)$

---

**Algorithme** : Descente par coordonnée

---

**Entrées** :  $f$ , nombre d'« époques »  $K$  (ou de « passes »)

Initialisation :  $k = 0$  et  $\theta^{(k)} = 0 \in \mathbb{R}^p$

**pour**  $k = 1, \dots, K$  **faire**

$$\theta_1^{(k)} \leftarrow \arg \min_{\theta_1 \in \mathbb{R}} f(\theta_1, \theta_2^{(k-1)}, \theta_3^{(k-1)}, \dots, \theta_{p-1}^{(k-1)}, \theta_p^{(k-1)})$$

$$\theta_2^{(k)} \leftarrow \arg \min_{\theta_2 \in \mathbb{R}} f(\theta_1^{(k)}, \theta_2, \theta_3^{(k-1)}, \dots, \theta_{p-1}^{(k-1)}, \theta_p^{(k-1)})$$

$$\theta_3^{(k)} \leftarrow \arg \min_{\theta_3 \in \mathbb{R}} f(\theta_1^{(k)}, \theta_2^{(k)}, \theta_3, \dots, \theta_{p-1}^{(k-1)}, \theta_p^{(k-1)})$$

$$\vdots$$

# Descente par coordonnée

Objectif : trouver une solution (approchée !) de  $\arg \min_{\theta \in \mathbb{R}^p} f(\theta)$

---

**Algorithme** : Descente par coordonnée

---

**Entrées** :  $f$ , nombre d'« époques »  $K$  (ou de « passes »)

Initialisation :  $k = 0$  et  $\theta^{(k)} = 0 \in \mathbb{R}^p$

**pour**  $k = 1, \dots, K$  **faire**

$$\theta_1^{(k)} \leftarrow \arg \min_{\theta_1 \in \mathbb{R}} f(\theta_1, \theta_2^{(k-1)}, \theta_3^{(k-1)}, \dots, \theta_{p-1}^{(k-1)}, \theta_p^{(k-1)})$$

$$\theta_2^{(k)} \leftarrow \arg \min_{\theta_2 \in \mathbb{R}} f(\theta_1^{(k)}, \theta_2, \theta_3^{(k-1)}, \dots, \theta_{p-1}^{(k-1)}, \theta_p^{(k-1)})$$

$$\theta_3^{(k)} \leftarrow \arg \min_{\theta_3 \in \mathbb{R}} f(\theta_1^{(k)}, \theta_2^{(k)}, \theta_3, \dots, \theta_{p-1}^{(k-1)}, \theta_p^{(k-1)})$$

$$\vdots$$

$$\theta_p^{(k)} \leftarrow \arg \min_{\theta_p \in \mathbb{R}} f(\theta_1^{(k)}, \theta_2^{(k)}, \theta_3^{(k)}, \dots, \theta_{p-1}^{(k)}, \theta_p)$$

# Descente par coordonnée

Objectif : trouver une solution (approchée !) de  $\arg \min_{\theta \in \mathbb{R}^p} f(\theta)$

---

**Algorithme** : Descente par coordonnée

---

**Entrées** :  $f$ , nombre d'« époques »  $K$  (ou de « passes »)

Initialisation :  $k = 0$  et  $\theta^{(k)} = 0 \in \mathbb{R}^p$

**pour**  $k = 1, \dots, K$  **faire**

$$\theta_1^{(k)} \leftarrow \arg \min_{\theta_1 \in \mathbb{R}} f(\theta_1, \theta_2^{(k-1)}, \theta_3^{(k-1)}, \dots, \theta_{p-1}^{(k-1)}, \theta_p^{(k-1)})$$

$$\theta_2^{(k)} \leftarrow \arg \min_{\theta_2 \in \mathbb{R}} f(\theta_1^{(k)}, \theta_2, \theta_3^{(k-1)}, \dots, \theta_{p-1}^{(k-1)}, \theta_p^{(k-1)})$$

$$\theta_3^{(k)} \leftarrow \arg \min_{\theta_3 \in \mathbb{R}} f(\theta_1^{(k)}, \theta_2^{(k)}, \theta_3, \dots, \theta_{p-1}^{(k-1)}, \theta_p^{(k-1)})$$

$$\vdots$$

$$\theta_p^{(k)} \leftarrow \arg \min_{\theta_p \in \mathbb{R}} f(\theta_1^{(k)}, \theta_2^{(k)}, \theta_3^{(k)}, \dots, \theta_{p-1}^{(k)}, \theta_p)$$

**Sorties** :  $\theta^{(K)}$

---

# Descente par coordonnée

Objectif : trouver une solution (approchée !) de  $\arg \min_{\theta \in \mathbb{R}^p} f(\theta)$

---

**Algorithme** : Descente par coordonnée

---

**Entrées** :  $f$ , nombre d'« époques »  $K$  (ou de « passes »)

Initialisation :  $k = 0$  et  $\theta^{(k)} = 0 \in \mathbb{R}^p$

**pour**  $k = 1, \dots, K$  **faire**

$$\theta_1^{(k)} \leftarrow \arg \min_{\theta_1 \in \mathbb{R}} f(\theta_1, \theta_2^{(k-1)}, \theta_3^{(k-1)}, \dots, \theta_{p-1}^{(k-1)}, \theta_p^{(k-1)})$$

$$\theta_2^{(k)} \leftarrow \arg \min_{\theta_2 \in \mathbb{R}} f(\theta_1^{(k)}, \theta_2, \theta_3^{(k-1)}, \dots, \theta_{p-1}^{(k-1)}, \theta_p^{(k-1)})$$

$$\theta_3^{(k)} \leftarrow \arg \min_{\theta_3 \in \mathbb{R}} f(\theta_1^{(k)}, \theta_2^{(k)}, \theta_3, \dots, \theta_{p-1}^{(k-1)}, \theta_p^{(k-1)})$$

$$\vdots$$


$$\theta_p^{(k)} \leftarrow \arg \min_{\theta_p \in \mathbb{R}} f(\theta_1^{(k)}, \theta_2^{(k)}, \theta_3^{(k)}, \dots, \theta_{p-1}^{(k)}, \theta_p)$$

**Sorties** :  $\theta^{(K)}$

---

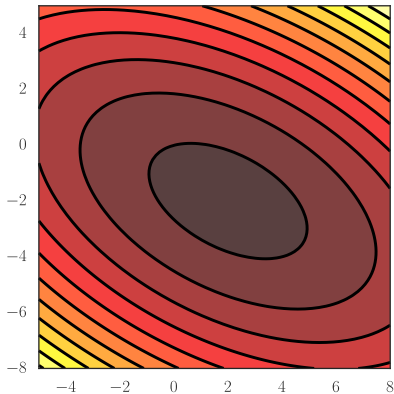
Critères d'arrêts : itérés stables, objectifs stables, saut de dualité ...

# Intérêt

- ▶ la descente par coordonnée peut être particulièrement utile quand  $p$  est grand
- ▶ parcours possible : cyclique, aléatoire, avec/sans remise, etc.
- ▶ on peut faire le même raisonnement par bloc : on ne met plus à jour une coordonnée, mais tout un groupe/bloc ( : *Block Coordinate Descent*)
- ▶ mathématiquement cet algorithme converge vers un minimum (sous certaines conditions : fonction lisse, ou bien non lisse mais séparable cf. Tseng (2001)), cf. la suite

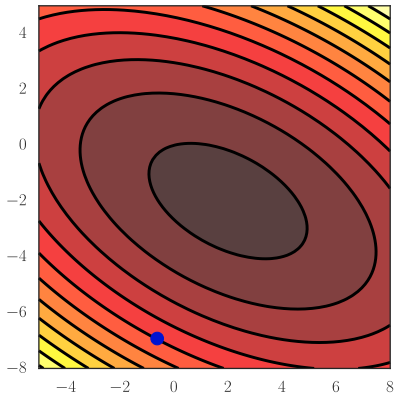
# Motivation (cas convexe)

- Convergence vers un minimum pour des fonctions lisses  
cf. Tseng (2001)



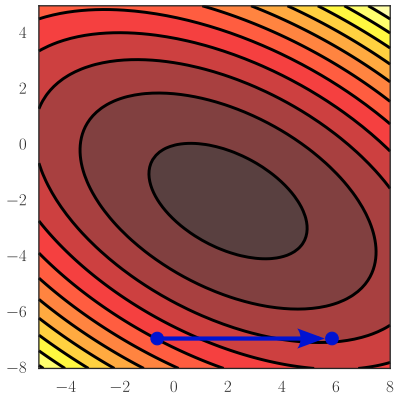
# Motivation (cas convexe)

- Convergence vers un minimum pour des fonctions lisses  
cf. Tseng (2001)



# Motivation (cas convexe)

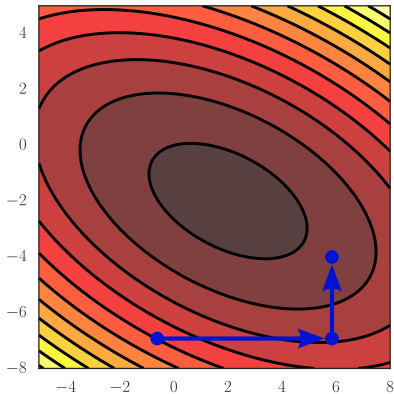
- Convergence vers un minimum pour des fonctions lisses  
cf. Tseng (2001)





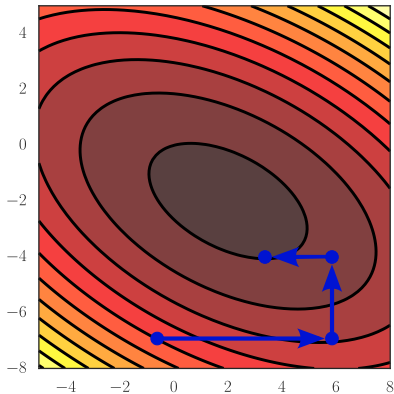
# Motivation (cas convexe)

- Convergence vers un minimum pour des fonctions lisses  
*cf.* Tseng (2001)



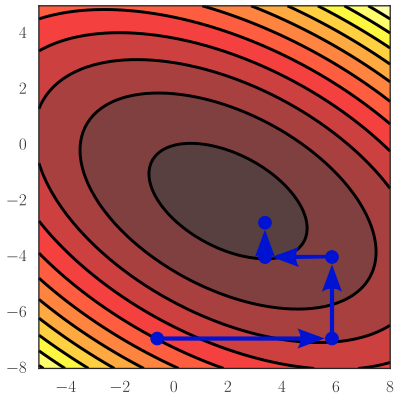
# Motivation (cas convexe)

- Convergence vers un minimum pour des fonctions lisses  
cf. Tseng (2001)



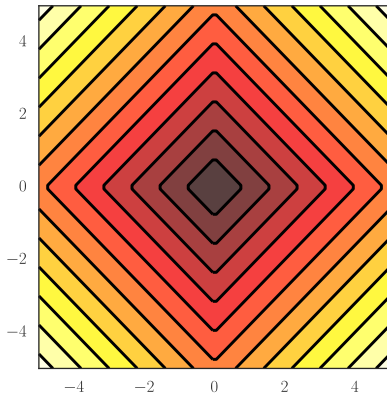
# Motivation (cas convexe)

- Convergence vers un minimum pour des fonctions lisses  
cf. Tseng (2001)



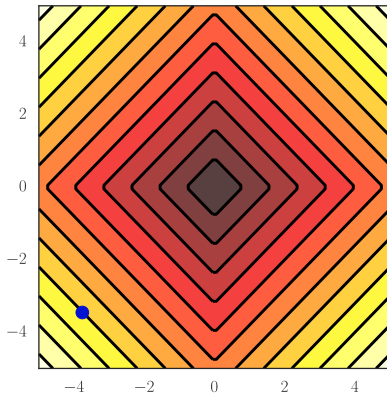
## Motivation (cas convexe)

- Convergence vers un minimum pour des fonctions non-lisses séparables cf. Tseng (2001)



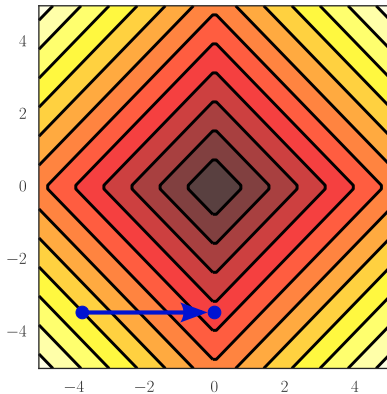
## Motivation (cas convexe)

- Convergence vers un minimum pour des fonctions non-lisses séparables cf. Tseng (2001)



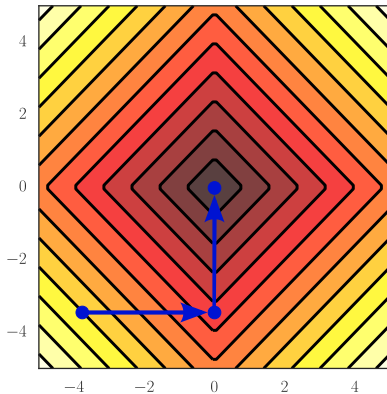
# Motivation (cas convexe)

- Convergence vers un minimum pour des fonctions non-lisses séparables cf. Tseng (2001)



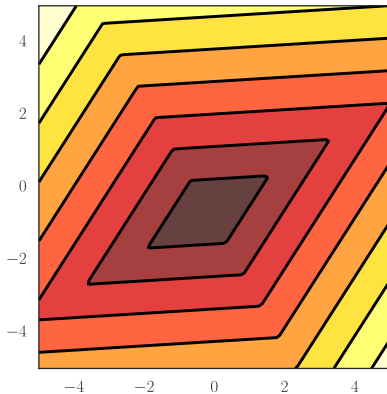
# Motivation (cas convexe)

- Convergence vers un minimum pour des fonctions non-lisses séparables cf. Tseng (2001)



## Motivation (cas convexe)

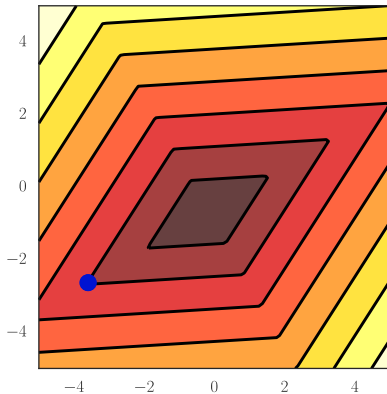
- Attention : pas (toujours) convergence vers un minimum pour des fonctions non-séparables/non-lisses





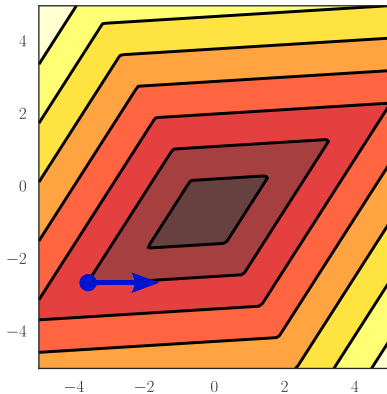
## Motivation (cas convexe)

- Attention : pas (toujours) convergence vers un minimum pour des fonctions non-séparables/non-lisses



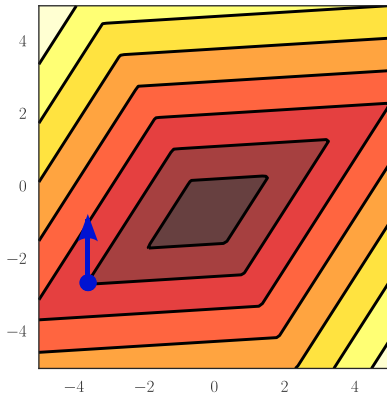
## Motivation (cas convexe)

- Attention : pas (toujours) convergence vers un minimum pour des fonctions non-séparables/non-lisses



## Motivation (cas convexe)

- Attention : pas (toujours) convergence vers un minimum pour des fonctions non-séparables/non-lisses



# Moindre carrés

$$\arg \min_{\boldsymbol{\theta} \in \mathbb{R}^p} f(\boldsymbol{\theta}) \text{ pour } f(\boldsymbol{\theta}) = \frac{1}{2} \|\mathbf{y} - X\boldsymbol{\theta}\|_2^2 = \frac{1}{2} \sum_{i=1}^n (y_i - \sum_{j=1}^p \theta_j X_{i,j})^2$$

$$\text{Rappel : } \nabla f(\boldsymbol{\theta}) = X^\top (X\boldsymbol{\theta} - \mathbf{y}) = \begin{pmatrix} \mathbf{x}_1^\top (X\boldsymbol{\theta} - \mathbf{y}) \\ \vdots \\ \mathbf{x}_p^\top (X\boldsymbol{\theta} - \mathbf{y}) \end{pmatrix} = \begin{pmatrix} \frac{\partial f}{\partial \theta_1}(\boldsymbol{\theta}) \\ \vdots \\ \frac{\partial f}{\partial \theta_p}(\boldsymbol{\theta}) \end{pmatrix}$$

Minimiser en  $\theta_j$  avec les autres  $\theta_k$ , pour  $k \neq j$ , fixes :

$$\begin{aligned} 0 &= \frac{\partial f}{\partial \theta_j}(\boldsymbol{\theta}) = \mathbf{x}_j^\top (X\boldsymbol{\theta} - \mathbf{y}) = \mathbf{x}_j^\top \left( \mathbf{x}_j \theta_j + \sum_{k \neq j} \mathbf{x}_k \theta_k - \mathbf{y} \right) \\ \Leftrightarrow \theta_j &= \frac{\mathbf{x}_j^\top (\mathbf{y} - \sum_{k \neq j} \mathbf{x}_k \theta_k)}{\mathbf{x}_j^\top \mathbf{x}_j} = \frac{\mathbf{x}_j^\top (\mathbf{y} - \sum_{k=1}^p \mathbf{x}_k \theta_k + \mathbf{x}_j \theta_j)}{\|\mathbf{x}_j\|_2^2} \end{aligned}$$

# CD pour les moindres carrés

Mise à jour intelligente : stocker les résidus courants  $r^{(k)}$  et les coefficients dans  $\theta^{(k)}$

# CD pour les moindres carrés

Mise à jour intelligente : stocker les résidus courants  $r^{(k)}$  et les coefficients dans  $\theta^{(k)}$

Pour chaque  $j \in \llbracket 1, p \rrbracket$ , faire :

# CD pour les moindres carrés

Mise à jour intelligente : stocker les résidus courants  $r^{(k)}$  et les coefficients dans  $\theta^{(k)}$

$$r^{\text{int}} \leftarrow r^{(k)} + \mathbf{x}_j \theta_j^{(k)}$$

Pour chaque  $j \in \llbracket 1, p \rrbracket$ , faire :

# CD pour les moindres carrés

Mise à jour intelligente : stocker les résidus courants  $r^{(k)}$  et les coefficients dans  $\theta^{(k)}$

$$r^{\text{int}} \leftarrow r^{(k)} + \mathbf{x}_j \theta_j^{(k)}$$

Pour chaque  $j \in \llbracket 1, p \rrbracket$ , faire :  $\theta_j^{(k+1)} \leftarrow \mathbf{x}_j^\top r^{\text{int}} / \|\mathbf{x}_j\|_2^2$



# CD pour les moindres carrés

Mise à jour intelligente : stocker les résidus courants  $r^{(k)}$  et les coefficients dans  $\theta^{(k)}$

$$r^{\text{int}} \leftarrow r^{(k)} + \mathbf{x}_j \theta_j^{(k)}$$

Pour chaque  $j \in \llbracket 1, p \rrbracket$ , faire :  $\theta_j^{(k+1)} \leftarrow \mathbf{x}_j^\top r^{\text{int}} / \|\mathbf{x}_j\|_2^2$

$$r^{(k+1)} \leftarrow r^{\text{int}} - \mathbf{x}_j \theta_j^{(k+1)}$$

# CD pour les moindres carrés

Mise à jour intelligente : stocker les résidus courants  $r^{(k)}$  et les coefficients dans  $\theta^{(k)}$

$$r^{\text{int}} \leftarrow r^{(k)} + \mathbf{x}_j \theta_j^{(k)}$$

Pour chaque  $j \in \llbracket 1, p \rrbracket$ , faire :  $\theta_j^{(k+1)} \leftarrow \mathbf{x}_j^\top r^{\text{int}} / \|\mathbf{x}_j\|_2^2$

$$r^{(k+1)} \leftarrow r^{\text{int}} - \mathbf{x}_j \theta_j^{(k+1)}$$

Impact mémoire faible :

- ▶ stocker un vecteur de résidus de taille  $n$
- ▶ stocker un vecteur d'estimation de taille  $p$

Rem:  $\|\mathbf{x}_j\|_2^2 = 1$  utile en optimisation ( $\neq$  en statistique)

## Ridge : descente par coordonnée

$$\arg \min_{\boldsymbol{\theta} \in \mathbb{R}^p} f(\boldsymbol{\theta}) \text{ pour } f(\boldsymbol{\theta}) = \frac{1}{2} \sum_{i=1}^n (y_i - \sum_{j=1}^p \theta_j X_{i,j})^2 + \frac{\lambda}{2} \sum_{j=1}^p \theta_j^2$$

$$\nabla f(\boldsymbol{\theta}) = X^\top (X\boldsymbol{\theta} - \mathbf{y}) + \lambda \boldsymbol{\theta} = \begin{pmatrix} \mathbf{x}_1^\top (X\boldsymbol{\theta} - \mathbf{y}) + \lambda \theta_1 \\ \vdots \\ \mathbf{x}_p^\top (X\boldsymbol{\theta} - \mathbf{y}) + \lambda \theta_p \end{pmatrix} = \begin{pmatrix} \frac{\partial f}{\partial \theta_1}(\boldsymbol{\theta}) \\ \vdots \\ \frac{\partial f}{\partial \theta_p}(\boldsymbol{\theta}) \end{pmatrix}$$

Minimise en  $\theta_j$  avec les autres  $\theta_k$  ( $k \neq j$ ) fixes

$$0 = \frac{\partial f}{\partial \theta_j}(\boldsymbol{\theta}) = \mathbf{x}_j^\top (X\boldsymbol{\theta} - \mathbf{y}) + \lambda \theta_j = \mathbf{x}_j^\top \left( \mathbf{x}_j \theta_j + \sum_{k \neq j} \mathbf{x}_k \theta_k - \mathbf{y} \right) + \lambda \theta_j$$

$$\Leftrightarrow \theta_j = \frac{\mathbf{x}_j^\top \left( \mathbf{y} - \sum_{k \neq j} \mathbf{x}_k \theta_k \right)}{\mathbf{x}_j^\top \mathbf{x}_j + \lambda} = \frac{\mathbf{x}_j^\top \left( \mathbf{y} - \sum_{k=1}^p \mathbf{x}_k \theta_k + \mathbf{x}_j \theta_j \right)}{\|\mathbf{x}_j\|_2^2 + \lambda}$$

## Ridge descent par coordonnée (II)

Mise à jour intelligente : stocker les résidus courants  $r^{(k)}$  et les coefficients dans  $\theta^{(k)}$

## Ridge descent par coordonnée (II)

Mise à jour intelligente : stocker les résidus courants  $r^{(k)}$  et les coefficients dans  $\theta^{(k)}$

Pour chaque  $j \in \llbracket 1, p \rrbracket$ , faire :

## Ridge descent par coordonnée (II)

Mise à jour intelligente : stocker les résidus courants  $r^{(k)}$  et les coefficients dans  $\boldsymbol{\theta}^{(k)}$

$$r^{\text{int}} \leftarrow r^{(k)} + \mathbf{x}_j \theta_j^{(k)}$$

Pour chaque  $j \in \llbracket 1, p \rrbracket$ , faire :

## Ridge descent par coordonnée (II)

Mise à jour intelligente : stocker les résidus courants  $r^{(k)}$  et les coefficients dans  $\theta^{(k)}$

$$r^{\text{int}} \leftarrow r^{(k)} + \mathbf{x}_j \theta_j^{(k)}$$

Pour chaque  $j \in \llbracket 1, p \rrbracket$ , faire :  $\theta_j^{(k+1)} \leftarrow \mathbf{x}_j^\top r^{\text{int}} / (\|\mathbf{x}_j\|_2^2 + \lambda)$

## Ridge descent par coordonnée (II)

Mise à jour intelligente : stocker les résidus courants  $r^{(k)}$  et les coefficients dans  $\theta^{(k)}$

$$r^{\text{int}} \leftarrow r^{(k)} + \mathbf{x}_j \theta_j^{(k)}$$

Pour chaque  $j \in \llbracket 1, p \rrbracket$ , faire :  $\theta_j^{(k+1)} \leftarrow \mathbf{x}_j^\top r^{\text{int}} / (\|\mathbf{x}_j\|_2^2 + \lambda)$

$$r^{(k+1)} \leftarrow r^{\text{int}} - \mathbf{x}_j \theta_j^{(k+1)}$$



## Ridge descent par coordonnée (II)

Mise à jour intelligente : stocker les résidus courants  $r^{(k)}$  et les coefficients dans  $\theta^{(k)}$

$$r^{\text{int}} \leftarrow r^{(k)} + \mathbf{x}_j \theta_j^{(k)}$$

Pour chaque  $j \in \llbracket 1, p \rrbracket$ , faire :  $\theta_j^{(k+1)} \leftarrow \mathbf{x}_j^\top r^{\text{int}} / (\|\mathbf{x}_j\|_2^2 + \lambda)$

$$r^{(k+1)} \leftarrow r^{\text{int}} - \mathbf{x}_j \theta_j^{(k+1)}$$

Impact mémoire faible :

- ▶ stocker un vecteur de résidus de taille  $n$
- ▶ stocker un vecteur d'estimation de taille  $p$

Rem:  $\|\mathbf{x}_j\|_2^2 = 1$  utile en optimisation ( $\neq$  en statistique)

## Lasso : descente par coordonnée

$$\arg \min_{\boldsymbol{\theta} \in \mathbb{R}^p} f(\boldsymbol{\theta}) \text{ pour } f(\boldsymbol{\theta}) = \frac{1}{2} \|\mathbf{y} - X\boldsymbol{\theta}\|^2 + \lambda \sum_{j=1}^p |\theta_j|$$

Minimise en  $\theta_j$  avec les autres  $\theta_k$  ( $k \neq j$ ) fixes

$$\hat{\theta}_j = \arg \min_{\theta_j \in \mathbb{R}} f(\theta_1, \dots, \theta_p)$$

$$= \arg \min_{\theta_j \in \mathbb{R}} \frac{1}{2} \|\mathbf{y} - \sum_{k \neq j} \theta_k \mathbf{x}_k - \mathbf{x}_j \theta_j\|^2 + \lambda \sum_{k \neq j} |\theta_k| + \lambda |\theta_j|$$

$$= \arg \min_{\theta_j \in \mathbb{R}} \frac{1}{2} \|\mathbf{x}_j\|^2 \theta_j^2 - \langle y - \sum_{k \neq j} \theta_k \mathbf{x}_k, \mathbf{x}_j \rangle \theta_j + \lambda |\theta_j|$$

$$= \arg \min_{\theta_j \in \mathbb{R}} \|\mathbf{x}_j\|^2 \left[ \frac{1}{2} \left( \theta_j - \|\mathbf{x}_j\|^{-2} \langle y - \sum_{k \neq j} \theta_k \mathbf{x}_k, \mathbf{x}_j \rangle \right)^2 + \frac{\lambda}{\|\mathbf{x}_j\|^2} |\theta_j| \right]$$

Rappel :  $\eta_{\text{ST}, \lambda}(z) = \arg \min_{t \in \mathbb{R}} \frac{1}{2} (z - t)^2 + \lambda |t|$

## Lasso : descente par coordonnée (II)

Solution :

$$\hat{\theta}_j = \eta_{\text{ST}, \lambda / \|\mathbf{x}_j\|^2} \left( \|\mathbf{x}_j\|^{-2} \left\langle y - \sum_{k \neq j} \theta_k \mathbf{x}_k, \mathbf{x}_j \right\rangle \right)$$

Mise à jour intelligente : stocker les résidus courants  $r^{(k)}$  et les coefficients dans  $\boldsymbol{\theta}^{(k)}$

## Lasso : descente par coordonnée (II)

Solution :

$$\hat{\theta}_j = \eta_{\text{ST}, \lambda / \|\mathbf{x}_j\|^2} \left( \|\mathbf{x}_j\|^{-2} \left\langle y - \sum_{k \neq j} \theta_k \mathbf{x}_k, \mathbf{x}_j \right\rangle \right)$$

Mise à jour intelligente : stocker les résidus courants  $r^{(k)}$  et les coefficients dans  $\boldsymbol{\theta}^{(k)}$

Pour chaque  $j \in \llbracket 1, p \rrbracket$ , faire :

## Lasso : descente par coordonnée (II)

Solution :

$$\hat{\theta}_j = \eta_{\text{ST}, \lambda / \|\mathbf{x}_j\|^2} \left( \|\mathbf{x}_j\|^{-2} \langle y - \sum_{k \neq j} \theta_k \mathbf{x}_k, \mathbf{x}_j \rangle \right)$$

Mise à jour intelligente : stocker les résidus courants  $r^{(k)}$  et les coefficients dans  $\boldsymbol{\theta}^{(k)}$

$$r^{\text{int}} \leftarrow r^{(k)} + \mathbf{x}_j \theta_j^{(k)}$$

Pour chaque  $j \in \llbracket 1, p \rrbracket$ , faire :

## Lasso : descente par coordonnée (II)

Solution :

$$\hat{\theta}_j = \eta_{\text{ST}, \lambda / \|\mathbf{x}_j\|^2} \left( \|\mathbf{x}_j\|^{-2} \langle y - \sum_{k \neq j} \theta_k \mathbf{x}_k, \mathbf{x}_j \rangle \right)$$

Mise à jour intelligente : stocker les résidus courants  $r^{(k)}$  et les coefficients dans  $\boldsymbol{\theta}^{(k)}$

$$r^{\text{int}} \leftarrow r^{(k)} + \mathbf{x}_j \theta_j^{(k)}$$

Pour chaque  $j \in \llbracket 1, p \rrbracket$ , faire :  $\theta_j^{(k+1)} \leftarrow \eta_{\text{ST}, \lambda / \|\mathbf{x}_j\|^2} (\mathbf{x}_j^\top r^{\text{int}} / \|\mathbf{x}_j\|^2)$

## Lasso : descente par coordonnée (II)

Solution :

$$\hat{\theta}_j = \eta_{\text{ST}, \lambda / \|\mathbf{x}_j\|^2} \left( \|\mathbf{x}_j\|^{-2} \langle y - \sum_{k \neq j} \theta_k \mathbf{x}_k, \mathbf{x}_j \rangle \right)$$

Mise à jour intelligente : stocker les résidus courants  $r^{(k)}$  et les coefficients dans  $\boldsymbol{\theta}^{(k)}$

$$r^{\text{int}} \leftarrow r^{(k)} + \mathbf{x}_j \theta_j^{(k)}$$

Pour chaque  $j \in \llbracket 1, p \rrbracket$ , faire :

$$\theta_j^{(k+1)} \leftarrow \eta_{\text{ST}, \lambda / \|\mathbf{x}_j\|^2} (\mathbf{x}_j^\top r^{\text{int}} / \|\mathbf{x}_j\|^2)$$

$$r^{(k+1)} \leftarrow r^{\text{int}} - \mathbf{x}_j \theta_j^{(k+1)}$$

## Lasso : descente par coordonnée (II)

Solution :

$$\hat{\theta}_j = \eta_{\text{ST}, \lambda / \|\mathbf{x}_j\|^2} \left( \|\mathbf{x}_j\|^{-2} \langle y - \sum_{k \neq j} \theta_k \mathbf{x}_k, \mathbf{x}_j \rangle \right)$$

Mise à jour intelligente : stocker les résidus courants  $r^{(k)}$  et les coefficients dans  $\theta^{(k)}$

$$r^{\text{int}} \leftarrow r^{(k)} + \mathbf{x}_j \theta_j^{(k)}$$

Pour chaque  $j \in \llbracket 1, p \rrbracket$ , faire :

$$\theta_j^{(k+1)} \leftarrow \eta_{\text{ST}, \lambda / \|\mathbf{x}_j\|^2} (\mathbf{x}_j^\top r^{\text{int}} / \|\mathbf{x}_j\|^2)$$

$$r^{(k+1)} \leftarrow r^{\text{int}} - \mathbf{x}_j \theta_j^{(k+1)}$$

Impact mémoire faible :

- ▶ stocker un vecteur de résidus de taille  $n$
- ▶ stocker un vecteur d'estimation de taille  $p$

Rem:  $\|\mathbf{x}_j\|_2^2 = 1$  utile en optimisation ( $\neq$  en statistique)



## Descente par coordonnée : cas non-convexe

$$\hat{\boldsymbol{\theta}}_{\lambda,\gamma}^{\text{pen}} = \arg \min_{\boldsymbol{\theta} \in \mathbb{R}^p} \left( \underbrace{\frac{1}{2} \|\mathbf{y} - X\boldsymbol{\theta}\|_2^2}_{\text{attache aux données}} + \underbrace{\sum_{j=1}^p \text{pen}_{\lambda,\gamma}(|\theta_j|)}_{\text{régularisation}} \right)$$

Même approche (mais sans garantie de convergence), si  $\|\mathbf{x}_j\|_2^2 = 1$

## Descente par coordonnée : cas non-convexe

$$\hat{\boldsymbol{\theta}}_{\lambda,\gamma}^{\text{pen}} = \arg \min_{\boldsymbol{\theta} \in \mathbb{R}^p} \left( \underbrace{\frac{1}{2} \|\mathbf{y} - X\boldsymbol{\theta}\|_2^2}_{\text{attache aux données}} + \underbrace{\sum_{j=1}^p \text{pen}_{\lambda,\gamma}(|\theta_j|)}_{\text{régularisation}} \right)$$

Même approche (mais sans garantie de convergence), si  $\|\mathbf{x}_j\|_2^2 = 1$

Pour chaque  $j \in \llbracket 1, p \rrbracket$ , faire :

## Descente par coordonnée : cas non-convexe

$$\hat{\boldsymbol{\theta}}_{\lambda,\gamma}^{\text{pen}} = \arg \min_{\boldsymbol{\theta} \in \mathbb{R}^p} \left( \underbrace{\frac{1}{2} \|\mathbf{y} - X\boldsymbol{\theta}\|_2^2}_{\text{attache aux données}} + \underbrace{\sum_{j=1}^p \text{pen}_{\lambda,\gamma}(|\theta_j|)}_{\text{régularisation}} \right)$$

Même approche (mais sans garantie de convergence), si  $\|\mathbf{x}_j\|_2^2 = 1$

$$r^{\text{int}} \leftarrow r^{(k)} + \mathbf{x}_j \theta_j^{(k)}$$

Pour chaque  $j \in \llbracket 1, p \rrbracket$ , faire :

## Descente par coordonnée : cas non-convexe

$$\hat{\boldsymbol{\theta}}_{\lambda, \gamma}^{\text{pen}} = \arg \min_{\boldsymbol{\theta} \in \mathbb{R}^p} \left( \underbrace{\frac{1}{2} \|\mathbf{y} - X\boldsymbol{\theta}\|_2^2}_{\text{attache aux données}} + \underbrace{\sum_{j=1}^p \text{pen}_{\lambda, \gamma}(|\theta_j|)}_{\text{régularisation}} \right)$$

Même approche (mais sans garantie de convergence), si  $\|\mathbf{x}_j\|_2^2 = 1$

$$r^{\text{int}} \leftarrow r^{(k)} + \mathbf{x}_j \theta_j^{(k)}$$

Pour chaque  $j \in \llbracket 1, p \rrbracket$ , faire :  $\theta_j^{(k+1)} \leftarrow \eta_{\text{pen}_{\lambda, \gamma}}(\mathbf{x}_j^\top r^{\text{int}})$

## Descente par coordonnée : cas non-convexe

$$\hat{\boldsymbol{\theta}}_{\lambda,\gamma}^{\text{pen}} = \arg \min_{\boldsymbol{\theta} \in \mathbb{R}^p} \left( \underbrace{\frac{1}{2} \|\mathbf{y} - X\boldsymbol{\theta}\|_2^2}_{\text{attache aux données}} + \underbrace{\sum_{j=1}^p \text{pen}_{\lambda,\gamma}(|\theta_j|)}_{\text{régularisation}} \right)$$

Même approche (mais sans garantie de convergence), si  $\|\mathbf{x}_j\|_2^2 = 1$

$$r^{\text{int}} \leftarrow r^{(k)} + \mathbf{x}_j \theta_j^{(k)}$$

Pour chaque  $j \in \llbracket 1, p \rrbracket$ , faire :  $\theta_j^{(k+1)} \leftarrow \eta_{\text{pen}_{\lambda,\gamma}}(\mathbf{x}_j^\top r^{\text{int}})$

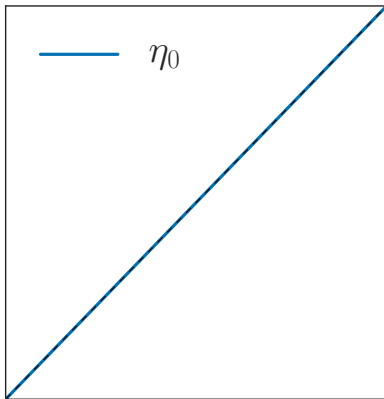
$$r^{(k+1)} \leftarrow r^{\text{int}} - \mathbf{x}_j \theta_j^{(k+1)}$$

$$\text{où } \eta_{\text{pen}_{\lambda,\gamma}}(z) = \arg \min_{t \in \mathbb{R}} \frac{1}{2}(z - t)^2 + \text{pen}_{\lambda,\gamma}(t)$$

voir par exemple [Breheny et Huang \(2011\)](#)

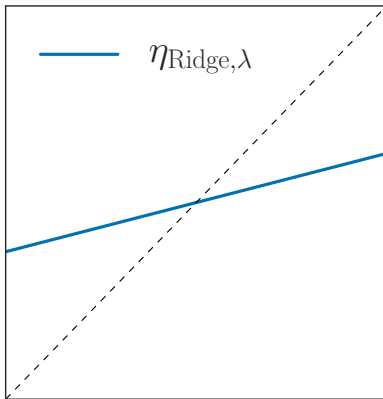
# Régularisation en 1D : Aucune

$$\eta_0(z) = z$$



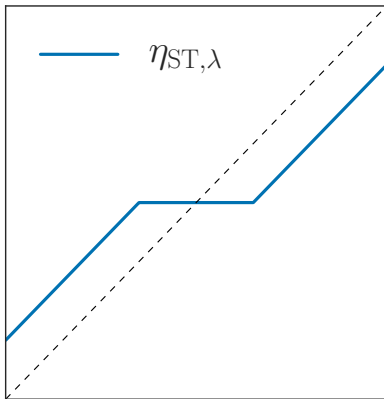
# Régularisation en 1D : Ridge

$$\eta_{\text{Ridge},\lambda}(z) = \frac{z}{1 + \lambda}$$



# Régularisation en 1D : Lasso

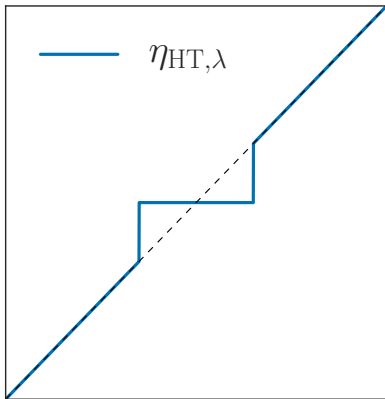
$$\eta_{\text{ST},\lambda}(z) = \text{sign}(z)(|z| - \lambda)_+$$





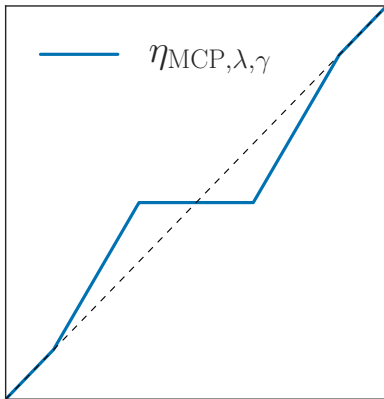
# Régularisation en 1D : $\ell_0$

$$\eta_{\text{HT}, \lambda^2/2}(z) = z \mathbb{1}_{|z| \geq \lambda}$$



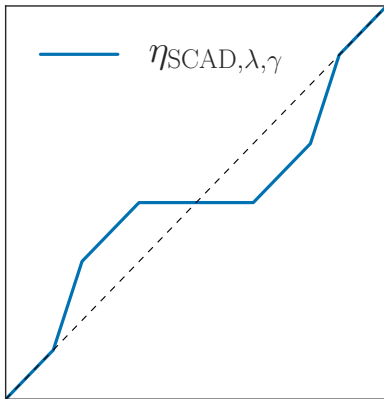
# Régularisation en 1D : MCP

$$\eta_{\text{MCP},\lambda,\gamma}(z) = \begin{cases} \text{sign}(z)(|z| - \lambda)_+ / (1 - 1/\gamma) & \text{si } |z| \leq \gamma\lambda \\ z & \text{si } |z| > \gamma\lambda \end{cases}$$



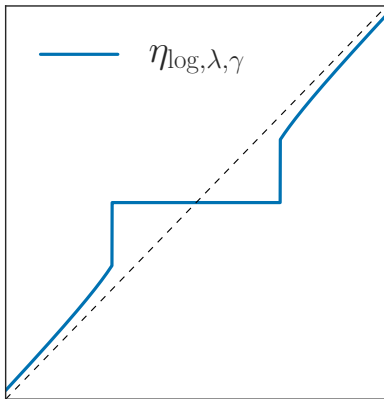
## Régularisation en 1D : SCAD

$$\eta_{\text{SCAD},\lambda,\gamma}(z) = \begin{cases} \text{sign}(z)(|z| - \lambda)_+ / (1 - 1/\gamma) & \text{si } |z| \leq 2\lambda \\ ([\gamma - 1]z - \text{sign}(z)\gamma\lambda) / (\gamma - 2) & \text{si } 2\lambda \leq |z| \leq \gamma\lambda \\ z & \text{si } |z| > \gamma\lambda \end{cases}$$



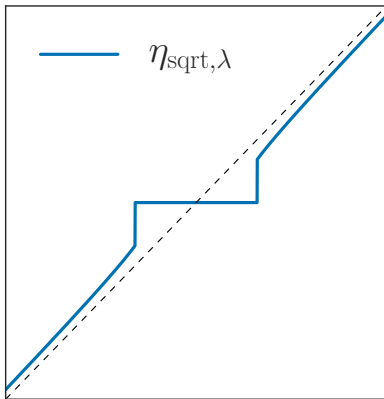
# Régularisation en 1D : $\log$

$$\eta_{\log, \lambda}(z) = \dots$$

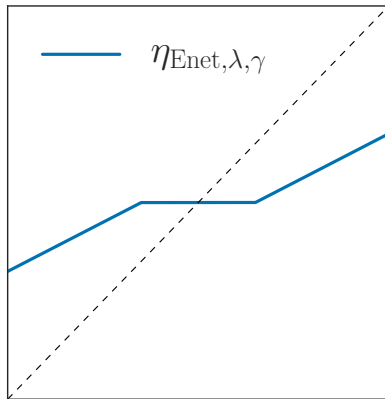


# Régularisation en 1D : sqrt

$$\eta_{\text{sqrt},\lambda}(z) = \dots$$



$$\eta_{Enet,\lambda,\gamma}(z) = \dots$$



# Lasso-Positif

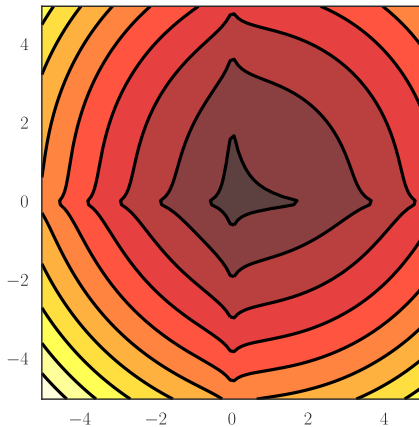
---

**Exo:** Proposer une manière de résoudre le problème Lasso avec une contrainte de positivité sur les coefficients

$$\hat{\boldsymbol{\theta}}_{\lambda}^{\text{Lasso}+} = \arg \min_{\boldsymbol{\theta} \in \mathbb{R}_+^p} \left( \underbrace{\frac{1}{2} \|\mathbf{y} - X\boldsymbol{\theta}\|_2^2}_{\text{attache aux données}} + \underbrace{\lambda \|\boldsymbol{\theta}\|_1}_{\text{régularisation}} \right)$$

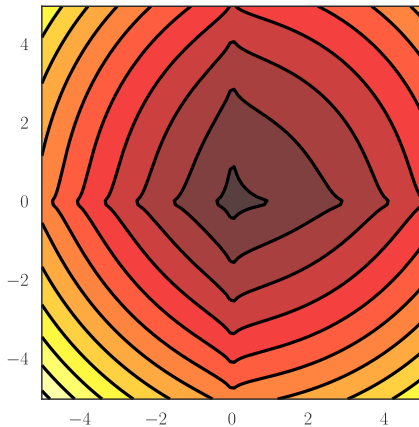
---

# Lignes de niveaux pour log





# Lignes de niveaux pour sqrt



# Sommaire


Optimisation pour le Lasso

Descente par coordonnée

Alternatives

# Optimisation : autres méthodes

D'autres algorithmes peuvent être utilisés pour construire une solution approchée du Lasso :

- ▶ LARS Efron *et al.* (2004) pour le chemin entier
- ▶ méthodes de gradient proximal, Forward-Backward, de type Seuillage Doux Itératif ( : *ISTA*, *FISTA*), cf. Beck et Teboulle(2009)

Ces dernières méthodes seront vues en INFMDI 341

# Références I

- ▶ P. Breheny and J. Huang.  
Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection.  
*5(1)* :232, 2011.
- ▶ A. Beck and M. Teboulle.  
A fast iterative shrinkage-thresholding algorithm for linear inverse problems.  
*SIAM J. Imaging Sci.*, 2(1) :183–202, 2009.
- ▶ B. Efron, T. Hastie, I. M. Johnstone, and R. Tibshirani.  
Least angle regression.  
*Ann. Statist.*, 32(2) :407–499, 2004.  
With discussion, and a rejoinder by the authors.
- ▶ P. Tseng.  
Convergence of a block coordinate descent method for nondifferentiable minimization.  
*J. Optim. Theory Appl.*, 109(3) :475–494, 2001.