SD204
Descente par coordonnée

Joseph Salmon
http://josephsalmon.eu
Télécom Paristech, Institut Mines-Télécom

Plan

Sommaire

Descente par coordonnée

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

Algorithme : Descente par coordonnée

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

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

Descente par coordonnée

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

Algorithme : Descente par coordonnée

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

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

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

Descente par coordonnée

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

Algorithme : Descente par coordonnée

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

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

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

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

Descente par coordonnée

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

Algorithme : Descente par coordonnée

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

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

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

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

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

Descente par coordonnée

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

Algorithme : Descente par coordonnée

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

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

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

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

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

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

Descente par coordonnée

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

Algorithme : Descente par coordonnée

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

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

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

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

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

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

$\vdots$

Descente par coordonnée

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

Algorithme : Descente par coordonnée

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

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

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

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

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

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

$$\vdots$$

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

Descente par coordonnée

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

Algorithme : Descente par coordonnée

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

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

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

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

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

$$\theta_3^{(k)} \leftarrow \underset{\theta_3 \in \mathbb{R}}{\arg\min}\, f(\theta^{(k)}, \theta_2^{(k)}, \theta_3, \ldots, \theta_{p-1}^{(k-1)}, \theta_p^{(k-1)})$$

$$\vdots$$

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

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

Descente par coordonnée

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

Algorithme : Descente par coordonnée

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

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

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

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

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

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

$$\vdots$$

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

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

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

Parcours possibles

On doit visiter toutes les coordonnées **régulièrement** pour assurer la convergence. Les parcours les plus courants sont les suivants :

- parcours cyclique (Gauss-Seidel)

Parcours possibles

On doit visiter toutes les coordonnées **régulièrement** pour assurer la convergence. Les parcours les plus courants sont les suivants :

- parcours cyclique (Gauss-Seidel)
- parcours aléatoire avec remise : on tire de manière uniforme *i.i.d.* les coordonnées à mettre à jour

Parcours possibles

On doit visiter toutes les coordonnées **régulièrement** pour assurer la convergence. Les parcours les plus courants sont les suivants :

- parcours cyclique (Gauss-Seidel)
- parcours aléatoire avec remise : on tire de manière uniforme *i.i.d.* les coordonnées à mettre à jour
- parcours aléatoire sans remise : on tire de manière *i.i.d.* une permutation aléatoire des coordonnées (: *shuffle*) après chaque époque

Parcours possibles

On doit visiter toutes les coordonnées **régulièrement** pour assurer la convergence. Les parcours les plus courants sont les suivants :

- parcours cyclique (Gauss-Seidel)
- parcours aléatoire avec remise : on tire de manière uniforme *i.i.d.* les coordonnées à mettre à jour
- parcours aléatoire sans remise : on tire de manière *i.i.d.* une permutation aléatoire des coordonnées (: *shuffle*) après chaque époque
- parcours glouton (Gauss-Southwell) : on cherche de manière itérative la coordonnée la meilleure (*e.g.*, celle qui fait le plus bouger ou qui diminue le plus la fonction objective)

Intérêt de la descente par coordonnée

- utilité quand p est (très) grand
- stratégie par “bloc” : on met à jour tout un groupe/bloc (: *Block Coordinate Descent*) de coordonnées
- la convergence est assurée vers un minimum pour les cas :

  1. Fonction lisse :
  $$\arg\min_{\theta} f(\theta)$$
  avec f convexe différentiable
  2. Fonction lisse + fonction séparable
  $$\arg\min_{\theta} f(\theta) + g(\theta)$$
  avec f convexe différentiable, et g convexe séparable : $g(\theta) = \sum_{j=1}^{p} g_j(\theta_j)$, *cf.* Tseng (2001)

Intérêt de la descente par coordonnée

- utilité quand p est (très) grand
- stratégie par "bloc" : on met à jour tout un groupe/bloc (: *Block Coordinate Descent*) de coordonnées
- la convergence est assurée vers un minimum pour les cas :

1. Fonction lisse :

$$\arg\min_{\theta} f(\theta)$$

   avec f convexe différentiable

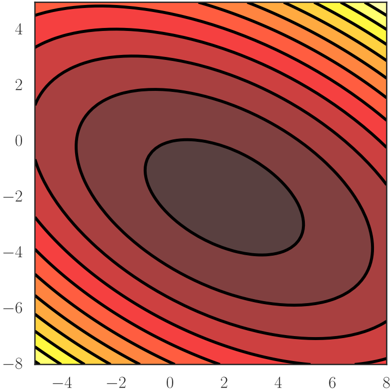
2. Fonction lisse + fonction séparable

$$\arg\min_{\theta} f(\theta) + g(\theta)$$

   avec f convexe différentiable, et g convexe séparable : $g(\theta) = \sum_{j=1}^{p} g_j(\theta_j)$, *cf.* Tseng (2001)

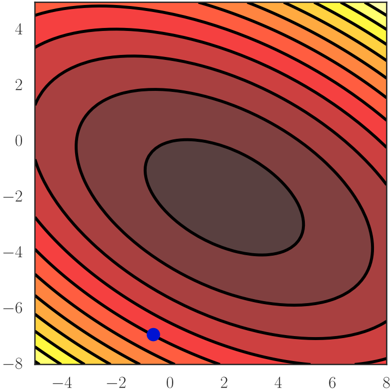
Motivation (cas convexe)

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



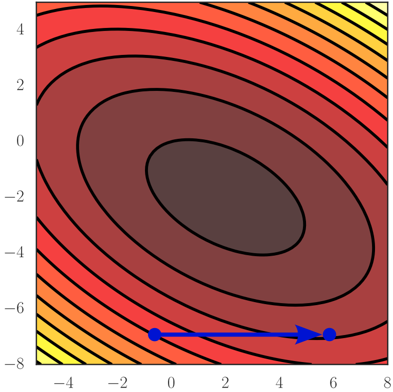
Motivation (cas convexe)

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



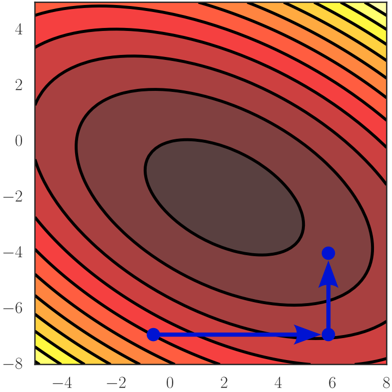
Motivation (cas convexe)

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



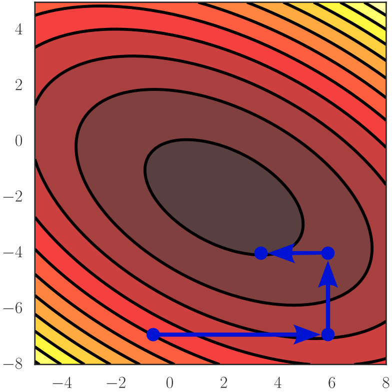
Motivation (cas convexe)

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



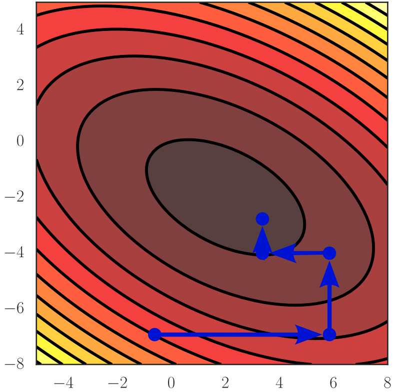
Motivation (cas convexe)

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



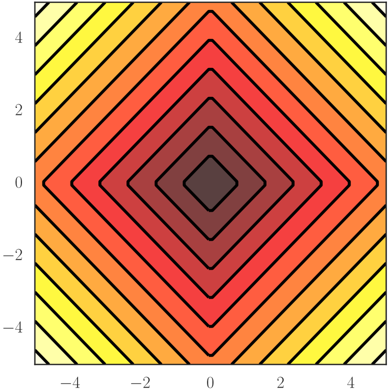
Motivation (cas convexe)

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



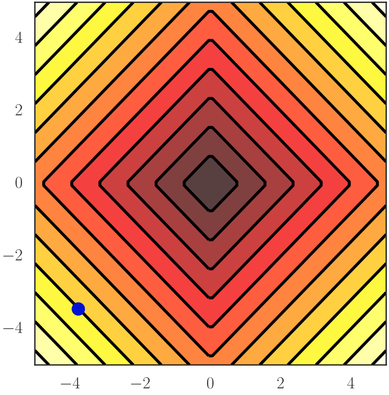
Motivation (cas convexe)

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



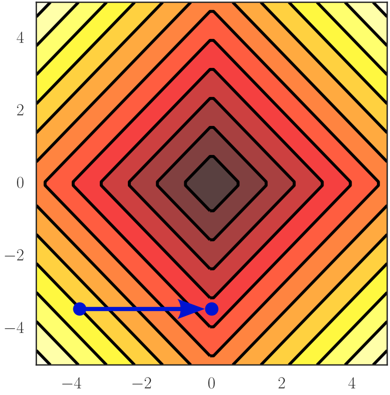
Motivation (cas convexe)

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



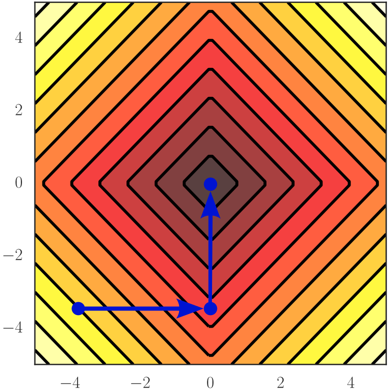
Motivation (cas convexe)

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



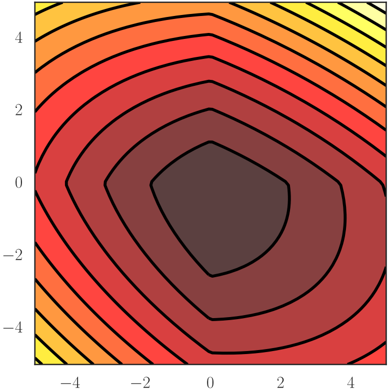
Motivation (cas convexe)

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



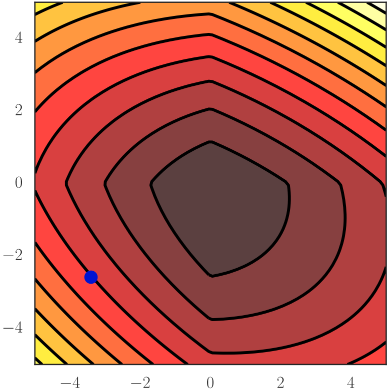
Motivation (cas convexe)

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



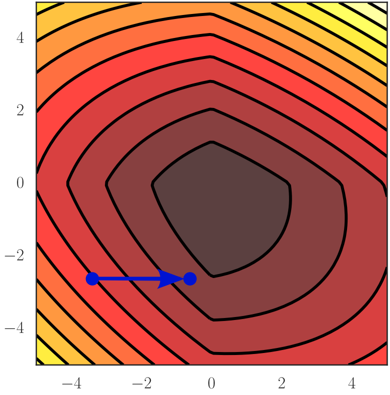
Motivation (cas convexe)

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



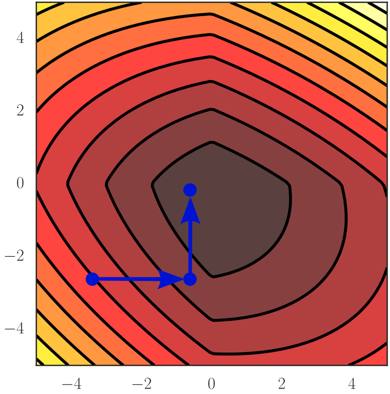
Motivation (cas convexe)

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



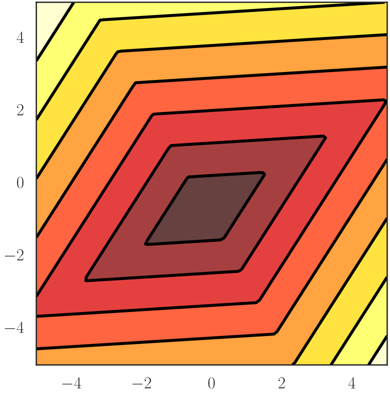
Motivation (cas convexe)

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



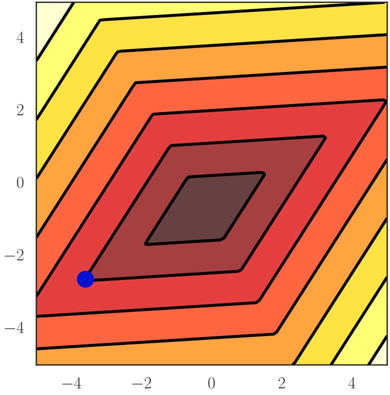
Motivation (cas convexe)

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



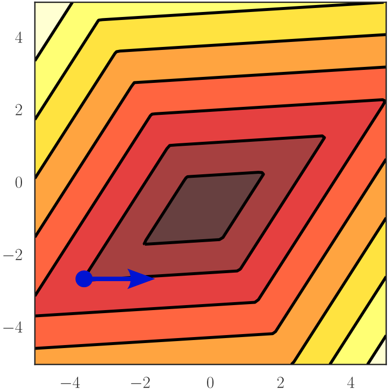
Motivation (cas convexe)

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



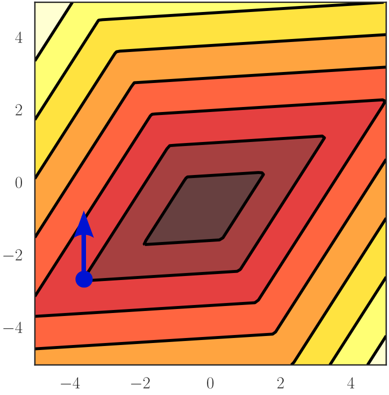
Motivation (cas convexe)

- 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



Sommaire

Moindre carrés

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

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

Moindre carrés

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

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 θ_j en fixant θ_k pour $k \neq j \Leftrightarrow$ annuler la j^{e} dérivée partielle

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

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

Moindre carrés

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

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 θ_j en fixant θ_k pour $k \neq j$ $\Leftrightarrow$ annuler la j^{e} dérivée partielle

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

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

CD pour les moindres carrés

Mise à jour intelligente : stocker à l'itération k les résidus courants $r^{(k)} = y - X\boldsymbol{\theta}^{(k)}$ et les coefficients $\boldsymbol{\theta}^{(k)}$

CD pour les moindres carrés

Mise à jour intelligente : stocker à l'itération k les résidus courants $r^{(k)} = y - X\boldsymbol{\theta}^{(k)}$ et les coefficients $\boldsymbol{\theta}^{(k)}$

Pour chaque $j \in [\![1, p]\!]$, faire :

CD pour les moindres carrés

Mise à jour intelligente : stocker à l'itération k les résidus courants $r^{(k)} = y - X\boldsymbol{\theta}^{(k)}$ et les coefficients $\boldsymbol{\theta}^{(k)}$

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

Pour chaque $j \in [\![1, p]\!]$, faire :

CD pour les moindres carrés

Mise à jour intelligente : stocker à l'itération k les résidus courants $r^{(k)} = y - X\boldsymbol{\theta}^{(k)}$ et les coefficients $\boldsymbol{\theta}^{(k)}$

$$\text{Pour chaque } j \in [\![1, p]\!], \text{ faire : } \begin{aligned} r^{\text{int}} &\leftarrow r^{(k)} + \mathbf{x}_j \theta_j^{(k)} \\ \theta_j^{(k+1)} &\leftarrow \mathbf{x}_j^\top r^{\text{int}} / \|\mathbf{x}_j\|_2^2 \end{aligned}$$

CD pour les moindres carrés

Mise à jour intelligente : stocker à l'itération k les résidus courants $r^{(k)} = y - X\boldsymbol{\theta}^{(k)}$ et les coefficients $\boldsymbol{\theta}^{(k)}$

$$\text{Pour chaque } j \in [\![1, p]\!], \text{ faire : } \begin{aligned} r^{\text{int}} &\leftarrow r^{(k)} + \mathbf{x}_j \theta_j^{(k)} \\ \theta_j^{(k+1)} &\leftarrow \mathbf{x}_j^\top r^{\text{int}} / \|\mathbf{x}_j\|_2^2 \\ r^{(k+1)} &\leftarrow r^{\text{int}} - \mathbf{x}_j \theta_j^{(k+1)} \end{aligned}$$

CD pour les moindres carrés

Mise à jour intelligente : stocker à l'itération k les résidus courants $r^{(k)} = y - X\boldsymbol{\theta}^{(k)}$ et les coefficients $\boldsymbol{\theta}^{(k)}$

$$\text{Pour chaque } j \in [\![1, p]\!], \text{ faire : } \begin{aligned} r^{\text{int}} &\leftarrow r^{(k)} + \mathbf{x}_j \theta_j^{(k)} \\ \theta_j^{(k+1)} &\leftarrow \mathbf{x}_j^\top r^{\text{int}} / \|\mathbf{x}_j\|_2^2 \\ r^{(k+1)} &\leftarrow r^{\text{int}} - \mathbf{x}_j \theta_j^{(k+1)} \end{aligned}$$

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)

Sommaire

Ridge : descente par coordonnée

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

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

Ridge : descente par coordonnée

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

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

Minimiser en θ_j en fixant θ_k pour $k \neq j$ $\Leftrightarrow$ annuler la j^{e} dérivée partielle

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

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

Ridge : descente par coordonnée

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

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

Minimiser en θ_j en fixant θ_k pour $k \neq j$ $\Leftrightarrow$ annuler la j^{e} dérivée partielle

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

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

Ridge descente par coordonnée (II)

Mise à jour intelligente : stocker à l'itération k les résidus courants $r^{(k)} = y - X\boldsymbol{\theta}^{(k)}$ et les coefficients $\boldsymbol{\theta}^{(k)}$

Ridge descente par coordonnée (II)

Mise à jour intelligente : stocker à l'itération k les résidus courants $r^{(k)} = y - X\boldsymbol{\theta}^{(k)}$ et les coefficients $\boldsymbol{\theta}^{(k)}$

Pour chaque $j \in [\![1, p]\!],$ faire :

Ridge descente par coordonnée (II)

Mise à jour intelligente : stocker à l'itération k les résidus courants $r^{(k)} = y - X\boldsymbol{\theta}^{(k)}$ et les coefficients $\boldsymbol{\theta}^{(k)}$

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

Pour chaque $j \in [\![1, p]\!]$, faire :

Ridge descente par coordonnée (II)

Mise à jour intelligente : stocker à l'itération k les résidus courants $r^{(k)} = y - X\boldsymbol{\theta}^{(k)}$ et les coefficients $\boldsymbol{\theta}^{(k)}$

$$\text{Pour chaque } j \in [\![1, p]\!], \text{ faire : } \begin{aligned} r^{\text{int}} &\leftarrow r^{(k)} + \mathbf{x}_j \theta_j^{(k)} \\ \theta_j^{(k+1)} &\leftarrow \mathbf{x}_j^\top r^{\text{int}} / (\|\mathbf{x}_j\|_2^2 + \lambda) \end{aligned}$$

Ridge descente par coordonnée (II)

Mise à jour intelligente : stocker à l'itération k les résidus courants $r^{(k)} = y - X\boldsymbol{\theta}^{(k)}$ et les coefficients $\boldsymbol{\theta}^{(k)}$

$$\text{Pour chaque } j \in [\![1, p]\!], \text{ faire : } \begin{aligned} r^{\text{int}} &\leftarrow r^{(k)} + \mathbf{x}_j \theta_j^{(k)} \\ \theta_j^{(k+1)} &\leftarrow \mathbf{x}_j^\top r^{\text{int}} / (\|\mathbf{x}_j\|_2^2 + \lambda) \\ r^{(k+1)} &\leftarrow r^{\text{int}} - \mathbf{x}_j \theta_j^{(k+1)} \end{aligned}$$

Ridge descente par coordonnée (II)

Mise à jour intelligente : stocker à l'itération k les résidus courants $r^{(k)} = y - X\boldsymbol{\theta}^{(k)}$ et les coefficients $\boldsymbol{\theta}^{(k)}$

$$\text{Pour chaque } j \in [\![1, p]\!], \text{ faire : } \begin{aligned} r^{\text{int}} &\leftarrow r^{(k)} + \mathbf{x}_j \theta_j^{(k)} \\ \theta_j^{(k+1)} &\leftarrow \mathbf{x}_j^\top r^{\text{int}} / (\|\mathbf{x}_j\|_2^2 + \lambda) \\ r^{(k+1)} &\leftarrow r^{\text{int}} - \mathbf{x}_j \theta_j^{(k+1)} \end{aligned}$$

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)

Sommaire

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 θ_j avec les autres θ_k $(k \neq j)$ fixes :

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 θ_j avec les autres θ_k $(k \neq j)$ fixes :
$\hat{\theta}_j = \arg\min_{\theta_j \in \mathbb{R}} f(\theta_1, \ldots, \theta_p)$

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 θ_j avec les autres θ_k $(k \neq j)$ fixes :

$$\begin{aligned}\hat{\theta}_j &= \arg\min_{\theta_j\in\mathbb{R}} f(\theta_1, \ldots, \theta_p) \\ &= \arg\min_{\theta_j\in\mathbb{R}} \frac{1}{2}\|\mathbf{y} - \sum_{k\neq j} \theta_k \mathbf{x}_k - \mathbf{x}_j\theta_j\|^2 + \lambda \sum_{k\neq j} |\theta_k| + \lambda|\theta_j|\end{aligned}$$

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 θ_j avec les autres θ_k $(k \neq j)$ fixes :

$$\begin{aligned}
\hat{\theta}_j &= \arg\min_{\theta_j\in\mathbb{R}} f(\theta_1, \dots, \theta_p) \\
&= \arg\min_{\theta_j\in\mathbb{R}} \frac{1}{2}\|\mathbf{y} - \sum_{k\neq j} \theta_k \mathbf{x}_k - \mathbf{x}_j\theta_j\|^2 + \lambda \sum_{k\neq j} |\theta_k| + \lambda|\theta_j| \\
&= \arg\min_{\theta_j\in\mathbb{R}} \frac{1}{2}\|\mathbf{x}_j\|^2\theta_j^2 - \langle y - \sum_{k\neq j} \theta_k \mathbf{x}_k, \mathbf{x}_j\rangle\theta_j + \lambda|\theta_j|
\end{aligned}$$

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 θ_j avec les autres θ_k $(k \neq j)$ fixes :

$$\begin{aligned}
\hat{\theta}_j &= \arg\min_{\theta_j\in\mathbb{R}} f(\theta_1,\ldots,\theta_p) \\
&= \arg\min_{\theta_j\in\mathbb{R}} \frac{1}{2}\|\mathbf{y} - \sum_{k\neq j}\theta_k\mathbf{x}_k - \mathbf{x}_j\theta_j\|^2 + \lambda\sum_{k\neq j}|\theta_k| + \lambda|\theta_j| \\
&= \arg\min_{\theta_j\in\mathbb{R}} \frac{1}{2}\|\mathbf{x}_j\|^2\theta_j^2 - \langle y - \sum_{k\neq j}\theta_k\mathbf{x}_k, \mathbf{x}_j\rangle\theta_j + \lambda|\theta_j| \\
&= \arg\min_{\theta_j\in\mathbb{R}} \|\mathbf{x}_j\|^2\left[\frac{1}{2}\left(\theta_j - \|\mathbf{x}_j\|^{-2}\langle y - \sum_{k\neq j}\theta_k\mathbf{x}_k, \mathbf{x}_j\rangle\right)^2 + \frac{\lambda}{\|\mathbf{x}_j\|^2}|\theta_j|\right]
\end{aligned}$$

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 θ_j avec les autres θ_k $(k \neq j)$ fixes :

$$\begin{aligned}
\hat{\theta}_j &= \arg\min_{\theta_j\in\mathbb{R}} f(\theta_1, \dots, \theta_p) \\
&= \arg\min_{\theta_j\in\mathbb{R}} \frac{1}{2}\|\mathbf{y} - \sum_{k\neq j}\theta_k\mathbf{x}_k - \mathbf{x}_j\theta_j\|^2 + \lambda\sum_{k\neq j}|\theta_k| + \lambda|\theta_j| \\
&= \arg\min_{\theta_j\in\mathbb{R}} \frac{1}{2}\|\mathbf{x}_j\|^2\theta_j^2 - \langle y - \sum_{k\neq j}\theta_k\mathbf{x}_k, \mathbf{x}_j\rangle\theta_j + \lambda|\theta_j| \\
&= \arg\min_{\theta_j\in\mathbb{R}} \|\mathbf{x}_j\|^2\left[\frac{1}{2}\left(\theta_j - \|\mathbf{x}_j\|^{-2}\langle y - \sum_{k\neq j}\theta_k\mathbf{x}_k, \mathbf{x}_j\rangle\right)^2 + \frac{\lambda}{\|\mathbf{x}_j\|^2}|\theta_j|\right]
\end{aligned}$$

<u>Rappel</u> : $\quad \eta_{\mathrm{ST},\lambda}(z) = \arg\min_{t\in\mathbb{R}} \frac{1}{2}(z-t)^2 + \lambda|t|$

Lasso : descente par coordonnée (II)

Solution :

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

Mise à jour intelligente : stocker à l'itération k les résidus courants $r^{(k)} = y - X\boldsymbol{\theta}^{(k)}$ et les coefficients $\boldsymbol{\theta}^{(k)}$

Lasso : descente par coordonnée (II)

Solution :

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

<u>Mise à jour intelligente</u> : stocker à l'itération k les résidus courants $r^{(k)} = y - X\boldsymbol{\theta}^{(k)}$ et les coefficients $\boldsymbol{\theta}^{(k)}$

Pour chaque $j \in [\![1, p]\!]$, faire :

Lasso : descente par coordonnée (II)

Solution :

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

Mise à jour intelligente : stocker à l'itération k les résidus courants $r^{(k)} = y - X\boldsymbol{\theta}^{(k)}$ et les coefficients $\boldsymbol{\theta}^{(k)}$

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

Pour chaque $j \in [\![1, p]\!]$, faire :

Lasso : descente par coordonnée (II)

Solution :

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

Mise à jour intelligente : stocker à l'itération k les résidus courants $r^{(k)} = y - X\boldsymbol{\theta}^{(k)}$ et les coefficients $\boldsymbol{\theta}^{(k)}$

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

Pour chaque $j \in [\![1, p]\!]$, faire : $\theta_j^{(k+1)} \leftarrow \eta_{\mathrm{ST},\lambda/\|\mathbf{x_j}\|^2}\left(\mathbf{x}_j^\top r^{\mathrm{int}}/\|\mathbf{x}_j\|^2\right)$

Lasso : descente par coordonnée (II)

Solution :

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

Mise à jour intelligente : stocker à l'itération k les résidus courants $r^{(k)} = y - X\boldsymbol{\theta}^{(k)}$ et les coefficients $\boldsymbol{\theta}^{(k)}$

Pour chaque $j \in [\![1, p]\!]$, faire :
$$\begin{aligned} r^{\mathrm{int}} &\leftarrow r^{(k)} + \mathbf{x}_j \theta_j^{(k)} \\ \theta_j^{(k+1)} &\leftarrow \eta_{\mathrm{ST},\lambda/\|\mathbf{x_j}\|^2}\left(\mathbf{x}_j^\top r^{\mathrm{int}}/\|\mathbf{x}_j\|^2\right) \\ r^{(k+1)} &\leftarrow r^{\mathrm{int}} - \mathbf{x}_j \theta_j^{(k+1)} \end{aligned}$$

Lasso : descente par coordonnée (II)

Solution :

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

Mise à jour intelligente : stocker à l'itération k les résidus courants $r^{(k)} = y - X\boldsymbol{\theta}^{(k)}$ et les coefficients $\boldsymbol{\theta}^{(k)}$

Pour chaque $j \in [\![1, p]\!]$, faire :
$$\begin{aligned} r^{\mathrm{int}} &\leftarrow r^{(k)} + \mathbf{x}_j\theta_j^{(k)} \\ \theta_j^{(k+1)} &\leftarrow \eta_{\mathrm{ST},\lambda/\|\mathbf{x_j}\|^2}\left(\mathbf{x}_j^\top r^{\mathrm{int}}/\|\mathbf{x}_j\|^2\right) \\ r^{(k+1)} &\leftarrow r^{\mathrm{int}} - \mathbf{x}_j\theta_j^{(k+1)} \end{aligned}$$

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)

Sommaire

Descente par coordonnée : cas non-convexe

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

Même approche mais sans garantie de convergence globale

Descente par coordonnée : cas non-convexe

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

Même approche mais sans garantie de convergence globale

Pour chaque $j \in [\![1, p]\!]$, faire :

Descente par coordonnée : cas non-convexe

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

Même approche mais sans garantie de convergence globale

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

Pour chaque $j \in [\![1, p]\!]$, faire :

Descente par coordonnée : cas non-convexe

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

Même approche mais sans garantie de convergence globale

$$\text{Pour chaque } j \in [\![1, p]\!], \text{ faire : } \begin{aligned} r^{\text{int}} &\leftarrow r^{(k)} + \mathbf{x}_j \theta_j^{(k)} \\ \theta_j^{(k+1)} &\leftarrow \eta_{\text{pen}_{\lambda,\gamma}}\left(\mathbf{x}_j^\top r^{\text{int}}\right) \end{aligned}$$

Descente par coordonnée : cas non-convexe

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

Même approche mais sans garantie de convergence globale

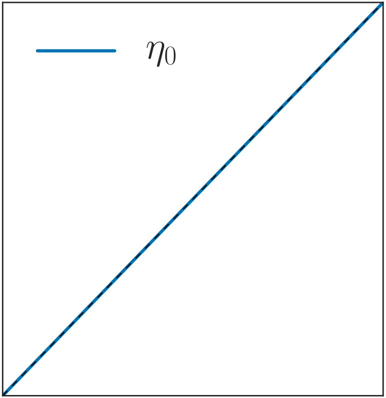
$$\text{Pour chaque } j \in [\![1,p]\!], \text{ faire : } \begin{aligned} r^{\text{int}} &\leftarrow r^{(k)} + \mathbf{x}_j\theta_j^{(k)} \\ \theta_j^{(k+1)} &\leftarrow \eta_{\text{pen}_{\lambda,\gamma}}\left(\mathbf{x}_j^\top r^{\text{int}}\right) \\ r^{(k+1)} &\leftarrow r^{\text{int}} - \mathbf{x}_j\theta_j^{(k+1)} \end{aligned}$$

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

et $\|\mathbf{x}_j\|_2^2 = 1$; voir par exemple Breheny et Huang (2011)

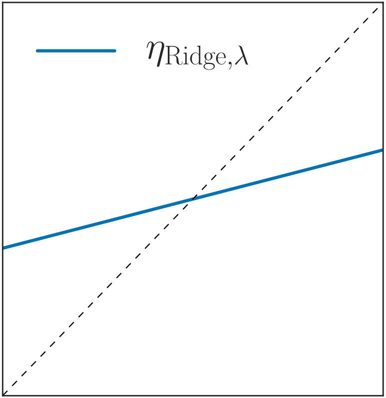
Régularisation en 1D : Aucune

$\eta_0(z) = z$



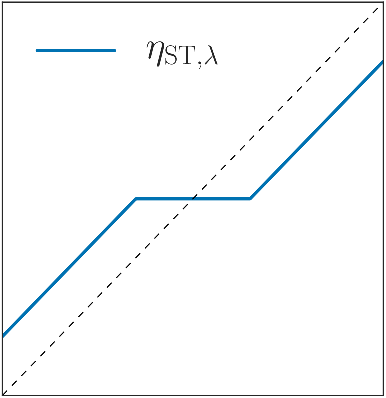
Régularisation en 1D : Ridge

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



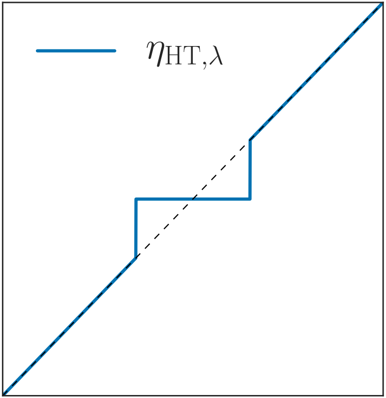
Régularisation en 1D : Lasso

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



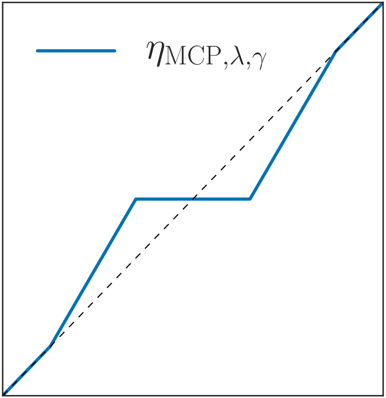
Régularisation en 1D : ℓ_0

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



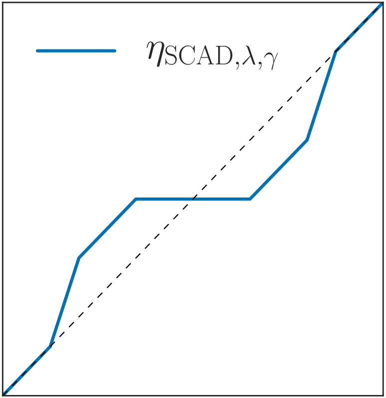
Régularisation en 1D : MCP

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



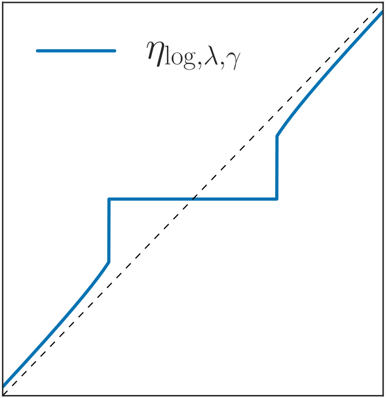
Régularisation en 1D : SCAD

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



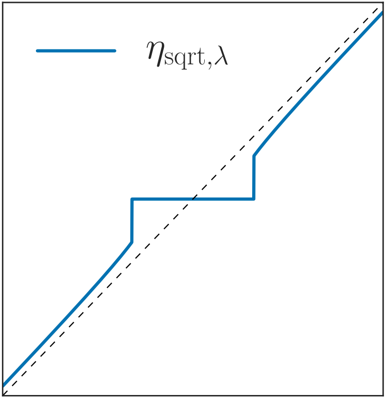
Régularisation en 1D : $\log$

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



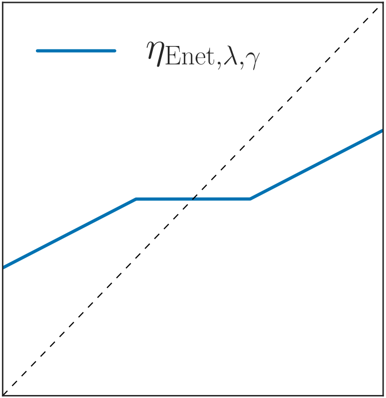
Régularisation en 1D : sqrt

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



Régularisation en 1D : Enet

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

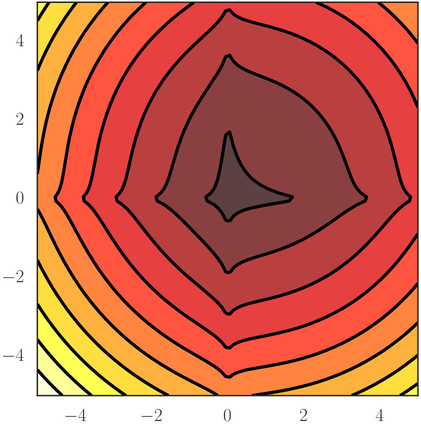


Lasso-Positif

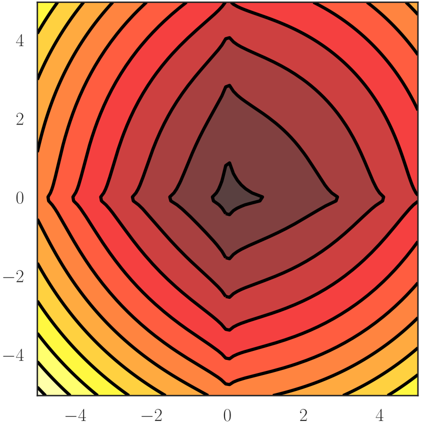
Exo: Proposer une manière de résoudre le problème Lasso avec une contrainte de positivité sur les coefficients

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

Lignes de niveaux pour log



Lignes de niveaux pour sqrt



Optimisation pour le Lasso : autres méthodes

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

- LARS Efron *et al.* (2004) pour le chemin entier. Celui-ci est affine par morceaux, et on peut calculer toutes les solutions par moindre carrées successifs
- méthodes de gradient proximal, Forward-Backward, de type Seuillage Doux Itératif (: *ISTA, FISTA*), *cf.* Beck et Teboulle(2009) : algorithme qui consiste à alterner mise à jour par descente de gradient sur la partie lisse (quadratique) et seuillage de l'itéré

Ces dernières méthodes seront vues dans des cours ultérieurs (*e.g.*, INFMDI 341)

Références I

- P. Breheny and J. Huang.

  Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection.

  Ann. Appl. Stat., 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.