Apprentissage CH3

→ Apprentissage supervisé

↳ Modèle de regression linéaire

$$h_\beta(x) = \beta_0 + \beta_1 x_1 + \dots + \beta_D x_D \qquad x \in \mathbb{R}^D \qquad \beta \in \mathbb{R}^{D+1}$$

→ Fonction coût

$$l(\beta) = \frac{1}{N} \sum_{i=1}^{N} \left(t^{(i)} - (\beta_0 + \beta_1 x_1^{(i)} + \dots + \beta_D x_D^{(i)}) \right)^2$$

→ 2 approches pour trouver le $\vec{\beta}$ optimal

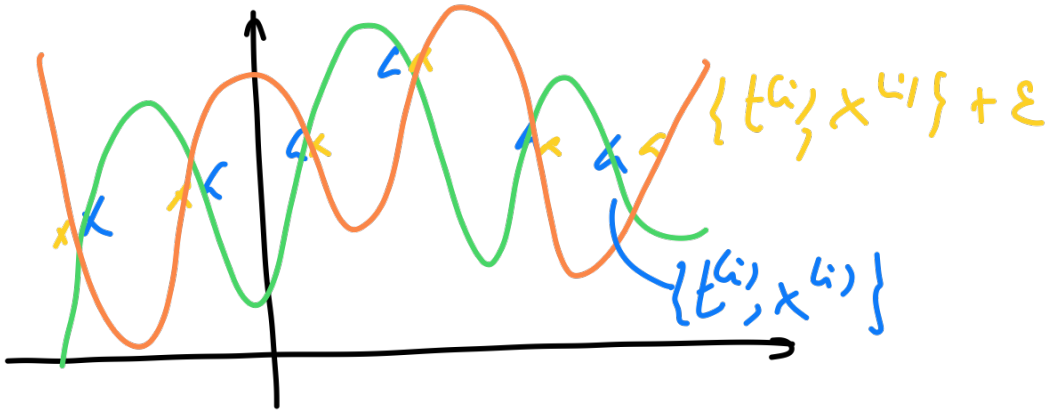
→ Descente de gradient

→ Calcul et accumulation des dérivées

+ résolution des equations normales

⟶ $\vec{\beta} = (\underline{\underline{X}}^T \underline{\underline{X}})^{-1} \underline{\underline{X}}^T \vec{t}$

? X^TX n'est pas inversible ou mal conditionné



$$\beta_2 = \frac{\langle t, c_1 \rangle}{\langle x_1, x_p \rangle} \quad \text{>>>}$$

→ Régularisation

#1 Sélection du meilleur sous ensemble (Best Subset Selection)

→ for $k = 2 : D+1$

Evaluer le modèle entraîné sur un sous ensemble de taille k des $D+1$ β_j

→ Sélectionner le sous ensemble qui donne la meilleure prédiction sur l'ensemble test (via cross validation)

$$\sum_{k=2}^{D+1} \binom{D+1}{k}$$

#2 Ajouter une pénalité sur les coefficients β_j à la fonction coût $\mathcal{L}(\beta)$ → SOMME DES CARRÉS DES β_j

$$\mathcal{L}(\beta) = \frac{1}{N}\sum_{i=1}^{N}\left(t^{(i)} - (\beta_0 + \beta_1 x_1^{(i)} + \dots + \beta_D x_D^{(i)})\right)^2 + \lambda \sum_{j=1}^{D} (\beta_j)^2$$

→ fidélité aux données

pénalité sur la complexité du modèle

→ Regression Ridge

$$\vec{\varepsilon} = (\vec{t} - \tilde{\underline{X}}\vec{\beta})$$

$$\mathcal{L}(\beta) = \frac{1}{N}\sum_{i=1}^{N}(\varepsilon^{(i)})^2 = \frac{1}{N}\|\vec{\varepsilon}\|^2 = \langle \vec{\varepsilon}, \vec{\varepsilon}\rangle \frac{1}{N}$$

$$= \vec{\varepsilon}^T\vec{\varepsilon}\,\frac{1}{N}$$

$$\mathcal{L}(\beta) = \frac{1}{N}(\vec{t} - \tilde{\underline{X}}\vec{\beta})^T(\vec{t} - \tilde{\underline{X}}\beta)$$

$$\frac{\partial \mathcal{L}(\beta)}{\partial \beta_0} = \frac{\partial}{\partial \beta_0} \frac{1}{N} \sum_{i=1}^{N} \left(t^{(i)} - \left(\beta_0 + \beta_1 x_1^{(i)} + \dots + \beta_D x_D^{(i)} \right) \right)^2$$

$$= \frac{2}{N} \sum_{i=1}^{N} \left(t^{(i)} - \left(\beta_0 + \beta_1 x_1^{(i)} + \dots + \beta_D x_D^{(i)} \right) \right) \cdot (-1) = 0$$

$$\rightarrow \quad \sum_{i=1}^{N} t^{(i)} - \sum_{i=1}^{N} \beta_0 - \sum_{i=1}^{N} \left(\beta_1 x_1^{(i)} + \dots + \beta_D x_D^{(i)} \right) = 0$$

à partir de β_1

$$0 = \sum_{i=1}^{N} t^{(i)} - N \beta_0 - \sum_{i=1}^{N} \beta^T \vec{x} \qquad \text{où } \beta = (\beta_1, \dots, \beta_D)$$

$$\beta_0 = \frac{1}{N} \sum_{i=1}^{N} t^{(i)} - \frac{1}{N} \sum_{i=1}^{N} \beta^T \vec{x}^{(i)}$$

$\tilde{X} = [1, x_1, \dots x_D]$

* $\tilde{\beta} = [\beta_0, \vec{\beta}]$

Si on centre les données

$$x_c^{(i)} \leftarrow x^{(i)} - \frac{1}{N}\sum_{j=1}^{N} x^{(j)} = x^{(i)} - \bar{x}$$

$$t_c^{(i)} \leftarrow t^{(i)} - \frac{1}{N}\sum_{j=1}^{N} t^{(j)} \quad — \; t^{(i)} - \bar{t}$$

Pour $x_c^{(i)}$ et $t_c^{(i)}$ si on calcule le β_0 on retrouve

$$\beta_0 = \frac{1}{N}\sum_{i=1}^{N} t_c^{(i)} - \frac{1}{N}\sum_{i=1}^{N} \beta^T x_c^{(i)} \leftarrow \quad ** = *(\text{données centrés})$$

$$= \frac{1}{N}\sum_{i=1}^{N}\left(t^{(i)} - \frac{1}{N}\sum_{j=1}^{N} t^{(j)}\right) - \frac{1}{N}\sum_{i=1}^{N} \beta^T\left(x^{(i)} - \frac{1}{N}\sum_{j=1}^{N} x^{(j)}\right)$$

$\bar{t}$

$$\frac{1}{N}\sum_{i=1}^{N} t^{(i)} - \frac{1}{N}\sum_{i=1}^{N} \cdot \frac{1}{N}\sum_{j=1}^{N} t^{(j)} = \frac{1}{N}\sum_{i=1}^{N} t^{(i)} - \frac{1}{N} N \cdot \bar{t}$$

$$= \underbrace{\frac{1}{N}\sum_{i=1}^{N} t^{(i)}}_{\bar{t}} - \bar{t}$$

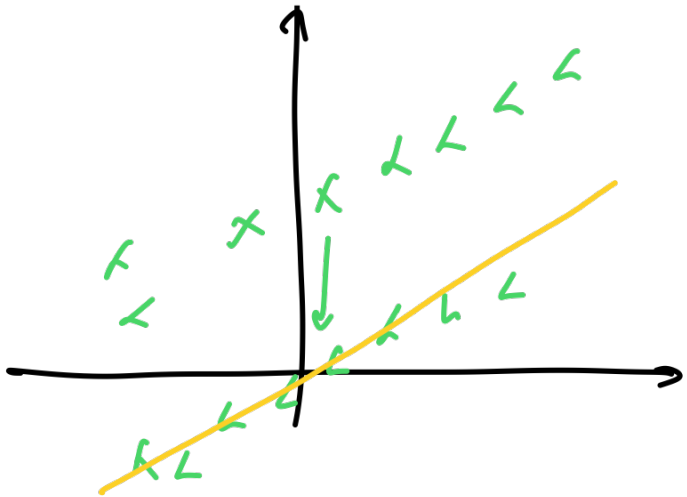
$$= \bar{t} - \bar{t} = 0$$

pour le second terme, on a de manière equivalente

$$-\underbrace{\frac{1}{N}\sum_{i=1}^{N}}\beta^T x^{(i)} + \frac{1}{N}\sum_{i=1}^{N}\beta^T\Big(\underbrace{\frac{1}{N}\sum_{j=1}^{N} x^{(j)}}_{\bar{x}}\Big)$$

$$-\frac{1}{N}\beta^T\sum_{i=1}^{N} x^{(i)} + \underbrace{\frac{1}{N}}\, N\cdot\beta^T\bar{x}$$

$$-\beta^T\bar{x} + \beta^T\bar{x} = 0$$



Pour des données centrées on se ramène à une fonction de coût du type

$$l(\beta) = \frac{1}{N} \sum_{i=1}^{N} \left(t^{(i)} - \left(\beta_1 x_1^{(i)} + \beta_2 x_2^{(i)} + \dots + \beta_D x_D^{(i)} \right) \right)^2 + \lambda \sum_{j=1}^{D} \beta_j^2$$

à partir de maintenant on utilisera $\hat{\beta}$ pour représenter le vecteur $[\beta_0, \beta_1, \dots, \beta_D]$ et $\vec{\beta} = [\beta_1, \beta_2 \dots, \beta_D]$

→ Pour la formulation de type Ridge, on a

$$l(\beta) = \frac{1}{N} \vec{\varepsilon}^T \vec{\varepsilon} = \frac{1}{N} \left(\vec{t} - \underline{\underline{X}} \vec{\beta} \right)^T \left(\vec{t} - \underline{\underline{X}} \vec{\beta} \right) + \lambda \vec{\beta}^T \vec{\beta}$$

$$\ell(\beta) = \frac{1}{N}\left(\vec{t}^T\vec{t} - \vec{\beta}^T \underline{\underline{X}}^T \vec{t} - \vec{t}^T \underline{\underline{X}} \beta + \vec{\beta}^T \underline{\underline{X}}^T \underline{\underline{X}} \vec{\beta} \right) + \lambda \beta^T \beta$$

$\frac{\partial \ell}{\partial \beta}$

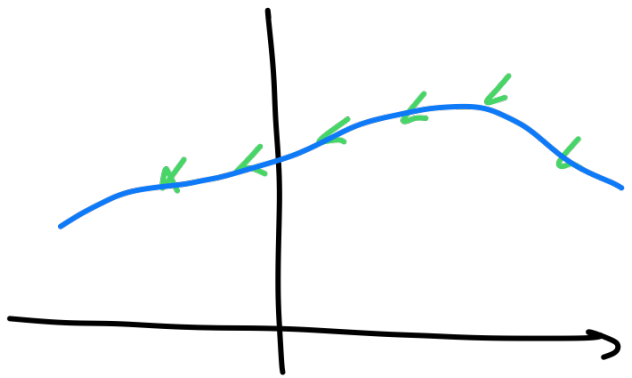
$$\frac{\partial}{\partial \beta}(v^T \beta) = v$$

$$\frac{\partial}{\partial \beta} = -2 \underline{\underline{X}}^T \vec{t} + 2 \underline{\underline{X}}^T \underline{\underline{X}} \beta \qquad 2\lambda\beta$$

$$\frac{\partial}{\partial \beta} \ell(\beta) = -2 \underline{\underline{X}}^T \vec{t} + 2 \underline{\underline{X}}^T \underline{\underline{X}} \beta + 2\lambda\beta = 0$$

$$\underline{\underline{X}}^T \vec{t} = \left(\underline{\underline{X}}^T \underline{\underline{X}} + \lambda \underline{\underline{I}}_D \right) \vec{\beta}$$

$$\vec{\beta} = \left(\underline{\underline{X}}^T \underline{\underline{X}} + \lambda \underline{\underline{I}}_D \right)^{-1} \underline{\underline{X}}^T \vec{t}$$



$A\vec{x} = \gamma\vec{x}$

x, γ vecteur et valeur propre de A

$$(A + \lambda I)\vec{x} = A\vec{x} + \lambda\vec{x}$$
$$= \gamma\vec{x} + \lambda\vec{x}$$
$$= (\gamma + \lambda)\vec{x}$$

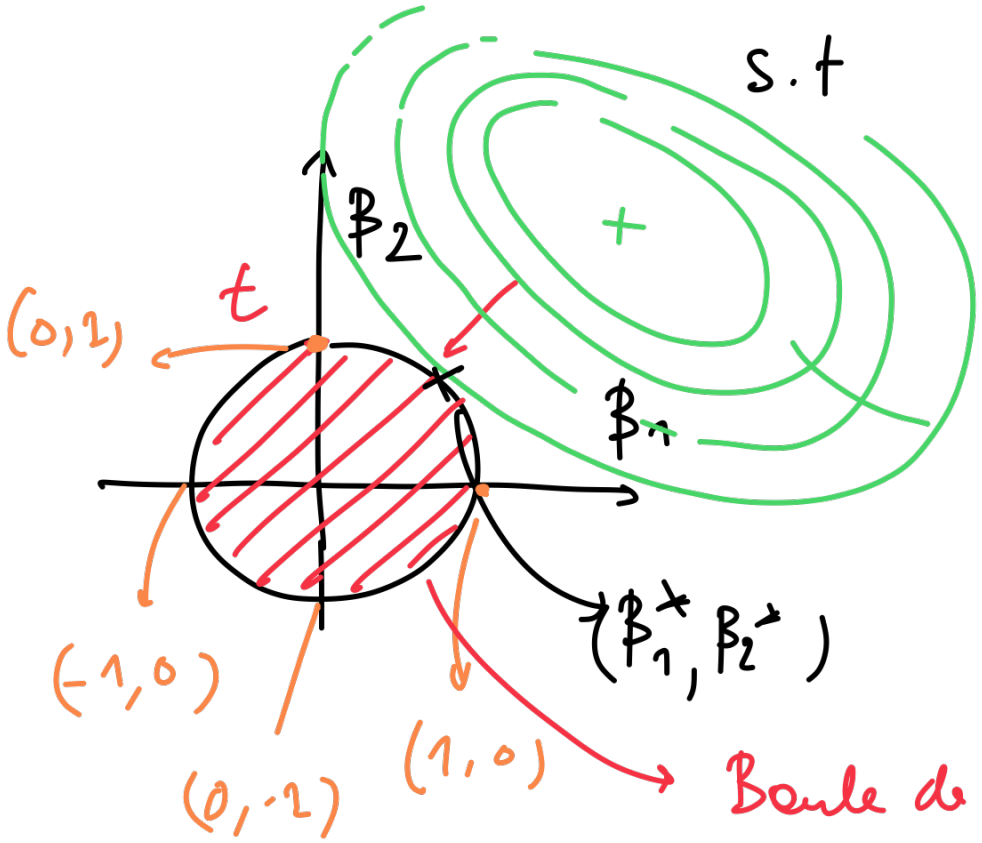
→ la plus petite valeur propre de

$(\underline{\underline{X}}^T\underline{\underline{X}} + \lambda I_D)$ est à priori donnée par $\lambda + \lambda_{min}(\underline{\underline{X}}^T\underline{\underline{X}})$

La formulation Ridge peut aussi s'écrire de façon contrainte

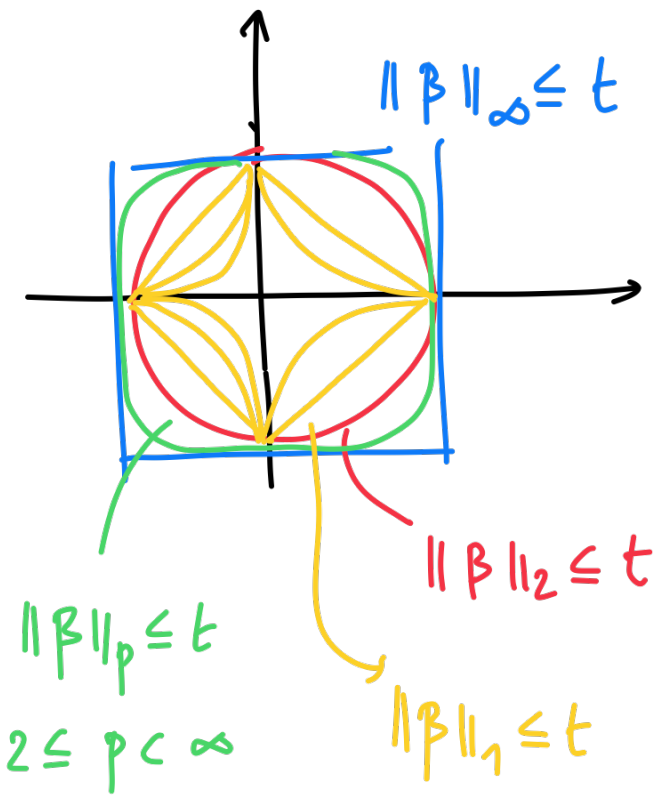
$$\min \ell(\beta) = \frac{1}{N} \sum_{i=1}^{N} \left(t^{(i)} - (\beta_1 x_1^{(i)} + \beta_2 x_2^{(i)} + \dots + \beta_D x_D^{(i)}) \right)^2$$

$$\text{s.t} \quad \sum_{j=1}^{D} \beta_j^2 \leq t \quad \leftarrow$$



$$\|\vec{\beta}\|_2^2 = \left(\sqrt{\beta_1^2 + \beta_2^2 + \dots + \beta_D^2} \right)^2$$

Boule de la norme l_2



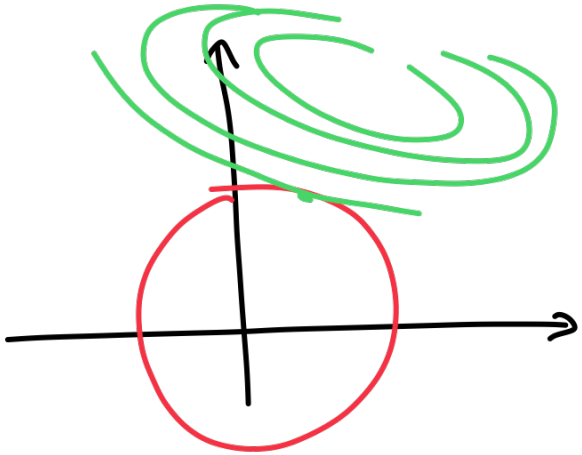
$$\|\vec{\beta}\|_p = \left(\sum_{j=1}^{2} |\beta_j|^p\right)^{1/p}$$

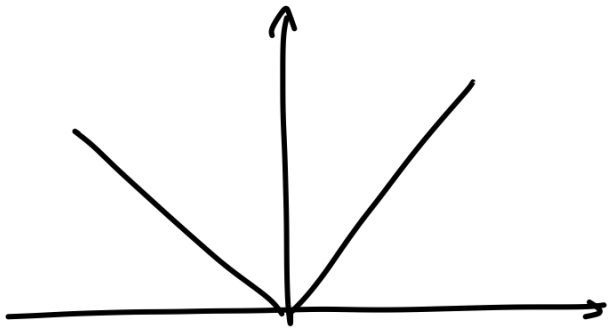
$$\|\vec{\beta}\|_\infty = \max_j |\beta_j|$$

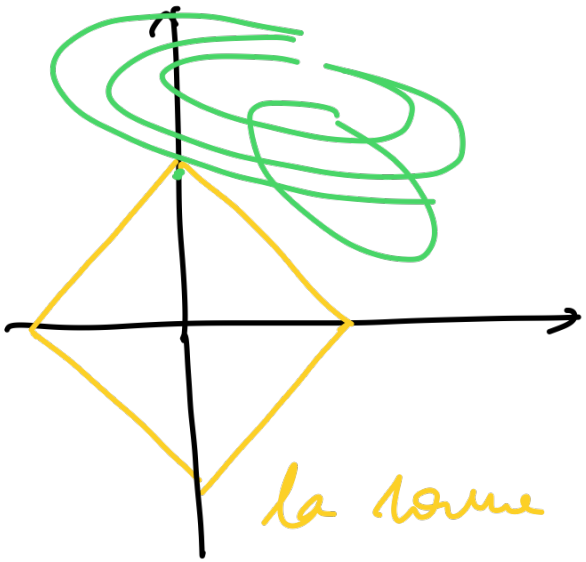
$$\|\vec{\beta}\|_\infty \le t$$

$p=2$ #3 Modèle de regression LASSO

$$\ell(\beta) = \frac{1}{N}\sum_{i=1}^{N}\left(t^{(i)} - (\beta_0 + \beta_1 x_1^{(i)} + \dots + \beta_D x_D^{(i)})\right)^2 + \lambda \sum_{j=1}^{D} |\beta_j|$$







la norme $\|\vec{\beta}\|_1$ est plus efficace en terme de sélection de caractéristiques