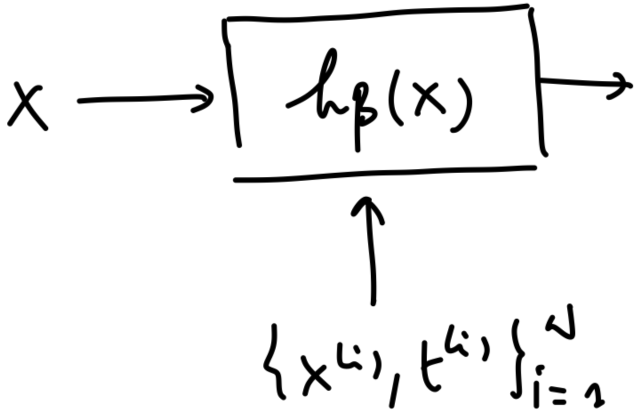
Apprentissage supervisé $\{x^{(i)}, t^{(i)}\}_{i=1}^{N}$



$h_\beta(x)$ t.q $h_\beta(x^{(i)}) \approx t^{(i)}$

→ Apprentissage : 2 approches :
- Descente de gradient
- Resolution des equations normales

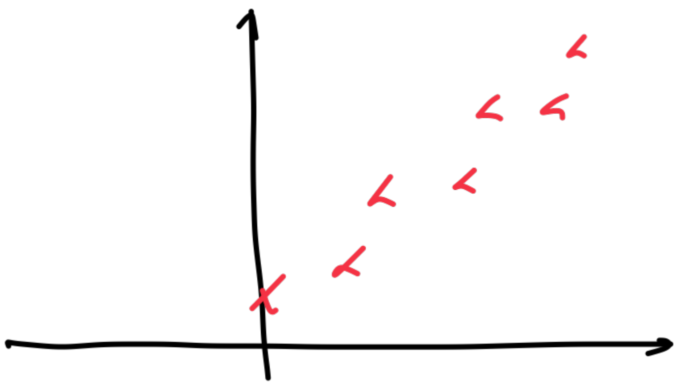
$$\ell(\beta) = \frac{1}{N} \sum_{i=1}^{N} \left(t^{(i)} - \left(\beta_0 + \beta_1 x_1^{(i)} \ldots\ldots \beta_D x_D^{(i)} \right) \right)^2$$

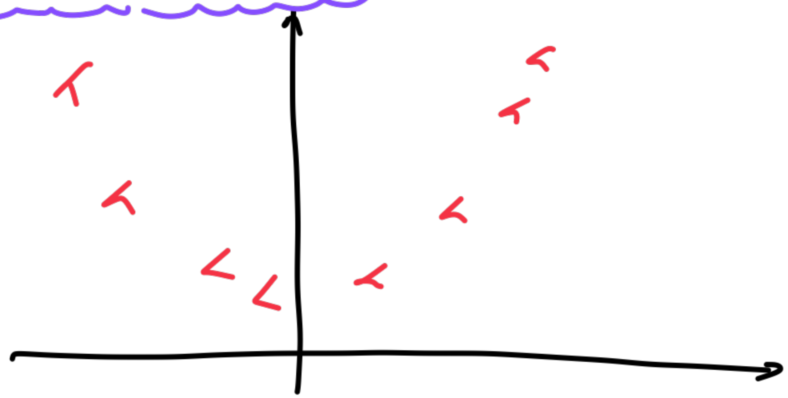
Descente de gradient : $\vec{\beta}^{(0)} \in \mathbb{R}^{D+1}$

for iter in MaxIter

$$\vec{\beta} \leftarrow \vec{\beta} - \eta \,\mathrm{grad}\, \ell(\beta)$$

Equations normales $\vec{\beta}_{OLS} = (\tilde{\underline{\underline{X}}}^T \tilde{\underline{\underline{X}}})^{-1} \tilde{\underline{\underline{X}}}^T t$





Pour D suffisamment grand, il existera toujours un modèle linéaire capable de "fitter" les observations

$\rightarrow$? $\underline{\underline{\tilde{X}}}^T\underline{\underline{\tilde{X}}}$ non inversible ? ex : quand certaines caractéristiques sont dependantes.

$\langle x, y\rangle = x^T y$

$\longrightarrow$ Première solution : Gram-Schmidt

Demarrer avec $c_0 = z_0$ la première colonne de $\tilde{X}$

For $j = 1, \dots, D$

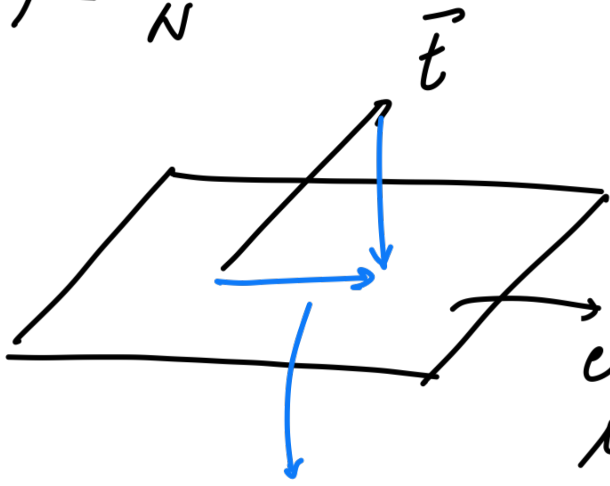
Calculer les coefficients $\hat{\gamma}_{\ell j} = \langle c_j, z_\ell\rangle / \langle z_\ell, z_\ell\rangle$

for $\ell = 0, \dots, j-1$

On défini la nouvelle colonne $\underbrace{z_j}_{z_D} = \underbrace{c_j}_{c_D} - \underbrace{\sum_{\ell=0}^{j-1} \hat{\gamma}_{\ell j} z_\ell}$

$$\min_{\beta} \frac{1}{N}\left(\vec{t} - \underline{\underline{\tilde{X}}}\beta\right)^T\left(\vec{t} - \underline{\underline{\tilde{X}}}\vec{\beta}\right) = \frac{1}{N}\vec{r}^{\,T}\vec{r}$$

$$\min_{\beta} \frac{1}{N}\, \| \underbrace{\vec{t} - \underline{\underline{\tilde{X}}}\beta} \|_2^2 \quad (*)$$



espace engendré par les colonnes de $\tilde{X}$

β t.q $\|\vec{t} - \underline{\underline{X}}\beta\|$ est la plus petite possible

la solution de $(*)$ peut être facilement calculée dans la base des vecteurs z_j via $\langle \vec{t}, z_j \rangle / \langle z_j, z_j \rangle$

$$\langle \vec{t}, z_j \rangle / \langle z_j, z_j \rangle = \alpha_j$$

On a $h_\beta(x) = \sum_j \alpha_j \underbrace{z_j} = \alpha_D z_D + \sum_{j=0}^{D-1} \alpha_j z_j$

$$= \underbrace{\alpha_D} c_D + \dots + \sum_{j=0}^{D-1} \alpha_j z_j$$

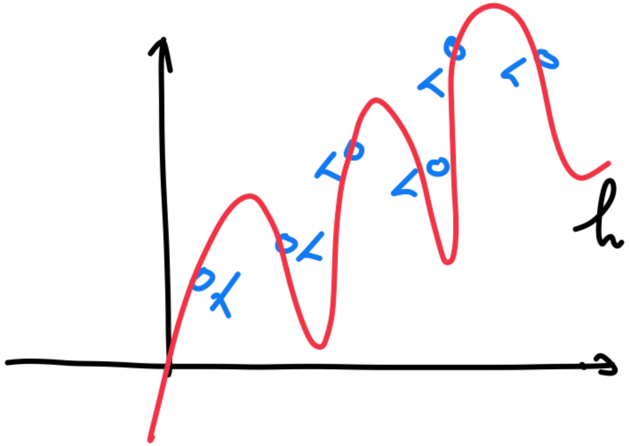
$$\beta_D = \alpha_D = \underbrace{\frac{\langle t, z_D \rangle}{\langle z_D, z_D \rangle}}$$

$$h_\beta(x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_D x_D$$

$$= \beta^T \tilde{x}$$

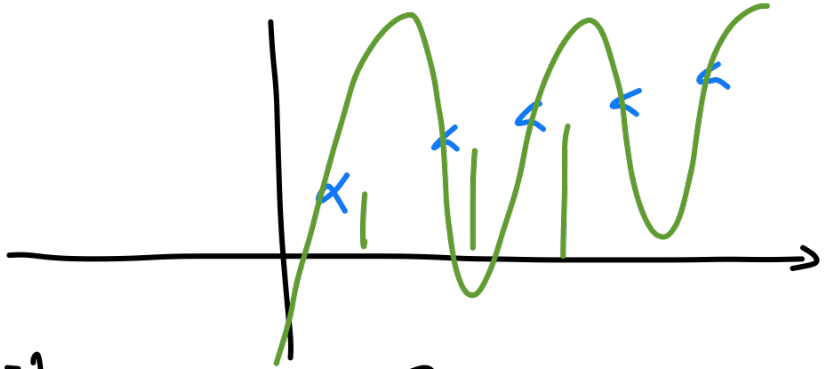
$$\overline{\{x^{(i)}, t^{(i)}\}}_{i=1}^{N}$$

$$\{x^{(i)}, t^{(i)} + \Delta\}_{i=1}^{N}$$



$$\overline{h_\beta(\tilde{x} + \Delta)} = \overline{\tilde{\beta}^T(\tilde{x} + \Delta)}$$

$$h_\beta(\tilde{x}) = \tilde{\beta}^T \tilde{x}$$

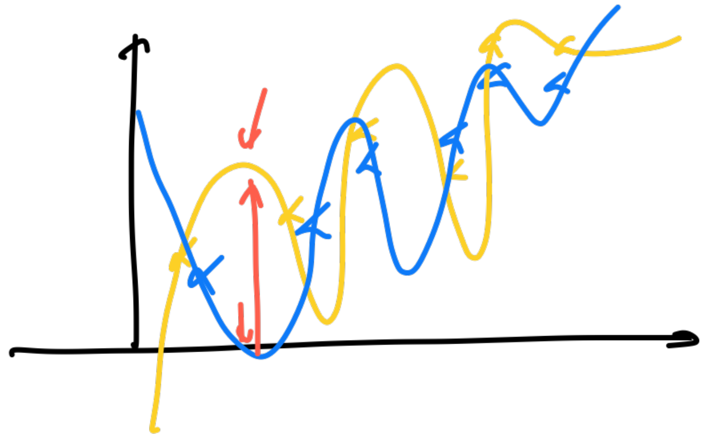


$$\beta_{OLS} = \underbrace{\overline{(\tilde{X}^T \tilde{X})}}^{-1} \tilde{X}^T t$$

$$\begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_{D+1} \end{bmatrix}^{-1} = \begin{bmatrix} \lambda_1^{-1} & & \\ & \ddots & \\ & & \lambda_{D+1}^{-1} \end{bmatrix}$$

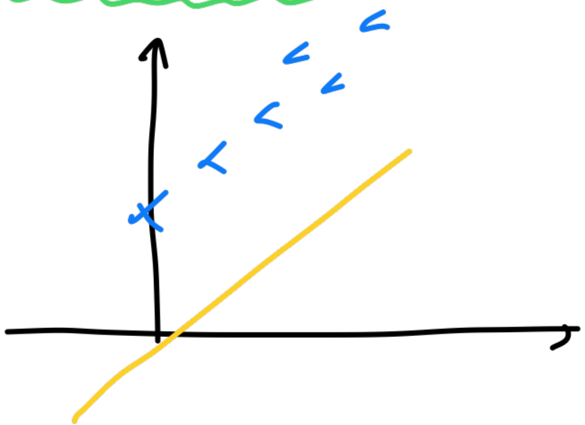
$$\beta^1_{OLS} = (\tilde{X}^T\tilde{X})^{-1}\tilde{X}^T t$$

$$\beta^2_{OLS} = \underbrace{(\tilde{X}^T\tilde{X})^{-1}}\tilde{X}^T (t+\Delta)$$

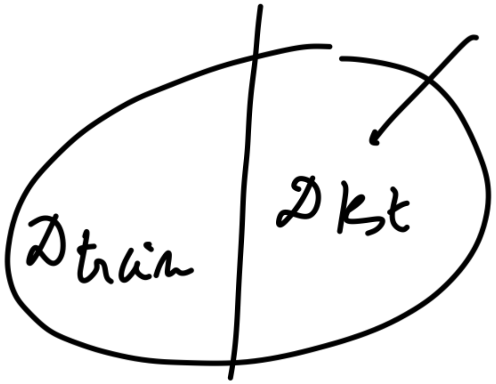


(Sélection de caractéristiques)

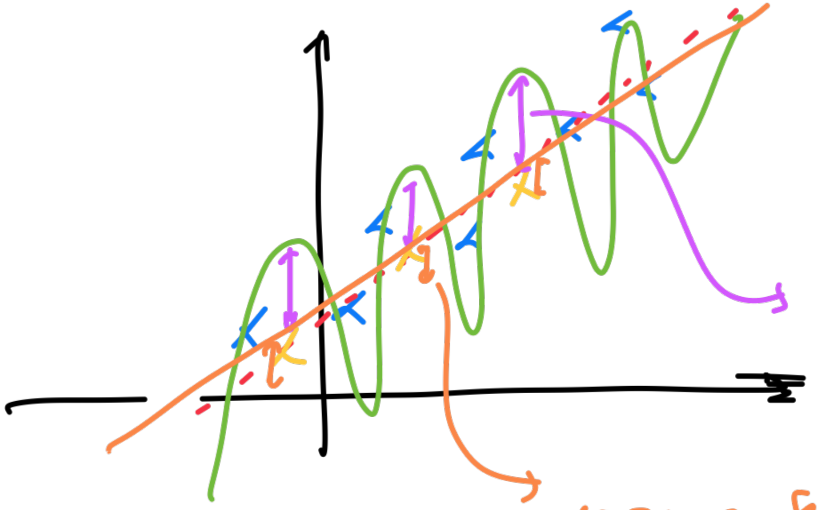
Approche 2 : Sélection du meilleur sous ensemble



$$\sum_{k=1}^{D} \binom{D}{k} \text{ sous ensembles}$$



→ 2 étapes : Pour chaque sous ensemble, on entraîne le modèle sur D_{train} et on calcule l'erreur de prédiction sur D_{test}



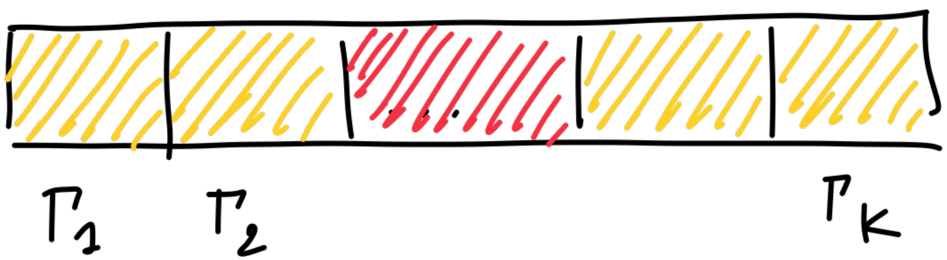
$$h_\beta(x) = \beta_0 + \beta_1 x_1 + \dots + \beta_D x_D$$

$D >>> 1$

erreur sur Dtest

erreur sur Dtest pour un modèle plus simple

Dans le cas où l'ensemble des observations est limité on se tourne vers la validation croisée (à K compartiments)



1 compartiment de test et K-1 compartiments d'entraînement

$$\text{erreur}_{CV} = \frac{1}{N} \sum_{i=1}^{N} \left(t^{(i)} - h_\beta^{-K(i)}(x^{(i)})\right)^2$$

$h_\beta^{-K(i)}$ = le modèle entraîné sur tous les compartiments sauf celui qui contient l'observation $\{x^{(i)}, t^{(i)}\}$

Approche #3 :

terme de fidélité aux données

pénalité sur la complexité du modèle

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

→ Modèle de Regression Ridge (Régularization)

$\beta_1 \quad \beta_2$ pénalité Ridge : $\beta_1^2 + \beta_2^2$

$$\min\ \ell(\beta) = \frac{1}{N}\sum_{i=1}^{N}\left(t^{(i)} - \left(\beta_0 + \beta_1 x_1^{(i)} + \beta_2 x_2^{(i)}\right)\right)^2 + \lambda \sum_{j=1}^{2} \beta_j^2$$

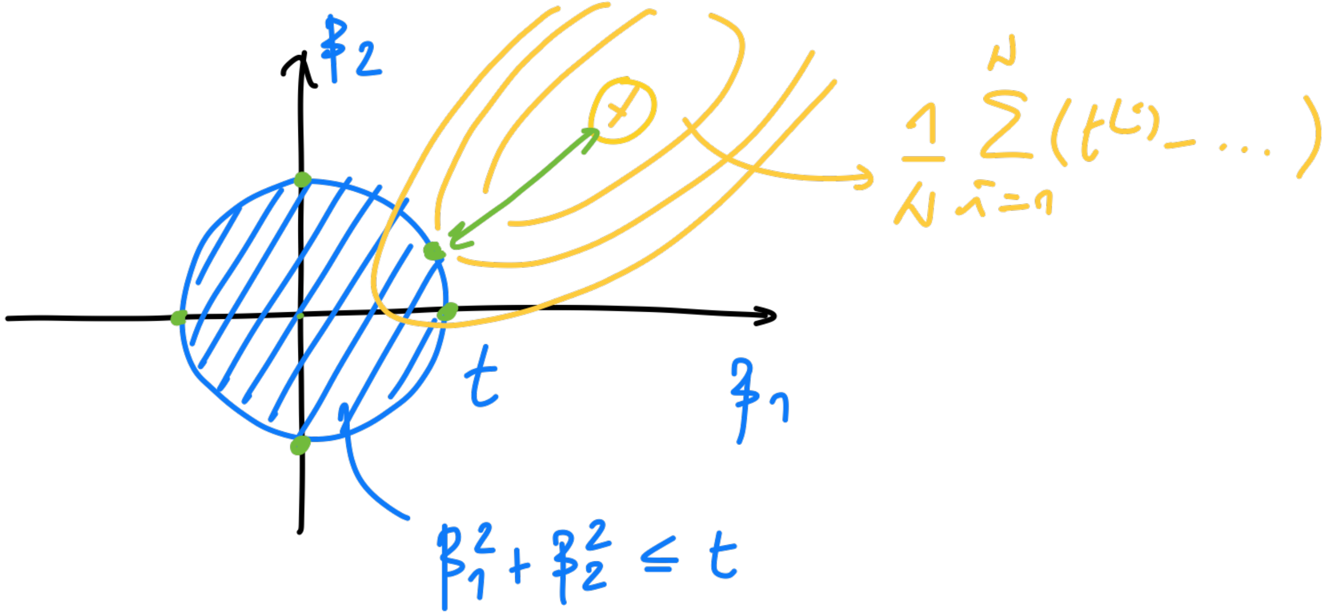
(formulation non contrainte)

$$\min\ \ell(\beta) = \frac{1}{N}\sum_{i=1}^{N}\left(t^{(i)} - \left(\beta_0 + \beta_1 x_1^{(i)} + \beta_2 x_2^{(i)}\right)\right)^2$$

tel que $\displaystyle\sum_{j=1}^{2} \beta_j^2 \leq t^2$

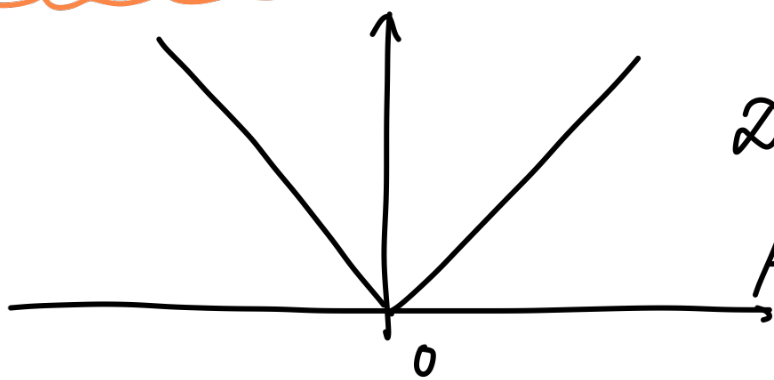
(Formulation contrainte)

$\beta_1^2 + \beta_2^2 \le t^2$



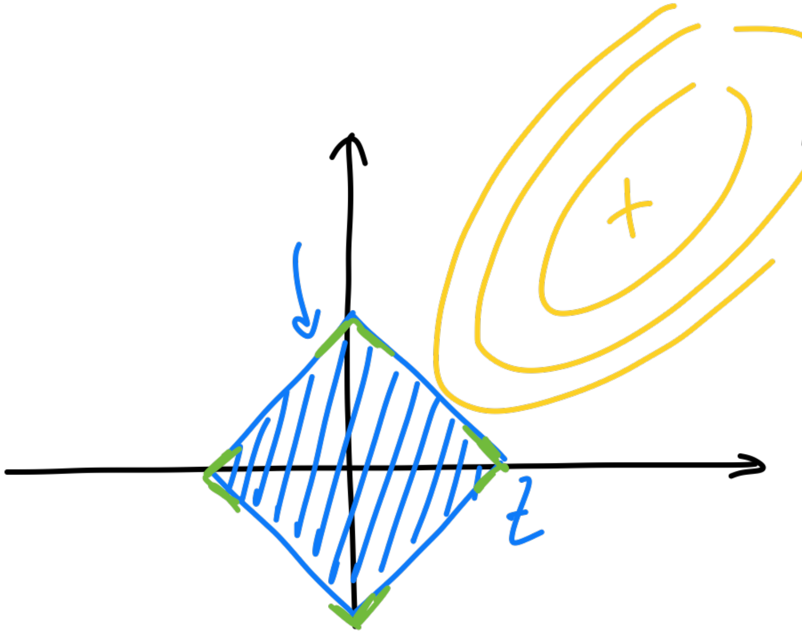
Approche 4: Regression LASSO

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



Désavantage :

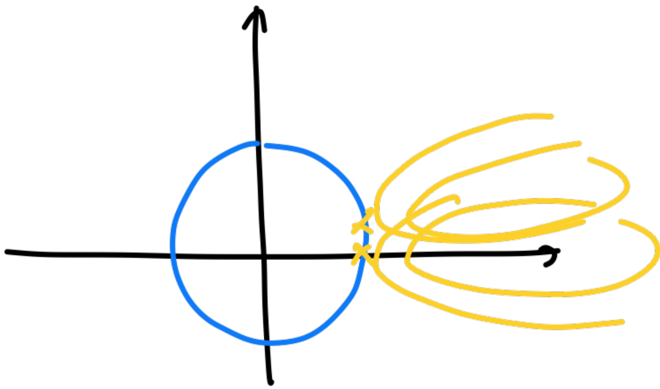
perte de la dérivabilité

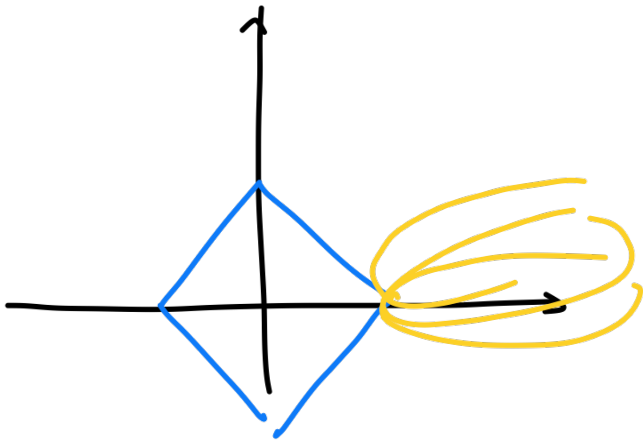


$$\min \ \ell(\beta) = \frac{1}{N} \sum_{i=1}^{N} \left(t^{(i)} - (\beta_0 + \beta_1 x_1^{(i)} + \dots + \beta_D x_D^{(i)}) \right)^2$$

$$\text{t.q} \quad \sum_{j=1}^{D} |\beta_j| \leq t$$

LASSO





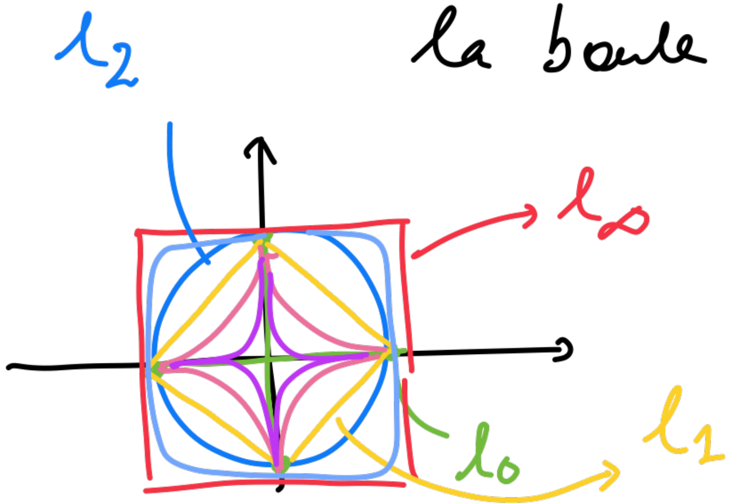
$$\|\vec{\beta}\|_2 = \sqrt{\sum_{j=1}^{D} \beta_j^2} \qquad \|\vec{\beta}\|_1 = \sum_{j=1}^{D} |\beta_j| \qquad \|\vec{\beta}\|_\infty = \max_j |\beta_j|$$

$$\boxed{\|\vec{\beta}\|_p = \left(\sum_{j=1}^{D} \beta_j^p \right)^{1/p}} \longrightarrow \text{norme } l_p$$

A chacune des normes l_p on associe la boule l_p

parmi lesquelles la boule l_1 $\quad \sum_{j=1}^{D} |\beta_j| \leq t$

la boule l_2 $\quad \sum_{j=1}^{D} |\beta_j|^2 \leq t$



Ridge → pour rappel : Fonction somme des résidus carrés

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

$$\rightarrow \quad l(\beta) = \frac{1}{N}\vec{t}^T\vec{t} - \frac{1}{N} t^T \underline{\underline{\tilde{X}}}\beta - \frac{1}{N}\vec{\beta}^T \underline{\underline{\tilde{X}}}^T \vec{t} + \frac{1}{N}\vec{\beta}^T \underline{\underline{\tilde{X}}}^T \underline{\underline{\tilde{X}}}\beta + \lambda \vec{\beta}^T\vec{\beta}$$

$$\operatorname{grad} l(\beta) = -\frac{2}{N}\tilde{X}^T\vec{t} + \frac{2}{N}\underline{\underline{\tilde{X}}}^T\underline{\underline{\tilde{X}}}\vec{\beta} + \frac{2\lambda'}{N}\vec{\beta}$$

$$\operatorname{grad} l(\beta) = 0 \Rightarrow \left(\underline{\underline{\tilde{X}}}^T\underline{\underline{\tilde{X}}} + \lambda I\right)\vec{\beta} = \tilde{X}^T\vec{t}$$

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

Problème avec OLS c'est que lorsqu'on a de petites valeurs propres, l'inverse sera grande.

$$\underline{\underline{\tilde{X}}}^T \underline{\underline{\tilde{X}}} v = \xi v \longrightarrow (\underline{\underline{\tilde{X}}}^T \underline{\underline{\tilde{X}}} + \lambda I) v$$

$$\underline{\underline{\tilde{X}}}^T \underline{\underline{\tilde{X}}} \vec{v} + \lambda \vec{v} = \xi \vec{v} + \lambda \vec{v}$$

$$= (\xi + \lambda) \vec{v}$$

Soit (v, ξ) une paire vecteur propre valeur propre de $\tilde{X}^T \tilde{X}$

$\tilde{X}^T \tilde{X}$ est semi définie positive $\rightarrow$ toutes les valeurs propres sont positives ou nulles

Même si ξ est petit pour un λ suffisamment grand, on peut contrôler l'amplitude de l'inverse, et donc des coefficients de regression.

Supposons que les données sont générées via

$$t^{(i)} = \beta_0 + \beta_1 x_1^{(i)} + \dots + \beta_D x_D^{(i)} + \varepsilon^{(i)}$$

$\varepsilon^{(i)} \sim \mathcal{N}(0, \sigma^2)$

$\varepsilon^{(i)}$ indépendants

identiquement distribués

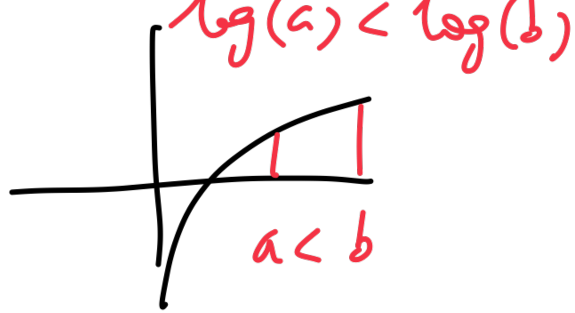
$$p(\varepsilon^{(i)}) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{(\varepsilon^{(i)})^2}{2\sigma^2}\right)$$

$$t^{(i)} \sim \mathcal{N}\left(\beta_0 + \beta_1 x_1^{(i)} + \dots + \beta_D x_D^{(i)}, \sigma^2\right)$$

$$\widehat{p(t^{(i)} \mid \beta)} = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{\left(t^{(i)} - \left(\beta_0 + \beta_1 x_1^{(i)} + \dots + \beta_D x_D^{(i)}\right)\right)^2}{2\sigma^2}\right)$$

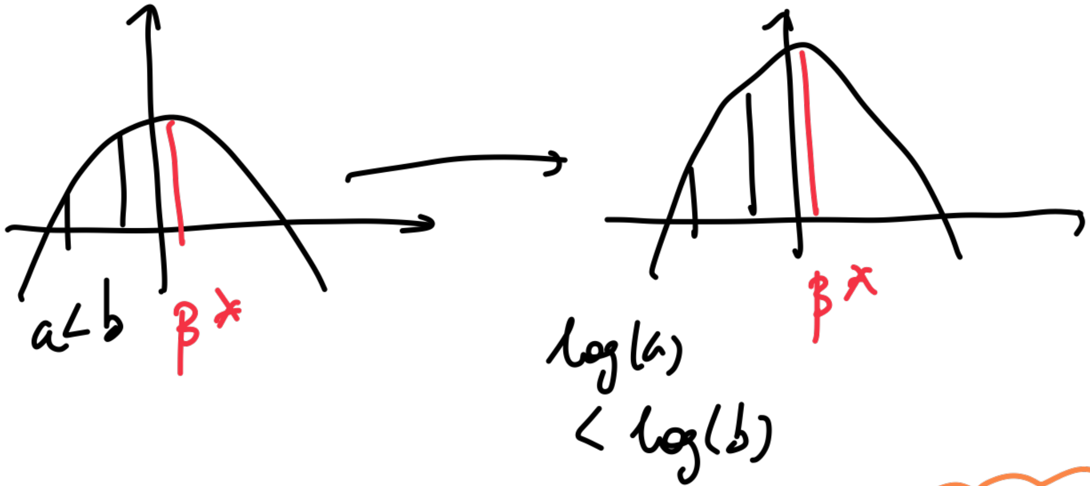
$$p\left(\{t^{(i)}\}_{i=1}^{N} \mid \beta\right) = \prod_{i=1}^{N} p(t^{(i)} \mid \beta)$$

$$= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{\left(t^{(i)} - \left(\beta_0 + \beta_1 x_1^{(i)} + \dots + \beta_D x_D^{(i)}\right)\right)^2}{2\sigma^2}\right)$$



$$\bar{\beta} = \underset{\beta}{\arg\max}\; p\left(\{t^{(i)}\}_{i=1}^{N} \mid \beta\right)$$

$$\vec{\beta} = \arg\max \log\left(\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{\left(t^{(i)}-\left(\beta_0+\beta_1 x_1^{(i)}+\dots+\beta_D x_D^{(i)}\right)\right)^2}{2\sigma^2}\right)\right)$$



Fonction de log vraisemblance

$$\vec{\beta} = \arg\max_{\beta} \boxed{\sum_{i=1}^{N} \cancel{\log\left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)} - \sum_{i=1}^{N} \frac{\left(t^{(i)}-\left(\beta_0+\beta_1 x_1^{(i)}+\dots+\beta_D x_D^{(i)}\right)\right)^2}{2\sigma^2}}$$

$$\boxed{\vec{\beta}_{MLE} = \arg\min_{\beta} \sum_{i=1}^{N} \left(t^{(i)}-\left(\beta_0+\beta_1 x_1^{(i)}+\dots+\beta_D x_D^{(i)}\right)\right)^2}$$

→ le Modèle de regression linéaire entrainé via la minimisation de la fonction de coût donnée par la somme des carrés des résidus correspond à l'estimateur de vraissemblance.

Aucune hypothèse sur le Modèle (i.e coefficients) → MLE

$p(t^{(i)} | \beta)$

Maximum a Posteriori : $p(\beta | t^{(i)})$

Règle de Bayes : $p(B|A) = \dfrac{p(A|B)\,p(B)}{p(A)}$

MAP : $p(\beta \mid t^{(i)}) = \dfrac{p(t^{(i)} \mid \beta)\, p(\beta)}{p(t^{(i)})}$ → ne dépend pas de β

probabilité à posteriori →

Étant donné $p(\beta \mid t^{(i)})$ on peut définir β_{MAP} comme l'estimateur qui maximise cette probabilité

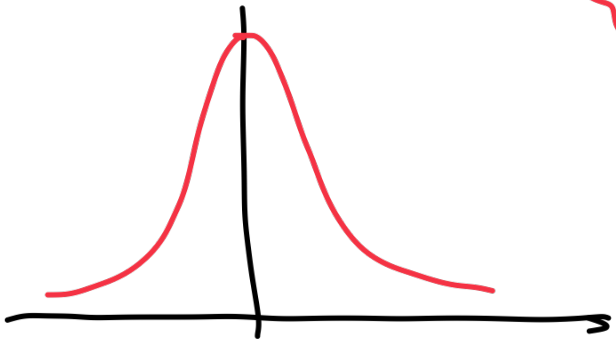
$$\beta_{MAP} = \operatorname{argmax}\ p(t^{(i)} \mid \beta)\, p(\beta)$$

→ A priori sur le modèle

Si $p(\beta)$ est uniforme on récupère le β_{MLE}.

→ Un premier exemple d'a priori sur β consistait à favoriser les coefficients qui soient pour la plupart

Proches de 0 $\longrightarrow$ $\boxed{p(\beta) = \prod_{j=1}^{D} p(\beta_j) = \prod_{j=1}^{D} \frac{1}{\sqrt{2\pi}\,\xi} \exp\left(-\frac{\beta_j^2}{2\xi^2}\right)}$



log

$$p(\beta \mid \{t^{(i)}\}_{i=1}^{N}) \propto \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(\frac{\left(t^{(i)} - (\beta_0 + \beta_1 x_1^{(i)} + \dots + \beta_D x_D^{(i)})\right)^2}{2\sigma^2} \right)$$

$$\times \prod_{j=1}^{D} \frac{1}{\sqrt{2\pi}\,\xi} \exp\left(-\frac{\beta_j^2}{2\xi^2}\right)$$

$$\vec{\beta}_{MAP} = \underset{\beta}{\operatorname{argmax}}\; p(\beta \mid \{t^{(i)}\}_{i=1}^{N})$$

$$\vec{\beta}_{MAP} = \arg\max_{\beta} \log\left(p(\beta \mid \{t^{(i)}\}_{i=1}^{N})\right)$$

$$= \sum_{i=1}^{N} \log\left(\cancel{\frac{1}{\sqrt{2\pi}\,\sigma}}\right) - \frac{\left(t^{(i)} - (\beta_0 + \beta_1 x_1^{(i)} + \dots + \beta_D x_D^{(i)})\right)^2}{2\sigma^2}$$

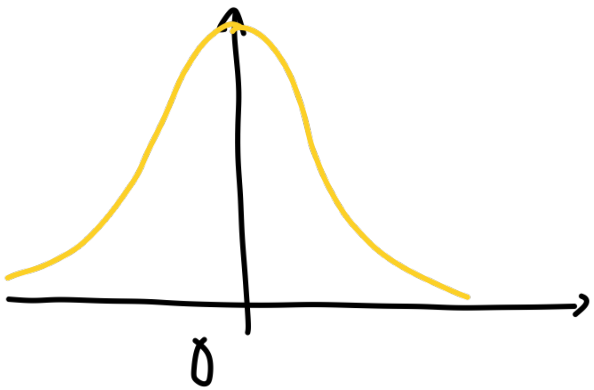
$$+ \sum_{j=1}^{D} \log\left(\cancel{\frac{1}{\sqrt{2\pi}\,\varsigma}}\right) - \frac{\beta_j^2}{2\varsigma^2}$$

Constants par rapport à β

$$\vec{\beta}_{MAP} = \arg\max_{\beta} - \sum_{i=1}^{N} \left(t^{(i)} - (\beta_0 + \beta_1 x_1^{(i)} + \dots + \beta_D x_D^{(i)})\right)^2 / 2\sigma^2 - \sum_{j=1}^{D} \beta_j^2 / 2\varsigma^2$$

$$= \underset{\beta}{\operatorname{argmax}} \ \frac{1}{2\sigma^2} \left(- \sum_{i=1}^{N} \left(t^{(i)} - \left(\beta_0 + \beta_1 x_1^{(i)} + \dots + \beta_D x_D^{(i)} \right) \right)^2 - \underbrace{\frac{2\sigma^2}{2\xi^2}}_{\lambda} \sum_{j=1}^{D} \beta_j^2 \right)$$

Formulation Ridge

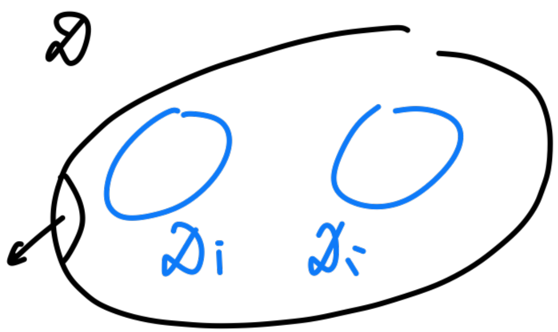


$$\frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(- \frac{x^2}{2\sigma^2} \right)$$

$$\frac{1}{2\lambda} \exp\left(\frac{-|x|}{\lambda} \right)$$

Dans le cas de la distribution de Laplace, le Maximum a posteriori retrouve l'estimateur LASSO

Equilibre "biais-variance"



Erreur quadratique moyenne $\{x, t(x)\}$

$$MSE(x) = \mathbb{E}_{\mathcal{D}_i} \left\{ \left(t(x) - h_{\mathcal{D}_i}(x) \right)^2 \right\}$$

$h_{\mathcal{D}_i}(x)$ = modèle entraîné sur les données de $\mathcal{D}_i$

$$= \mathbb{E}_{\mathcal{D}_i} \left\{ \left(t(x) - \mathbb{E}_{\mathcal{D}_i} h_{\mathcal{D}_i}(x) + \mathbb{E}_{\mathcal{D}_i} h_{\mathcal{D}_i}(x) - h_{\mathcal{D}_i}(x) \right)^2 \right\}$$

$$= \mathbb{E}_{D_i}\left\{\left(t(x) - \mathbb{E}_{D_i} h_{D_i}(x)\right)^2\right\} \leftarrow \text{bias}^2$$

$$+ \mathbb{E}_{D_i}\left\{\left(\mathbb{E}_{D_i} h_{D_i}(x) - h_{D_i}(x)\right)^2\right\} \leftarrow \text{Variance}$$

$$+ 2\,\mathbb{E}_{D_i}\left\{\left(t(x) - \mathbb{E}_{D_i} h_{D_i}(x)\right)\left(\mathbb{E}_{D_i} h_{D_i}(x) - h_{D_i}(x)\right)\right\}$$

$$= 2\left(t(x) - \mathbb{E}_{D_i} h_{D_i}(x)\right)\underbrace{\mathbb{E}_{D_i}\left\{\left(\mathbb{E}_{D_i} h_{D_i}(x) - h_{D_i}(x)\right)\right\}}$$

$$\mathbb{E}_{D_i} h_{D_i}(x) - \mathbb{E}_{D_i} h_{D_i}(x)$$

Complexité augmente → variance $\nearrow$

biais $\searrow$

Complexité diminue → variance $\searrow$

biais $\nearrow$