




## S2-UE4 - Apprentissage automatique 2

### Chap. 2 - Régression linéaire

---

Simon BERNARD  
[simon.bernard@univ-rouen.fr](mailto:simon.bernard@univ-rouen.fr)

## Introduction



## Introduction par l'exemple : Estimer l'altitude avec un thermomètre

- Expérience de Joseph D. Hooker en 1849
- Mesure de pression atmosphérique  $p_i$  et de température d'ébullition de l'eau  $t_i$  dans l'Himalaya
- Les lois de la physique disent que  $y_i = \ln(p_i)$  est (approx.) proportionnel à  $t_i$
- Donc :

$$y_i = \omega_0 + \omega_1 t_i + u_i$$

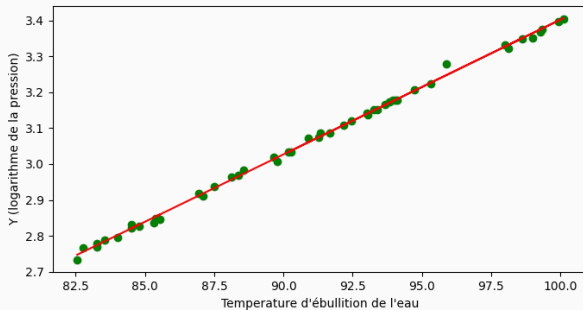
où  $u_i$  représente l'erreur de mesure (ou bruit)

- Objectif : Prédire la pression atmosphérique à partir de la température d'ébullition de l'eau (ce qui permet ensuite de déduire l'altitude, sans utiliser de baromètre)

## Introduction par l'exemple : Estimer l'altitude avec un thermomètre

Voici le tracé des points représentant les mesures, ainsi qu'une estimation de la droite :

$$y_i = \omega_0 + \omega_1 t_i$$



Les paramètres  $(\omega_0, \omega_1)$  sont estimés en minimisant l'erreur quadratique (moindre carrés)

La droite (définie par les paramètres estimés) est un modèle de regression linéaire

- Elle explique au mieux une grandeur  $Y$  (la réponse) en fonction d'autres grandeurs (variables explicatives)
- Elle permet également de séparer et quantifier les liens déterministes et les parties aléatoires ( $u_i$ )
- On peut par exemple supposer que les  $u_i$  suivent une distribution gaussienne centrée et estimer l'écart-type  $\sigma$
- Ce modèle permet ensuite de prédire n'importe quelle valeur de pression étant donné la température d'ébullition de l'eau

S'il y a plusieurs variables explicatives, on parle de régression linéaire multiple

- 1 variable explicative : **régression linéaire simple** (exemple précédent)
- Le modèle est une simple droite
- 2+ variables explicatives : **régression linéaire multiple**
- La relation entre la réponse et les variables explicatives prend la forme :

$$y_i = b + w_1 x_i^{(1)} + w_2 x_i^{(2)} + \dots + w_d x_i^{(d)} + u_i$$

Objectifs :

- Apprendre les paramètres  $w_j$  à partir de données exemples
- Déterminer les variables explicatives significatifs : la variable  $X^{(j)}$  a-t-elle une influence sur  $Y$  (i.e. est ce que  $w_j = 0$ )?
- Estimer l'erreur de prédiction du modèle

S'il y a plusieurs variables explicatives, on parle de régression linéaire multiple

- Le modèle  $h : \mathbb{R}^d \rightarrow \mathbb{R}$  à déterminer est de la forme

$$h(\mathbf{x}) = \sum_{i=1}^d w_i x^{(i)} + b = \mathbf{x}^\top \mathbf{w} + b = [\mathbf{x}^\top \ 1] \boldsymbol{\alpha}$$

avec

- $\mathbf{w} \in \mathbb{R}^d$ , un vecteur qui définit un hyperplan
- $b \in \mathbb{R}$  un biais qui déplace la fonction perpendiculairement à l'hyperplan
- $\boldsymbol{\alpha} = \begin{bmatrix} \mathbf{w} \\ b \end{bmatrix} \in \mathbb{R}^{d+1}$
- En pratique, on cherche à déterminer  $(\mathbf{w}, b)$  à partir de l'ensemble d'apprentissage  $\mathbf{X}$

## Représentation matricielle des données

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^\top & 1 \\ \mathbf{x}_2^\top & 1 \\ \vdots & \vdots \\ \mathbf{x}_i^\top & 1 \\ \vdots & \vdots \\ \mathbf{x}_n^\top & 1 \end{bmatrix} = \begin{bmatrix} x_1^{(1)} & x_1^{(2)} & \dots & x_1^{(j)} & \dots & x_1^{(d)} & 1 \\ x_2^{(1)} & x_2^{(2)} & \dots & x_2^{(j)} & \dots & x_2^{(d)} & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_i^{(1)} & x_i^{(2)} & \dots & x_i^{(j)} & \dots & x_i^{(d)} & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_n^{(1)} & x_n^{(2)} & \dots & x_n^{(j)} & \dots & x_n^{(d)} & 1 \end{bmatrix}, \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{bmatrix}$$

- $\mathbf{x}_i \in \mathbb{R}^d$  sont les observations (instances) pour  $i = 1, \dots, n$
- $y_i \in \mathbb{R}$  sont les valeurs observée à prédire (réponse) pour  $i = 1, \dots, n$
- $\mathbf{X} \in \mathbb{R}^{n \times (d+1)}$  telle que  $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n, \mathbf{e}]$  avec  $\mathbf{e} \in \mathbb{R}^d$  et  $e_i = 1, \forall i$
- $\mathbf{y} \in \mathbb{R}^n$  telle que  $\mathbf{y} = [y_1, y_2, \dots, y_n]^\top$



## Représentation matricielle du modèle

- Sous sa forme matricielle, la relation s'écrit

$$\mathbf{y} = \mathbf{X}\boldsymbol{\alpha} + \mathbf{u}$$

où  $\mathbf{u}$  est un vecteur de bruit

## Hypothèse sur $\mathbf{X}$ :

- On suppose que  $\mathbf{X}$  est de rang colonnes plein, c-à-d  $\mathbf{X}\mathbf{v} = 0$  ssi  $\mathbf{v} = 0$ .
- Dans ce cas,  $\mathbf{X}^T \mathbf{X}$  est inversible (nécessaire pour trouver l'estimateur des moindres carrés)
- Si ce n'est pas le cas, c-à-d  $\exists \mathbf{v} \neq 0$  tel que  $\mathbf{X}\mathbf{v} = 0$ , cela implique :
  - une des variables s'obtient par combinaison linéaire des autres : elle est inutile
  - pour tout estimateur  $\hat{\mathbf{w}}$ , l'estimateur  $\hat{\mathbf{w}} + \mathbf{v}$  est aussi bon : on ne peut pas estimer  $\mathbf{w}^*$  (sans hypothèses supplémentaires)

## Méthode des moindres carrées

---

Pour trouver  $h(\cdot)$ , nous cherchons à minimiser l'erreur de prédiction

- Nous rappelons que nous cherchons à estimer :

$$h(\mathbf{x}) = \mathbf{x}^\top \mathbf{w} + b$$

pour laquelle nous devons estimer les paramètres  $(\mathbf{w}, b)$

- Pour cela, on cherche à minimiser l'erreur de prédiction sur les exemples d'apprentissage, aussi appelé **résidu** :

$$\epsilon_i = y_i - h(\mathbf{x}_i)$$

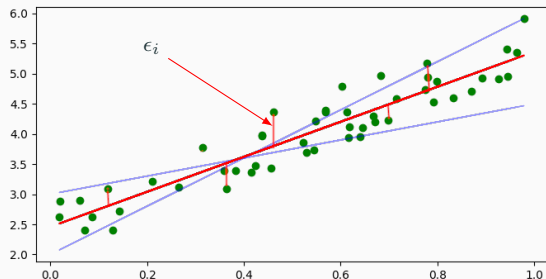
$$\epsilon_i = y_i - \mathbf{x}_i^\top \mathbf{w} - b$$

ou, sous la forme matricielle  $\boldsymbol{\epsilon} \in \mathbb{R}^n$ , tel que :

$$\boldsymbol{\epsilon} = \mathbf{y} - \mathbf{X}\boldsymbol{\alpha}$$

## Interprétation géométrique

Ce problème peut s'interpréter comme la recherche de l'hyperplan  $y = \mathbf{x}^\top \mathbf{w} + b$  passant "au mieux" (au sens des moindres carrés) parmi le nuage des observations  $(\mathbf{x}_i, y_i), i = 1, \dots, n$



Moindres carrés = minimiser le carré des résidus

- La méthode des moindres carrés consiste à minimiser la somme des résidus au carré :

$$\min_h \sum_{i=1}^n (y_i - h(\mathbf{x}_i))^2$$

$$\min_{(\mathbf{w}, b)} \sum_{i=1}^n (y_i - \mathbf{x}_i^\top \mathbf{w} - b)^2$$

ou encore :

$$\min_{\boldsymbol{\alpha}} \|\boldsymbol{\epsilon}\|^2 = \|\mathbf{y} - \mathbf{X}\boldsymbol{\alpha}\|^2$$

où  $\|\cdot\|^2$  est la norme euclidienne d'un vecteur telle que  $\|\boldsymbol{\epsilon}\|^2 = \sum_{i=1}^n \epsilon_i^2$

## Rappel d'optimisation

Nous voulons résoudre le problème d'optimisation

$$\min_{\boldsymbol{\alpha}} J(\boldsymbol{\alpha}) \quad \text{avec} \quad J(\boldsymbol{\alpha}) = \frac{1}{2} \|\mathbf{y} - \mathbf{X}\boldsymbol{\alpha}\|^2$$

Nous supposons pour le moment que la fonction  $J(\boldsymbol{\alpha})$  est convexe. Dans ce cas,  $\boldsymbol{\alpha}^*$  est un minimum de la fonction  $J(\boldsymbol{\alpha})$  si et seulement si :

$$\nabla J(\boldsymbol{\alpha}^*) = \mathbf{0}$$

où  $\nabla J(\boldsymbol{\alpha})$  est le gradient de la fonction en  $\boldsymbol{\alpha}$  tel que :

$$\nabla J(\boldsymbol{\alpha})_i = \frac{\partial J(\boldsymbol{\alpha})}{\partial \alpha_i}, \forall i$$

Note : le terme  $\frac{1}{2}$  sert juste à simplifier les calculs, mais il ne change rien au problème d'optimisation

Le problème de moindres carrés se réécrit sous la forme matricielle comme

$$\min_{\alpha} J(\alpha) \quad \text{avec} \quad J(\alpha) = \frac{1}{2} \|\mathbf{y} - \mathbf{X}\alpha\|^2$$

En développant :

$$\begin{aligned} J(\alpha) &= \frac{1}{2} \|\mathbf{y} - \mathbf{X}\alpha\|^2 \\ &= \frac{1}{2} (\mathbf{y} - \mathbf{X}\alpha)^\top (\mathbf{y} - \mathbf{X}\alpha) \\ &= \frac{1}{2} (\mathbf{y}^\top - (\mathbf{X}\alpha)^\top) (\mathbf{y} - \mathbf{X}\alpha) = \frac{1}{2} (\mathbf{y}^\top - \alpha^\top \mathbf{X}^\top) (\mathbf{y} - \mathbf{X}\alpha) \\ &= \frac{1}{2} \mathbf{y}^\top \mathbf{y} - \frac{1}{2} \mathbf{y}^\top \mathbf{X}\alpha - \frac{1}{2} \alpha^\top \mathbf{X}^\top \mathbf{y} + \frac{1}{2} \alpha^\top \mathbf{X}^\top \mathbf{X}\alpha \\ &= \frac{1}{2} \mathbf{y}^\top \mathbf{y} - \alpha^\top \mathbf{X}^\top \mathbf{y} + \frac{1}{2} \alpha^\top \mathbf{X}^\top \mathbf{X}\alpha \end{aligned}$$

car

- $(\mathbf{A} + \mathbf{B})^\top = \mathbf{A}^\top + \mathbf{B}^\top$  et  $(\mathbf{A}\mathbf{B})^\top = \mathbf{B}^\top \mathbf{A}^\top$
- $\mathbf{y}^\top \mathbf{X}\alpha = \alpha^\top \mathbf{X}^\top \mathbf{y}$  (qui est un réel)

Pour trouver la solution à ce problème d'optimisation, il faut calculer le gradient de  $J(\alpha)$

$$\begin{aligned}\frac{\partial J(\alpha)}{\partial \alpha_i} &= \frac{\partial}{\partial \alpha_i} \frac{1}{2} \mathbf{y}^\top \mathbf{y} - \alpha^\top \mathbf{X}^\top \mathbf{y} + \frac{1}{2} \alpha^\top \mathbf{X}^\top \mathbf{X} \alpha \\ &= \frac{\partial}{\partial \alpha_i} \frac{1}{2} \mathbf{y}^\top \mathbf{y} - \frac{\partial}{\partial \alpha_i} \alpha^\top \mathbf{X}^\top \mathbf{y} + \frac{\partial}{\partial \alpha_i} \frac{1}{2} \alpha^\top \mathbf{X}^\top \mathbf{X} \alpha\end{aligned}$$

En posant  $\mathbf{p} = \mathbf{X}^\top \mathbf{y}$  et  $\mathbf{M} = \mathbf{X}^\top \mathbf{X}$ , on a :

$$\begin{aligned}\frac{\partial}{\partial \alpha_i} \alpha^\top \mathbf{p} &= \frac{\partial}{\partial \alpha_i} \sum_{j=1}^{d+1} p_j \alpha_j = p_i \\ \frac{\partial}{\partial \alpha_i} \alpha^\top \mathbf{M} \alpha &= \frac{\partial}{\partial \alpha_i} \sum_{j=1}^{d+1} \sum_{k=1}^{d+1} \alpha_j \alpha_k M_{jk} = \sum_{j=1}^{d+1} \alpha_j M_{ji} + \sum_{k=1}^{d+1} \alpha_k M_{ik}\end{aligned}$$

Car  $(uv)' = uv' + u'v$  avec  $u = \alpha_j$  et  $v = \sum_{k=1}^{d+1} \alpha_k M_{jk}$



Pour trouver la solution à ce problème d'optimisation, il faut calculer le gradient de  $J(\alpha)$

$$\begin{aligned}\frac{\partial J(\alpha)}{\partial \alpha_i} &= \frac{\partial}{\partial \alpha_i} \frac{1}{2} \mathbf{y}^\top \mathbf{y} - \alpha^\top \mathbf{X}^\top \mathbf{y} + \frac{1}{2} \alpha^\top \mathbf{X}^\top \mathbf{X} \alpha \\ &= \frac{\partial}{\partial \alpha_i} \frac{1}{2} \mathbf{y}^\top \mathbf{y} - \frac{\partial}{\partial \alpha_i} \alpha^\top \mathbf{X}^\top \mathbf{y} + \frac{\partial}{\partial \alpha_i} \frac{1}{2} \alpha^\top \mathbf{X}^\top \mathbf{X} \alpha \\ &= 0 - p_i + \frac{1}{2} \sum_{j=1}^{d+1} (M_{ij} + M_{ji}) \alpha_j\end{aligned}$$

Ce qui donne sous la forme matricielle :

$$\begin{aligned}\nabla J(\alpha) &= -\mathbf{p} + \mathbf{M}\alpha \\ &= -\mathbf{X}^\top \mathbf{y} + \mathbf{X}^\top \mathbf{X} \alpha\end{aligned}$$

La minimisation de  $J(\alpha)$  est réalisée lorsque le gradient s'annule

$$\nabla J(\hat{\alpha}) = 0 \quad \Leftrightarrow \quad -\mathbf{X}^T \mathbf{y} + \mathbf{X}^T \mathbf{X} \hat{\alpha} = 0$$

La solution du problème de minimisation des moindres carrés est le vecteur  $\hat{\alpha}$  défini par :

$$\hat{\alpha} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

que l'on appelle **l'estimateur des moindres carrés**

Nous rappelons l'hypothèse sur  $\mathbf{X}$  :

- $\mathbf{X}$  est une matrice de rang  $d + 1$  et donc la matrice  $\mathbf{X}^T \mathbf{X}$  est inversible
- Dans le faits, ceci implique que  $n > d + 1$ , c-à-d que le nombre d'instances d'apprentissage est supérieur au nombre de variables explicatives (caractéristiques)

## Qualité du modèle

---

En apprentissage, on caractérise souvent la qualité d'un modèle par le compromis biais-variance

- Le **biais** est l'erreur intrinsèque, provenant d'hypothèses erronées dans l'algorithme d'apprentissage

E.g. si notre modèle est linéaire mais que la "vraie" relation ne l'est pas : erreur inévitable dû à cette inadéquation (biais non-nul)

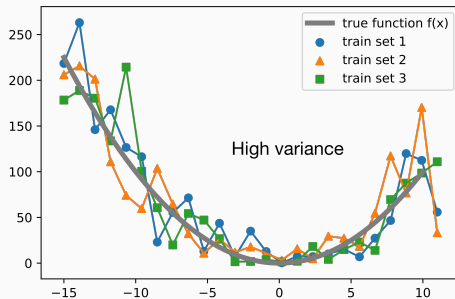
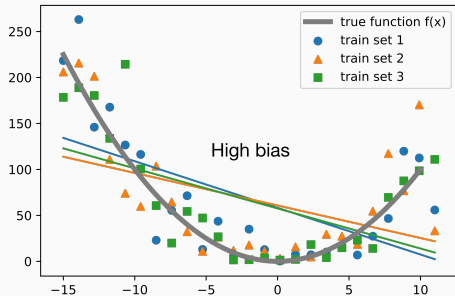
- La **variance** est l'erreur due à la sensibilité aux petites fluctuations de l'ensemble d'apprentissage

Si une méthode d'apprentissage donne 2 modèles très différents pour 2 ensembles d'apprentissage sensiblement différents, la variance sera élevée (+ augmente le risque de sur-apprentissage)

Un bon algorithme d'apprentissage vise un **bon compromis biais-variance** :

- Si la **capacité de modélisation augmente** = le biais diminue mais la variance augmente
- Si on **simplifie le modèle** = la variance diminue mais le biais augmente

En apprentissage, on caractérise souvent la qualité d'un modèle par le compromis biais-variance



## Décomposition biais-variance de l'erreur

- Formellement, le biais et la variance d'un estimateur  $\hat{\theta}$  d'une statistique  $\theta$  est défini par :

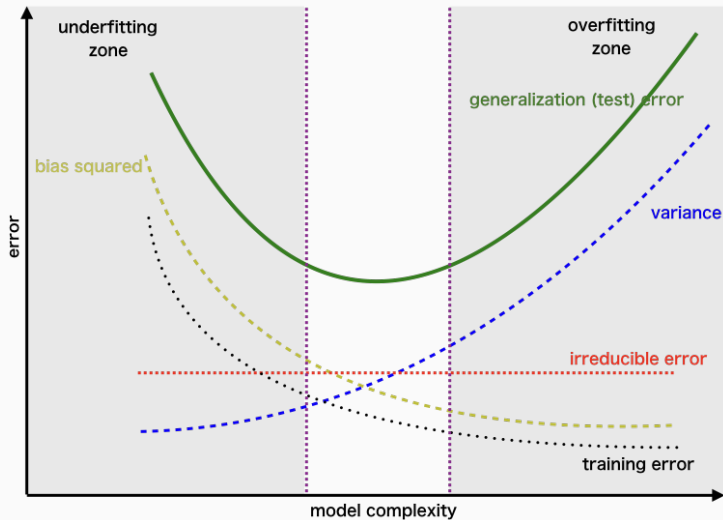
$$Bias(\hat{\theta}) = E[\hat{\theta}] - \theta \quad Var(\hat{\theta}) = E \left[ \left( E[\hat{\theta}] - \theta^2 \right)^2 \right]$$

- On peut montrer que l'erreur de prédiction attendue s'exprime en fonction du biais et de la variance du modèle appris  $\hat{h}$  :

$$\begin{aligned} E(\mathbf{x}) &= E \left[ \left( h(\mathbf{x}) - \hat{h}(\mathbf{x}) \right)^2 \right] \\ &= \left( h(\mathbf{x}) - E[\hat{h}(\mathbf{x})] \right)^2 + E \left[ \left( \hat{h}(\mathbf{x}) - E[\hat{h}(\mathbf{x})] \right)^2 \right] + E \left[ (y - h(\mathbf{x}))^2 \right] \\ &= Bias(\hat{h}(\mathbf{x}))^2 + Var(\hat{h}(\mathbf{x})) + \sigma^2 \end{aligned}$$

avec  $h(\mathbf{x})$  le "vrai" modèle et  $E[\hat{h}(\mathbf{x})]$  est l'espérance du modèle en fonction de  $\mathbf{X}$

- Le terme  $\sigma^2$  représente l'erreur incompressible, due au bruit (borne inférieure).



## Estimer le biais et la variance

- Les performances doivent être estimées sur des données de test (i.e. qui n'ont pas été utilisées en apprentissage)  
→ Séparer les données en 2 : un sous-ensemble d'apprentissage, un autre de test
- Les espérances mathématiques précédentes sont à estimer pour plusieurs modèles obtenus sur plusieurs ensembles d'apprentissages  
→ Les phases apprentissage/test sont répétées pour plusieurs de ces découpages
- L'erreur, le biais et la variance caractérisent un algorithme d'apprentissage  
→ L'algorithme d'apprentissage est le même pour toutes ces répétitions
- On ne peut pas tenir compte de  $\sigma^2$  dans ces estimations  
→  $h(\mathbf{x})$  est remplacé par les  $y$  associées aux données (supervision)



On peut vouloir estimer la qualité d'un seul modèle en mesurant ses performances

- Les deux mesures suivantes sont souvent utilisées pour estimer les performances d'un modèle de regression (sur des données de test)

## Erreur quadratique moyenne

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

- 0 quand la prédiction est parfait
- Pas une mesure normalisée (dépend de la variance de  $\mathbf{y}$ )

avec  $\mathbf{y}$  les valeurs à prédire et  $\hat{\mathbf{y}}$  les prédictions

## Coefficient de corrélation

$$r = \frac{\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sigma_y \sigma_{\hat{y}}}$$

- $\bar{y} = \frac{1}{n} \sum_i y_i$  moyenne de  $\mathbf{y}$  et  $\sigma_y$  sa variance
- 1 quand la prédiction est parfaite
- Mesure normalisée

## Méthode des moindres carrées régularisés

---

La méthode précédente s'appelle la méthode des moindres carrés "ordinaires"

- Pour rappel, on cherche à minimiser :

$$\min_{\boldsymbol{\alpha}} J(\boldsymbol{\alpha}) \quad \text{avec} \quad J(\boldsymbol{\alpha}) = \frac{1}{2} \|\mathbf{y} - \mathbf{X}\boldsymbol{\alpha}\|^2$$

- La solution du problème est :

$$\hat{\boldsymbol{\alpha}} = (\mathbf{X}^{\top} \mathbf{X})^{-1} \mathbf{X}^{\top} \mathbf{y}$$

- **Problème :**

- Lorsque  $n < d + 1$ , la matrice  $\mathbf{X}^{\top} \mathbf{X}$  n'est pas inversible
- La solution n'est pas unique : le problème est mal posé

- **Solution : régularisation**

L'idée est d'ajouter une contrainte sur les paramètres à estimer

- Cette contrainte prend la forme d'un terme qui s'ajoute à la fonction objective  $J(\alpha)$
- Du point de vue optim., ce terme permet de résoudre le problème de la matrice  $\mathbf{X}^T \mathbf{X}$  qui n'est pas inversible et de trouver une solution analytique
- Du point de vue ML, ce terme vise à **pénaliser les modèles trop complexes** :
  - Modèle complexe : fort risque de sur-apprentissage
  - Plusieurs solutions : on favorise la solution la moins complexe
- L'influence de ce terme est contrôlée par un hyperparamètre de régularisation  $\lambda$  (**coefficient de régularisation**). Pour  $\lambda = 0$ , on retrouve le problème d'optimisation des moindres carrés ordinaire.

Pour cette méthode, le régularisateur vise à minimiser la norme de  $\mathbf{w}$

$$\min_{\mathbf{w}, b} \quad \frac{1}{2} \sum_{i=1}^n (y_i - \mathbf{x}_i^\top \mathbf{w} - b)^2 + \frac{\lambda}{2} \|\mathbf{w}\|^2$$

- Cette régularisation est la **régularisation de Tikhonov**
- Elle a pour effet de promouvoir les paramètres  $\mathbf{w}$  de norme minimale et de rendre le problème strictement convexe
- L'hyperparamètre  $\lambda$  permet de limiter le sur-apprentissage s'il est choisi judicieusement
- $\lambda = 0$  permet de revenir à la régression des moindres carrés
- La méthode de regression résultante s'appelle la **regression ridge** (ou regression de crête)

## Version matricielle

$$\min_{\boldsymbol{\alpha}} J(\boldsymbol{\alpha}) \quad \text{avec} \quad J(\boldsymbol{\alpha}) = \frac{1}{2} \|\mathbf{y} - \mathbf{X}\boldsymbol{\alpha}\|^2 + \frac{\lambda}{2} \boldsymbol{\alpha}^\top \mathbf{S} \boldsymbol{\alpha}$$

avec  $\mathbf{S} \in \mathbb{R}^{(d+1) \times (d+1)}$ , une matrice dont le terme général est :

$$S_{i,j} = \begin{cases} 1 & \text{si } i = j \text{ et } i \leq d \\ 0 & \text{sinon} \end{cases}$$

$\mathbf{S}$  est donc une matrice diagonale unitaire dont le dernier terme diagonal est nul. On trouve donc :

$$\boldsymbol{\alpha}^\top \mathbf{S} \boldsymbol{\alpha} = \sum_{i,j=1}^{d+1} \alpha_i \alpha_j S_{i,j} = \sum_{i=1}^d \alpha_i^2 = \sum_{i=1}^d \mathbf{w}_i^2 = \|\mathbf{w}\|^2$$

Dérivées partielles de  $J(\alpha)$ 

En reprenant la ré-écriture de la méthode des MCO, on a :

$$\begin{aligned} J(\alpha) &= \frac{1}{2} \mathbf{y}^\top \mathbf{y} - \alpha^\top \mathbf{X}^\top \mathbf{y} + \frac{1}{2} \alpha^\top \mathbf{X}^\top \mathbf{X} \alpha + \frac{\lambda}{2} \alpha^\top \mathbf{S} \alpha \\ &= \frac{1}{2} \mathbf{y}^\top \mathbf{y} - \alpha^\top \mathbf{X}^\top \mathbf{y} + \frac{1}{2} \alpha^\top (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{S}) \alpha \end{aligned}$$

En posant toujours  $\mathbf{p} = \mathbf{X}^\top \mathbf{y}$ , mais cette fois  $\mathbf{M} = \mathbf{X}^\top \mathbf{X} + \lambda \mathbf{S}$ , on retrouve :

$$\begin{aligned} \frac{\partial J(\alpha)}{\partial \alpha_i} &= \frac{\partial}{\partial \alpha_i} \frac{1}{2} \mathbf{y}^\top \mathbf{y} - \frac{\partial}{\partial \alpha_i} \alpha^\top \mathbf{X}^\top \mathbf{y} + \frac{\partial}{\partial \alpha_i} \frac{1}{2} \alpha^\top (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{S}) \alpha \\ &= 0 - p_i + \frac{1}{2} \sum_{j=1}^{d+1} (M_{ij} + M_{ji}) \alpha_j \end{aligned}$$

ou sous la forme matricielle

$$\begin{aligned} \nabla J(\alpha) &= -\mathbf{p} + \mathbf{M} \alpha \\ &= -\mathbf{X}^\top \mathbf{y} + (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{S}) \alpha \end{aligned}$$

## Minimisation de $J(\alpha)$

La minimisation de  $J(\alpha)$  est réalisée lorsque le gradient s'annule :

$$\nabla J(\hat{\alpha}) = 0 \quad \Leftrightarrow \quad -\mathbf{X}^T \mathbf{y} + (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{S}) \hat{\alpha} = 0$$

La solution du problème de minimisation des moindres carrés est le vecteur  $\hat{\alpha}$  défini par :

$$\hat{\alpha} = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{S})^{-1} \mathbf{X}^T \mathbf{y}$$

## Régularisation :

- La matrice  $\mathbf{S}$  ajoute  $\lambda$  à tous les éléments de la diagonale de  $\mathbf{X}^T \mathbf{X}$ , ce qui la rend inversible.
- Le problème est maintenant bien posé et une solution unique existe



## Trouver la valeur de $\lambda$

- $\lambda$  contrôle la "quantité" de régularisation
- Pour chaque valeur de  $\lambda$  on a une solution différente
- On souhaite bien sûr **trouver la valeur de  $\lambda$  qui minimise la performance** (MSE)
- Il existe des calculs basés sur des hypothèses simplificatrices mais ils ne tiennent généralement pas compte de la qualité sur les données étudiées (i.e. n'utilise pas la supervision)
- Solution : **mesurer les performances sur des données de tests pour plusieurs valeurs de  $\lambda$  et retenir celle qui a permis d'obtenir les meilleurs résultats**
- Techniques pour fiabiliser : *cross-validation, bootstrap, leave-one-out*

Nous reviendrons sur ces techniques dans le chapitre sur la sélection de modèles...

Une autre méthode de régularisation populaire est la méthode LASSO

- Norme  $\ell_p$  :

$$\|\mathbf{x}\|_p = \left( \sum_{i=1}^n |x_i|^p \right)^{1/p}$$

- Regression Ridge : basé sur la norme  $\ell_2$  (norme euclidienne)
- Regression LASSO (*least absolute shrinkage and selection operator*) : basé sur la norme  $\ell_1$  :

$$\min_{\mathbf{w}, b} \quad \frac{1}{2} \sum_{i=1}^n (y_i - \mathbf{x}_i^\top \mathbf{w} - b)^2 + \frac{\lambda}{2} \|\mathbf{w}\|_1$$

- Contrairement à la regression Ridge, LASSO peut sélectionner les variables explicatives en autorisant certains  $w_i$  à 0
- Mais pas de calcul direct pour LASSO : algorithmes itératifs qui font évoluer les paramètres jusqu'à obtenir une solution

## Combiner les deux méthodes : Elastic net

$$\min_{\mathbf{w}, b} \quad \frac{1}{2} \sum_{i=1}^n (y_i - \mathbf{x}_i^\top \mathbf{w} - b)^2 + \frac{\lambda_1}{2} \|\mathbf{w}\|_2^2 + \frac{\lambda_2}{2} \|\mathbf{w}\|_1$$

avec  $(\lambda_1 = 0 \text{ et } \lambda_2 > 0)$  : Ridge;  $(\lambda_1 > 0 \text{ et } \lambda_2 = 0)$  : LASSO;  $(\lambda_1 = 0 \text{ et } \lambda_2 = 0)$  : MCO

ou

$$\min_{\mathbf{w}, b} \quad \frac{1}{2} \sum_{i=1}^n (y_i - \mathbf{x}_i^\top \mathbf{w} - b)^2 + \frac{\lambda}{2} (\omega \|\mathbf{w}\|_2^2 + (1 - \omega) \|\mathbf{w}\|_1)$$

avec  $(\omega = 0)$  : Ridge;  $(\omega = 1)$  : LASSO

- Corrige un défaut de LASSO quand  $d$  est grand et  $n$  petit : sélection des variables non pertinentes ( $n$  au maximum)
- Pas de calcul direct mais plusieurs algorithmes itératifs.

## Annexes

---

On veut calculer :

$$\frac{\partial J(\boldsymbol{\alpha})}{\partial \alpha_i} = \frac{\partial}{\partial \alpha_i} \frac{1}{2} \mathbf{y}^\top \mathbf{y} - \frac{\partial}{\partial \alpha_i} \boldsymbol{\alpha}^\top \mathbf{X}^\top \mathbf{y} + \frac{\partial}{\partial \alpha_i} \frac{1}{2} \boldsymbol{\alpha}^\top \mathbf{X}^\top \mathbf{X} \boldsymbol{\alpha}$$

Pour calculer le troisième terme, on pose  $\mathbf{M} = \mathbf{X}^\top \mathbf{X}$  et on utilise  $(uv)' = uv' + u'v$  avec  $u = \alpha_j$  et  $v = \sum_{k=1}^{d+1} \alpha_k M_{jk}$  :

$$\begin{aligned} \frac{\partial}{\partial \alpha_i} \boldsymbol{\alpha}^\top \mathbf{M} \boldsymbol{\alpha} &= \frac{\partial}{\partial \alpha_i} \sum_{j=1}^{d+1} \sum_{k=1}^{d+1} \alpha_j \alpha_k M_{jk} = \sum_{j=1}^{d+1} \frac{\partial}{\partial \alpha_i} \left( \alpha_j \sum_{k=1}^{d+1} \alpha_k M_{jk} \right) = \sum_{j=1}^{d+1} (uv)' \\ &= \sum_{j=1}^{d+1} (u'v + uv') = \sum_{j=1}^{d+1} u'v + \sum_{j=1}^{d+1} uv' \end{aligned}$$

Comme  $u' = 0$  pour  $j \neq i$  et  $u' = 1$  pour  $j = i$ , on a

$$\sum_{j=1}^{d+1} u'v = v \text{ avec } j = i \Rightarrow \sum_{j=1}^{d+1} u'v = \sum_{k=1}^{d+1} \alpha_k M_{ik}$$

Et comme  $v' = 0$  pour  $k \neq i$  et  $v' = M_{ji}$  pour  $k = i$

$$\sum_{j=1}^{d+1} uv' = \sum_{j=1}^{d+1} u M_{ji} \Rightarrow \sum_{j=1}^{d+1} uv' = \sum_{j=1}^{d+1} \alpha_j M_{ji}$$

Et donc

$$\frac{\partial}{\partial \alpha_i} \boldsymbol{\alpha}^\top \mathbf{M} \boldsymbol{\alpha} = \sum_{j=1}^{d+1} \alpha_j M_{ji} + \sum_{k=1}^{d+1} \alpha_k M_{ik}$$