Chapitre 8 : Régression Linéaire et Régularisation

26 février 2026

Plan du chapitre

Le Phénomène de Surapprentissage (Overfitting)

Un modèle trop flexible conduit inévitablement au **surapprentissage**.

Le Compromis Biais-Variance

L'erreur de généralisation d'un estimateur $\hat{f}$ se décompose analytiquement :

$$\mathbb{E}\Big[(y - \hat{f}(\mathbf{x}))^2\Big] = \underbrace{\Big(\mathbb{E}[\hat{f}(\mathbf{x})] - f(\mathbf{x})\Big)^2}_{\text{Biais}^2} + \underbrace{\mathsf{Var}(\hat{f}(\mathbf{x}))}_{\text{Variance}} + \underbrace{\sigma^2}_{\text{Bruit irréductible}}$$

Dans les modèles linéaires, le surapprentissage se manifeste par **l'explosion de la norme du vecteur des poids** $\|\mathbf{w}\|$.
Les coefficients deviennent grands, avec des signes opposés, pour annuler localement l'erreur sur l'échantillon d'apprentissage.

Le Cadre Général de la Régularisation

Soit la matrice de design $\mathbf{X} \in \mathbb{R}^{n \times d}$ et le vecteur cible $\mathbf{y} \in \mathbb{R}^n$.
L'estimateur des Moindres Carrés Ordinaires (MCO / OLS) minimise $\frac{1}{2}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2$.
Si les variables sont fortement corrélées (multicolinéarité) ou si $d > n$, la matrice $\mathbf{X}^\top \mathbf{X}$ est singulière ou mal conditionnée. L'estimateur analytique $\hat{\mathbf{w}}_{\text{MCO}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$ est instable ou n'existe pas.

Le Problème d'Optimisation Régularisé

Pour stabiliser le système, on ajoute un terme de **pénalité structurelle** $\Omega(\mathbf{w})$:

$$\hat{\mathbf{w}} = \underset{\mathbf{w} \in \mathbb{R}^d}{\arg\min} \underbrace{\frac{1}{2}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2}_{\text{Ajustement aux données}} + \underbrace{\lambda \Omega(\mathbf{w})}_{\text{Régularisation}}$$

L'hyperparamètre $\lambda \geq 0$ contrôle le compromis entre le biais (fort si λ grand) et la variance (forte si λ proche de zéro).

Les Séparatrices Affines et l'Intercept

Nous considérons la classe des fonctions affines :

$$h_{\mathbf{w},b}(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle + b \quad \text{avec } \mathbf{w} \in \mathbb{R}^d \text{ et } b \in \mathbb{R}$$

L'équation $\langle \mathbf{w}, \mathbf{x} \rangle + b = 0$ définit un **hyperplan** affine de dimension $d - 1$, orthogonal au vecteur normal $\mathbf{w}$.

L'astuce de l'Intercept (Biais)

Pour simplifier l'algèbre linéaire, on intègre systématiquement le biais b dans le vecteur des poids en ajoutant une coordonnée constante 1 au vecteur $\mathbf{x}$:

$$h_{\mathbf{w}}(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$$

On note désormais d la dimension totale incluant l'intercept : $\mathbf{w} \in \mathbb{R}^d$, $\mathbf{x} \in \mathbb{R}^d$. Le paramètre $w_0 = b$ est appelé **intercept**.

Régression Linéaire Multiple sous forme Matricielle

Pour n observations et d caractéristiques, on définit :

- $\mathbf{X} \in \mathbb{R}^{n \times d}$ la **matrice de design** (contenant les exemples $\mathbf{x}_i^\top$ en lignes).
- $\mathbf{y} \in \mathbb{R}^n$ le vecteur cible continus.

Estimateur des Moindres Carrés Ordinaires (MCO)

L'apprentissage consiste à minimiser l'Erreur Quadratique Moyenne (EQM / MSE) :

$$\mathscr{L}(\mathbf{w}) = \frac{1}{2}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 = \frac{1}{2}(\mathbf{y} - \mathbf{X}\mathbf{w})^\top(\mathbf{y} - \mathbf{X}\mathbf{w})$$

En annulant le gradient $\nabla_{\mathbf{w}}\mathscr{L}(\mathbf{w}) = -\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\mathbf{w}) = \mathbf{0}$, on obtient :

$$(\mathbf{X}^\top\mathbf{X})\mathbf{w} = \mathbf{X}^\top\mathbf{y} \quad \Longrightarrow \quad \hat{\mathbf{w}}_{\text{MCO}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$$

Limites de l'Estimateur MCO

L'estimateur analytique $\hat{\mathbf{w}}_{\text{MCO}}$ est vulnérable en pratique, notamment en grande dimension :

1. **Singularité matricielle ($n < d$) :** Si le nombre de variables dépasse le nombre d'observations, la matrice $\mathbf{X}^\top \mathbf{X}$ n'est pas de rang plein. Elle est **non-inversible**.
2. **Multicolinéarité (Instabilité) :** Si deux variables sont fortement corrélées, certaines valeurs propres de $\mathbf{X}^\top \mathbf{X}$ tendent vers 0. Lors de l'inversion, la **variance de l'estimateur explose**.

Pour forcer l'ajustement au bruit, les poids $\hat{w}_j$ prennent des valeurs très grandes de signes opposés (ex : $+10^6$ et -10^6) pour s'annuler artificiellement.

Solution : Pénaliser la norme du vecteur $\mathbf{w}$ via la **Régularisation**.

Régularisation ℓ_2 (Ridge, Tikhonov)

On introduit un **biais** dans l'estimation pour réduire fortement la **variance**.
On pénalise par la norme Euclidienne au carré :

$$\mathscr{L}_{\text{Ridge}}(\mathbf{w}) = \frac{1}{2}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \frac{\lambda}{2}\|\mathbf{w}\|_2^2$$

où $\lambda > 0$ contrôle l'intensité de la pénalité. En dérivant et en annulant le gradient :

$$-\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\mathbf{w}) + \lambda\mathbf{w} = \mathbf{0} \quad \Longrightarrow \quad \hat{\mathbf{w}}_{\text{Ridge}} = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y}$$

Restauration de l'inversibilité

L'ajout de $\lambda\mathbf{I}$ sur la diagonale garantit que toutes les valeurs propres de la matrice à inverser sont décalées et minorées par $\lambda > 0$.

Le problème de singularité est résolu.

Ridge : Preuve du Filtrage par la SVD

Comment le paramètre λ réduit-il concrètement la variance ?
Utilisons la Décomposition en Valeurs Singulières (SVD) : $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$. Soient σ_j^2 les valeurs propres de $\mathbf{X}^\top\mathbf{X}$. Les prédictions du modèle Ridge deviennent :

Facteur d'atténuation spectral (*Shrinkage*)

$$\hat{\mathbf{y}}_{\text{Ridge}} = \mathbf{X}\hat{\mathbf{w}}_{\text{Ridge}} = \sum_{j=1}^{\min(n,d)} \underbrace{\left(\frac{\sigma_j^2}{\sigma_j^2 + \lambda}\right)}_{\text{Shrinkage Factor}} \mathbf{u}_j \mathbf{u}_j^\top \mathbf{y}$$

- Si $\sigma_j^2 \gg \lambda$ (L'axe j porte une forte variance de signal utile) : Le facteur ≈ 1. L'information est préservée.
- Si $\sigma_j^2 \ll \lambda$ (L'axe j est dominé par du bruit colinéaire) : Le facteur tend vers 0.
  Ridge écrase l'influence des axes les moins informatifs.

Régularisation ℓ_1 (LASSO)

Le **LASSO** (*Least Absolute Shrinkage and Selection Operator*, Tibshirani 1996) utilise la norme de Manhattan :

$$\Omega(\mathbf{w}) = \|\mathbf{w}\|_1 = \sum_{j=1}^{d} |w_j|.$$

La norme ℓ_1 n'est pas différentiable en 0. Il n'existe pas de solution analytique globale sous forme matricielle. On doit recourir à la théorie convexe non-lisse et à la notion de **sous-différentiel** $\partial|w_j|$.

$$\partial|w_j| = \begin{cases} \{1\} & \text{si } w_j > 0 \\ \{-1\} & \text{si } w_j < 0 \\ [-1, 1] & \text{si } w_j = 0 \end{cases}$$

C'est précisément ce caractère non-lisse (la singularité en zéro) qui donne au LASSO sa capacité à générer de la **parcimonie**.

Régularisation ℓ_1 (LASSO)

Le risque quadratique pénalisé par la norme ℓ_1 s'écrit

$$\mathscr{L}_{\text{LASSO}}(\mathbf{w}) = \frac{1}{2}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \lambda\|\mathbf{w}\|_1$$

La valeur absolue n'est **pas dérivable en zéro**. Il n'y a **pas de solution matricielle fermée**. L'optimisation requiert des algorithmes non-lisses exploitant le **sous-différentiel** $\partial|w_j|$ (qui vaut la plage $[-1, 1]$ en 0).

Parcimonie (Sparsity)

Contrairement à Ridge qui réduit les coefficients sans jamais les annuler, la géométrie du LASSO force les coefficients des variables inutiles à valoir **exactement zéro**.
Il effectue simultanément la régularisation ET la **sélection de variables**.

Seuillage Doux (Soft-Thresholding)

Pour prouver l'émergence de la parcimonie, étudions le cas idéal où les caractéristiques sont **orthonormées** ($\mathbf{X}^\top \mathbf{X} = \mathbf{I}_d$). En annulant le sous-gradient, on obtient une solution analytique coordonnée par coordonnée :

Solution analytique du LASSO (cas orthogonal)

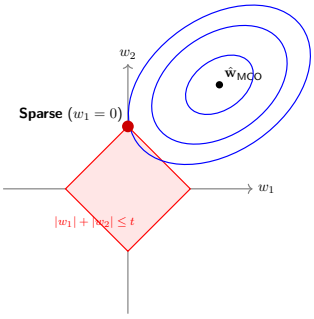
$$\hat{w}_j^{\text{LASSO}} = \text{sign}(\hat{w}_j^{\text{MCO}}) \cdot \max\left(0, |\hat{w}_j^{\text{MCO}}| - \lambda\right)$$

Mécanique du Rasoir d'Occam : Si l'estimation classique initiale $\hat{w}_j^{\text{MCO}}$ d'une variable est en-dessous du seuil λ (faible pouvoir prédictif), la soustraction devient négative et la fonction $\max(0, \cdot)$ **l'annule purement et simplement** ($w_j = 0$). Le modèle s'élague lui-même.

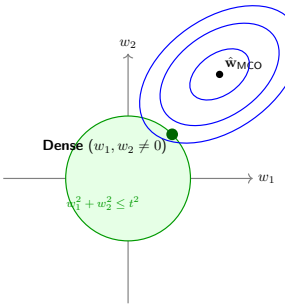
Géométrie des Contraintes (ℓ_1 vs ℓ_2)

Par dualité de Lagrange, le problème régularisé équivaut à minimiser l'erreur MCO sous contrainte spatiale rigide $\|\mathbf{w}\|_p \leq t$.

Contrainte ℓ_1 (LASSO)



Contrainte ℓ_2 (RIDGE)

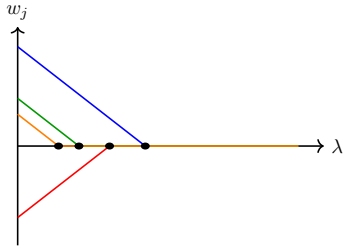


Intuition : La boule ℓ_1 est un polytope avec des "coins" posés sur les axes. En gonflant, les isocontours de la MSE rencontreront le domaine presque toujours sur ces singularités, forçant l'annulation des autres variables.

Chemins de Régularisation (Regularization Paths)

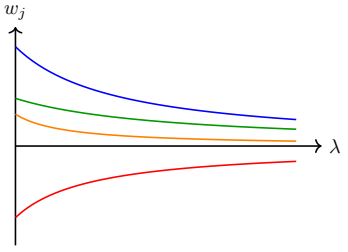
Traçons l'évolution dynamique de la valeur des coefficients w_j à mesure que la force de pénalité λ augmente.

Chemins LASSO (ℓ_1)



Atténuation linéaire : chaque variable est **ciblée et tuée net** à un seuil critique de λ pour libérer de la parcimonie.

Chemins Ridge (ℓ_2)



Atténuation asymptotique : les poids décroissent tous vers 0 de manière proportionnelle **sans jamais s'annuler**.

Synthèse et Ouverture : L'Elastic-Net

Bien que très utilisé, le LASSO présente deux limites théoriques :

1. En génomique par exemple ($d \gg n$), il sature géométriquement et ne peut sélectionner au maximum que n variables.
2. Face à un groupe de variables **fortement corrélées**, LASSO en choisit une au hasard et élimine les autres. Ridge, au contraire, distribue le poids équitablement (Effet de groupe).

Le compromis parfait : Elastic-Net (Zou & Hastie, 2005)

On combine linéairement les deux pénalités via un hyperparamètre de mélange $\alpha \in [0, 1]$:

$$P(\mathbf{w}) = \alpha \|\mathbf{w}\|_1 + \frac{1-\alpha}{2} \|\mathbf{w}\|_2^2$$

Géométriquement, l'Elastic-Net est un diamant aux arêtes strictement convexes, induisant la parcimonie tout en stabilisant les groupes de variables colinéaires.

Extension à une régression polynomiale

Comment modéliser une relation fortement non-linéaire tout en conservant l'élégance analytique du modèle linéaire ?

On applique un plongement des caractéristiques (*Feature Mapping*) via une fonction Ψ qui projette $x \in \mathbb{R}$ dans $\mathbb{R}^{s+1}$:

$$\Psi : x \longmapsto (1, x, x^2, \dots, x^s)^\top$$

L'ajustement d'un polynôme de degré s s'écrit alors comme un simple produit scalaire :

$$p(x) = w_0 + w_1 x + \cdots + w_s x^s = \mathbf{w}^\top \Psi(x)$$

Astuce fondamentale

Une régression polynomiale (fortement non-linéaire en x) reste un problème d'optimisation **strictement linéaire et convexe** par rapport aux paramètres $\mathbf{w}$.

L'équation MCO s'applique telle quelle.

Illustration : Biais, Variance et Surapprentissage