Modèles linéaires

Du risque empirique à la régularisation bayésienne

Pierre-Luc Bacon

IFT6390 – Fondements de l'apprentissage machine

Plan de la présentation

1. **Minimisation du risque empirique (MRE)**
2. **Prédicteur de Bayes optimal** : pourquoi la perte L2 donne la moyenne conditionnelle
3. **Moindres carrés ordinaires (MCO)** : solution analytique
4. **Décomposition en valeurs singulières (DVS)** : comprendre et résoudre MCO
5. **Généralisation** : surapprentissage et compromis biais-variance
6. **Régression Ridge** : régularisation L2
7. **Cadre probabiliste** : prédiction bayésienne
8. **Du maximum de vraisemblance au MAP** : lien entre Ridge et a priori gaussien

Apprentissage supervisé : le problème

Données d'entraînement : $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$ tirées i.i.d. de $p(\mathbf{x}, y)$

Objectif : Trouver une fonction $f \in \mathcal{H}$ qui prédit bien sur de **nouvelles** données

Classe d'hypothèses $\mathcal{H}$, notre espace de recherche :

- Fonctions linéaires : $f(\mathbf{x}) = \boldsymbol{\theta}^\top \mathbf{x} + b$
- Polynômes de degré k : $f(x) = \theta_0 + \theta_1 x + \cdots + \theta_k x^k$
- Réseaux de neurones : $f(\mathbf{x}) = \sigma(\boldsymbol{\Theta}_2 \sigma(\boldsymbol{\Theta}_1 \mathbf{x}))$

Le choix de $\mathcal{H}$ encode nos **hypothèses** sur la forme de la relation entre $\mathbf{x}$ et y.

Le risque : ce que nous voulons minimiser

Pour une fonction de perte $\ell(y, \hat{y})$, le **risque** (ou erreur de généralisation) est :

$$\mathcal{R}(f) = \mathbb{E}_{p(\mathbf{x}, y)}[\ell(y, f(\mathbf{x}))] = \int \ell(y, f(\mathbf{x}))\, p(\mathbf{x}, y)\, d\mathbf{x}\, dy$$

Cette quantité représente l'erreur moyenne sur **toutes les données possibles**, pondérée par leur probabilité.

Exemples de pertes :

Type	Formule	Usage
Quadratique	$\ell(y, \hat{y}) = (y - \hat{y})^2$	Régression
0-1	$\ell(y, \hat{y}) = \mathbf{1}[y \neq \hat{y}]$	Classification

Problème : La distribution $p(\mathbf{x}, y)$ est **inconnue**.

Le risque empirique : ce que nous pouvons calculer

Puisque le risque est inaccessible, nous l'**approximons** par la moyenne sur les données disponibles :

$$\hat{\mathcal{R}}(f, \mathcal{D}) = \frac{1}{N} \sum_{i=1}^{N} \ell(y_i, f(\mathbf{x}_i))$$

Perte	Risque empirique	Nom usuel
$(y - \hat{y})^2$	$\frac{1}{N} \sum_i (y_i - f(\mathbf{x}_i))^2$	Erreur quadratique moyenne
$\mathbf{1}[y \neq \hat{y}]$	$\frac{1}{N} \sum_i \mathbf{1}[y_i \neq f(\mathbf{x}_i)]$	Taux d'erreur

Propriété : Par la loi des grands nombres, $\hat{\mathcal{R}}(f) \xrightarrow{N \to \infty} \mathcal{R}(f)$

Le risque empirique est un estimateur **sans biais** du vrai risque.

Principe de minimisation du risque empirique (MRE)

Idée : Choisir la fonction qui minimise l'erreur sur les données d'entraînement

$$\hat{f} = \arg\min_{f \in \mathcal{H}} \hat{\mathcal{R}}(f, \mathcal{D}) = \arg\min_{f \in \mathcal{H}} \frac{1}{N} \sum_{i=1}^{N} \ell(y_i, f(\mathbf{x}_i))$$

Espoir : Si $\hat{\mathcal{R}}(\hat{f})$ est faible, alors $\mathcal{R}(\hat{f})$ l'est aussi.

Réalité : Ce n'est pas toujours le cas. L'écart $\mathcal{R}(\hat{f}) - \hat{\mathcal{R}}(\hat{f})$ peut être grand.

- Cet écart dépend de la **taille** de l'échantillon N
- Il dépend aussi de la **complexité** de la classe $\mathcal{H}$
- Ce phénomène s'appelle le **surapprentissage**

Le prédicteur de Bayes optimal

Question : Si nous connaissions la vraie distribution $p(\mathbf{x}, y)$, quel serait le meilleur prédicteur?

Le **prédicteur de Bayes optimal** minimise le risque pour chaque $\mathbf{x}$ individuellement :

$$f^*(\mathbf{x}) = \arg\min_{\hat{y}} \mathbb{E}_{p(y|\mathbf{x})}[\ell(y, \hat{y})]$$

Ce prédicteur constitue un **repère théorique** : aucun algorithme ne peut faire mieux, car il suppose l'accès à la vraie distribution conditionnelle $p(y|\mathbf{x})$.

La différence $\mathcal{R}(\hat{f}) - \mathcal{R}(f^*)$ mesure ce que nous perdons en ne connaissant pas p.

Cas de la perte quadratique (L2)

Pour $\ell(y, \hat{y}) = (y - \hat{y})^2$, développons l'espérance conditionnelle :

$$\mathbb{E}[(y - \hat{y})^2 | \mathbf{x}] = \mathbb{E}[y^2 | \mathbf{x}] - 2\hat{y}\mathbb{E}[y | \mathbf{x}] + \hat{y}^2$$

Cette expression est une **parabole** en $\hat{y}$ (convexe, car le coefficient de $\hat{y}^2$ est positif).

Condition d'optimalité, dérivée nulle :

$$\frac{\partial}{\partial \hat{y}} \mathbb{E}[(y - \hat{y})^2 | \mathbf{x}] = -2\mathbb{E}[y | \mathbf{x}] + 2\hat{y} = 0$$

$$\boxed{\hat{y}^* = \mathbb{E}[y | \mathbf{x}]}$$

Pour la perte L2, le prédicteur de Bayes optimal est la moyenne conditionnelle.

Prédicteurs optimaux selon la perte

Chaque fonction de perte définit son propre prédicteur optimal :

Perte	Formule	Prédicteur optimal
Quadratique	$(y - \hat{y})^2$	Moyenne : $\mathbb{E}[y\|\mathbf{x}]$
Absolue	$\|y - \hat{y}\|$	Médiane : $\text{med}(y\|\mathbf{x})$
0-1 (classif.)	$\mathbf{1}[y \neq \hat{y}]$	Mode : $\arg\max_c p(y = c\|\mathbf{x})$

Risque de Bayes, l'erreur irréductible :

$$\mathcal{R}^* = \mathcal{R}(f^*) = \mathbb{E}[\text{Var}(y|\mathbf{x})]$$

Ce risque représente le bruit intrinsèque dans les données. **Aucun algorithme ne peut faire mieux**, peu importe la quantité de données ou la puissance de calcul.

Le modèle linéaire

Nous cherchons une fonction de la forme :

$$f(\mathbf{x}; \boldsymbol{\theta}) = \boldsymbol{\theta}^\top \mathbf{x} = \sum_{j=1}^{d} \theta_j x_j$$

Notation matricielle pour N exemples et d caractéristiques :

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^\top \\ \mathbf{x}_2^\top \\ \vdots \\ \mathbf{x}_N^\top \end{bmatrix} \in \mathbb{R}^{N \times d}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} \in \mathbb{R}^N$$

Les prédictions pour tous les exemples s'écrivent $\hat{\mathbf{y}} = \mathbf{X}\boldsymbol{\theta}$.

Le biais b peut être absorbé en ajoutant une colonne de 1 à $\mathbf{X}$.

Fonction objectif : somme des carrés des résidus

Nous voulons minimiser la **somme des carrés des résidus** (SCR) :

$$\mathrm{SCR}(\boldsymbol{\theta}) = \sum_{n=1}^{N} (y_n - \boldsymbol{\theta}^\top \mathbf{x}_n)^2 = \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2$$

En développant la norme au carré :

$$\mathrm{SCR}(\boldsymbol{\theta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) = \mathbf{y}^\top \mathbf{y} - 2\boldsymbol{\theta}^\top \mathbf{X}^\top \mathbf{y} + \boldsymbol{\theta}^\top \mathbf{X}^\top \mathbf{X}\boldsymbol{\theta}$$

Cette fonction est **quadratique convexe** en $\boldsymbol{\theta}$:

- Terme constant : $\mathbf{y}^\top \mathbf{y}$
- Terme linéaire : $-2\mathbf{X}^\top \mathbf{y}$
- Terme quadratique : $\mathbf{X}^\top \mathbf{X}$ (matrice semi-définie positive)

Dérivation : les équations normales

Gradient par rapport à $\boldsymbol{\theta}$:

$$\nabla_{\boldsymbol{\theta}} \mathrm{SCR} = -2\mathbf{X}^\top \mathbf{y} + 2\mathbf{X}^\top \mathbf{X}\boldsymbol{\theta}$$

Condition d'optimalité, gradient nul :

$$-2\mathbf{X}^\top \mathbf{y} + 2\mathbf{X}^\top \mathbf{X}\boldsymbol{\theta} = 0$$
$$\mathbf{X}^\top \mathbf{X}\boldsymbol{\theta} = \mathbf{X}^\top \mathbf{y}$$

Ces équations sont les **équations normales**. Elles expriment que le résidu $\mathbf{r} = \mathbf{y} - \mathbf{X}\boldsymbol{\theta}$ est **orthogonal** à l'espace colonnes de $\mathbf{X}$:

$$\mathbf{X}^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) = \mathbf{0}$$

Solution des moindres carrés ordinaires (MCO)

Si $\mathbf{X}^\top \mathbf{X}$ est inversible (rang plein), la solution est :

$$\boxed{\hat{\boldsymbol{\theta}}_{\text{MCO}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}}$$

Interprétation géométrique :

- $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\theta}}$ est la **projection orthogonale** de $\mathbf{y}$ sur l'espace colonnes de $\mathbf{X}$
- Le résidu $\mathbf{r} = \mathbf{y} - \hat{\mathbf{y}}$ est perpendiculaire à cet espace
- La **matrice chapeau** $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top$ projette : $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$

Complexité : $O(Nd^2 + d^3)$, dominée par l'inversion de $\mathbf{X}^\top \mathbf{X}$

Décomposition en valeurs singulières (DVS)

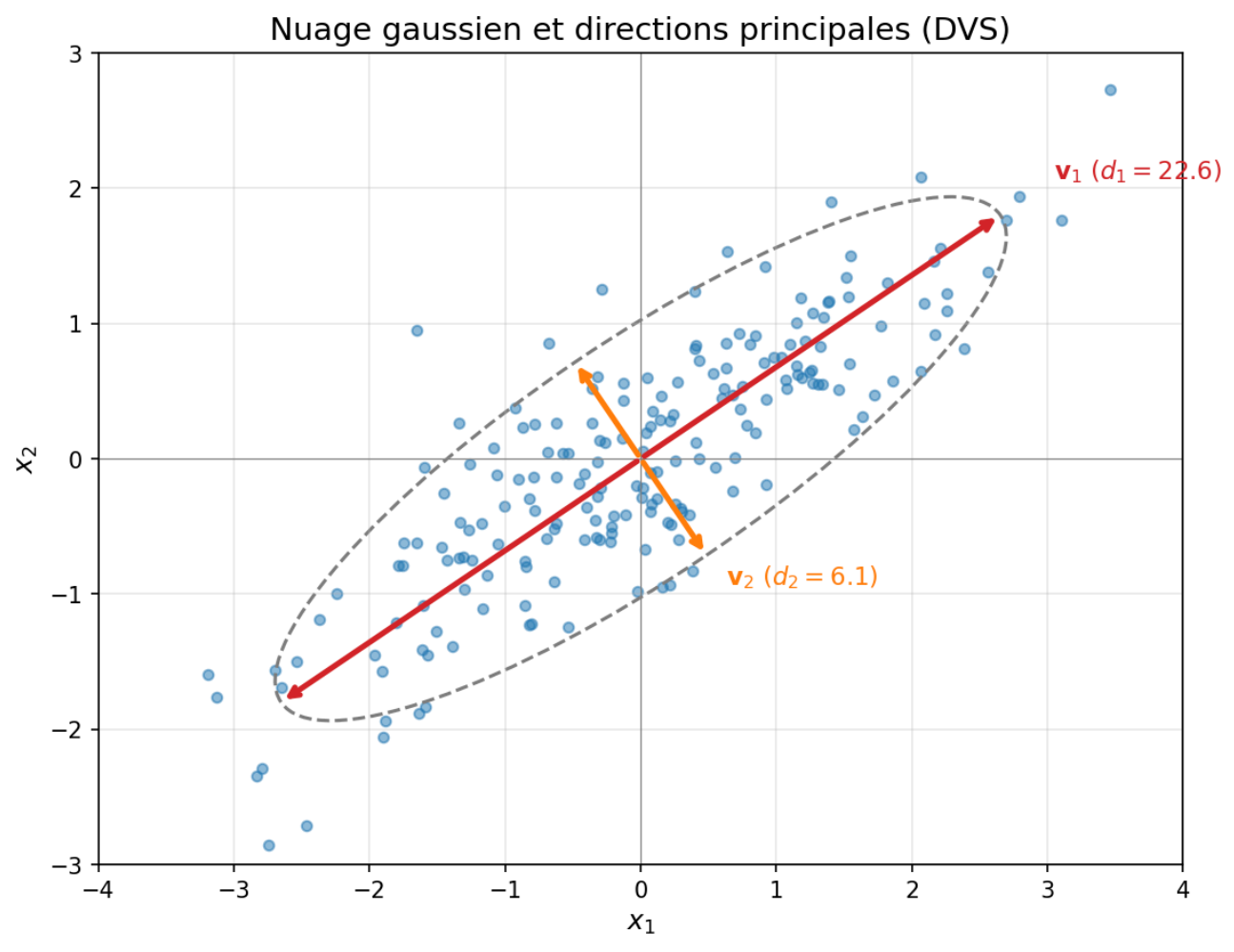
Toute matrice $\mathbf{X} \in \mathbb{R}^{N \times d}$ admet une décomposition :

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$$

Matrice	Dimension	Propriétés	Interprétation
$\mathbf{U}$	$N \times d$	Colonnes orthonormales	Directions dans l'espace des observations
$\mathbf{D}$	$d \times d$	Diagonale, $d_1 \geq \cdots \geq d_d \geq 0$	Valeurs singulières (amplitudes)
$\mathbf{V}$	$d \times d$	Orthogonale	Directions principales (espace des caractéristiques)

Lien avec les valeurs propres : Les colonnes de $\mathbf{V}$ sont les vecteurs propres de $\mathbf{X}^\top\mathbf{X}$, et d_j^2 sont les valeurs propres correspondantes.

Géométrie de la DVS : ellipse des données



Les **vecteurs singuliers** $\mathbf{v}_j$ sont les axes naturels du nuage de données. Les **valeurs singulières** d_j mesurent la **dispersion des données** le long de chaque axe.

Deux variances : ne pas confondre!

Le mot « variance » a **deux sens distincts** :

Type	Définition	Proportionnelle à	Quand d_j est grand
Variance des données	Dispersion le long de $\mathbf{v}_j$	d_j^2	Élevée
Variance d'estimation	Incertitude sur $\hat{\theta}_j$	$1/d_j^2$	Faible

Ces deux variances sont inversement reliées!

- Grande dispersion des données → beaucoup d'information → estimé précis
- Petite dispersion des données → peu d'information → estimé incertain

Ridge réduit la **variance d'estimation** (incertitude sur $\hat{\boldsymbol{\theta}}$) en rétrécissant les directions où d_j est petit, c'est-à-dire là où les données manquent de dispersion.

Interprétation géométrique de la DVS

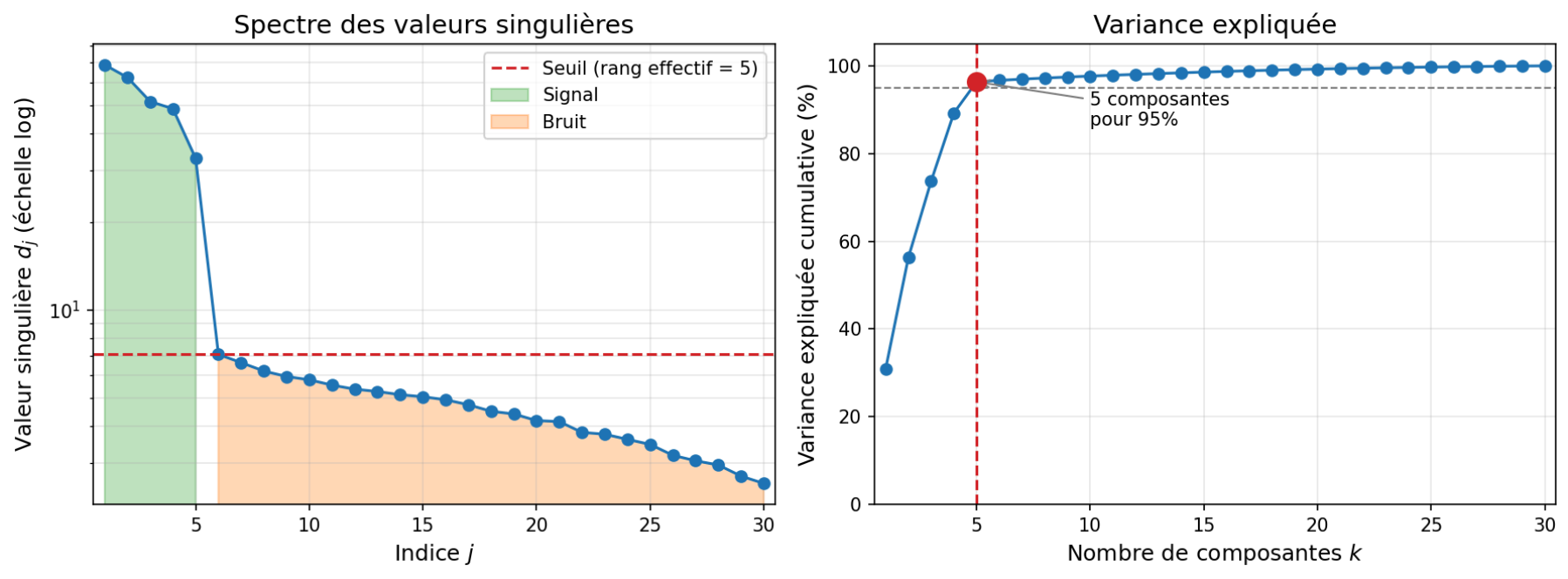
La DVS décompose la transformation $\mathbf{X}$ en trois étapes :

1. $\mathbf{V}^\top$: Rotation dans l'espace des caractéristiques (vers les axes principaux)
2. $\mathbf{D}$: Étirement ou compression le long de chaque axe (par d_j)
3. $\mathbf{U}$: Rotation dans l'espace des observations

Valeur singulière	Signal	Variance des données	Variance d'estimation
d_j grand	Fort	Élevée (données dispersées)	Faible (estimé précis)
d_j petit	Faible	Faible (peu de dispersion)	Élevée (estimé incertain)

Conditionnement : Le ratio $\kappa = d_1/d_d$ mesure la difficulté numérique. Si κ est grand, $\mathbf{X}^\top\mathbf{X}$ est mal conditionnée.

Spectre des valeurs singulières et rang effectif



Le **rang effectif** : nombre de valeurs singulières significatives (au-dessus du bruit).

Solution MCO via DVS

En utilisant $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$, la solution MCO devient :

$$\hat{\boldsymbol{\theta}}_{\text{MCO}} = \mathbf{V}\mathbf{D}^{-1}\mathbf{U}^\top\mathbf{y} = \sum_{j=1}^{d} \frac{\mathbf{u}_j^\top \mathbf{y}}{d_j}\mathbf{v}_j$$

Décomposition terme par terme :

- $\mathbf{u}_j^\top \mathbf{y}$: Projection de $\mathbf{y}$ sur la j-ème direction
- $1/d_j$: Normalisation par l'amplitude de cette direction
- $\mathbf{v}_j$: Direction correspondante dans l'espace des paramètres

Problème : Si $d_j \approx 0$, nous divisons par un petit nombre, ce qui cause une **amplification du bruit**.

Instabilité numérique de MCO

$$\hat{\boldsymbol{\theta}}_{\text{MCO}} = \sum_{j=1}^{d} \frac{\mathbf{u}_j^\top \mathbf{y}}{d_j} \mathbf{v}_j$$

Situation	Conséquence
d_j petit	Division par petit nombre, coefficients énormes
Caractéristiques corrélées	Valeurs singulières proches de 0
$d \approx N$	Matrice $\mathbf{X}^\top \mathbf{X}$ proche de singulière
$d > N$	Infinité de solutions (système sous-déterminé)

Exemple : Si $d_j = 0{,}001$ et $\mathbf{u}_j^\top \mathbf{y} = 0{,}1$, la contribution est $100 \cdot \mathbf{v}_j$. Le bruit est amplifié 1000 fois.

Solution : Régularisation (Ridge), qui pénalise les directions à faible signal.

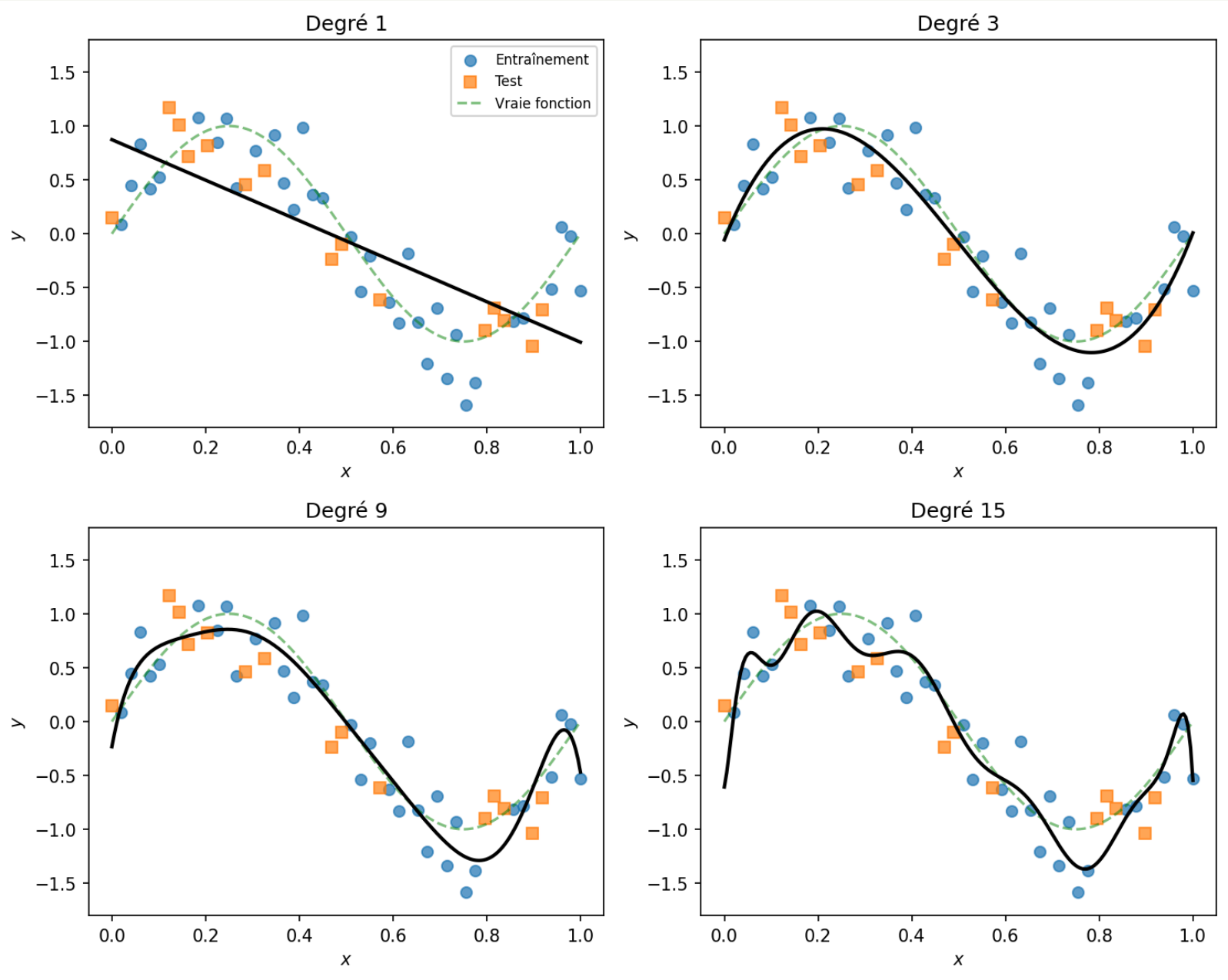
L'écart de généralisation

$$\text{Écart} = \underbrace{\mathcal{R}(f)}_{\text{Erreur sur nouvelles données}} - \underbrace{\hat{\mathcal{R}}(f; \mathcal{D}_{\text{train}})}_{\text{Erreur d'entraînement}}$$

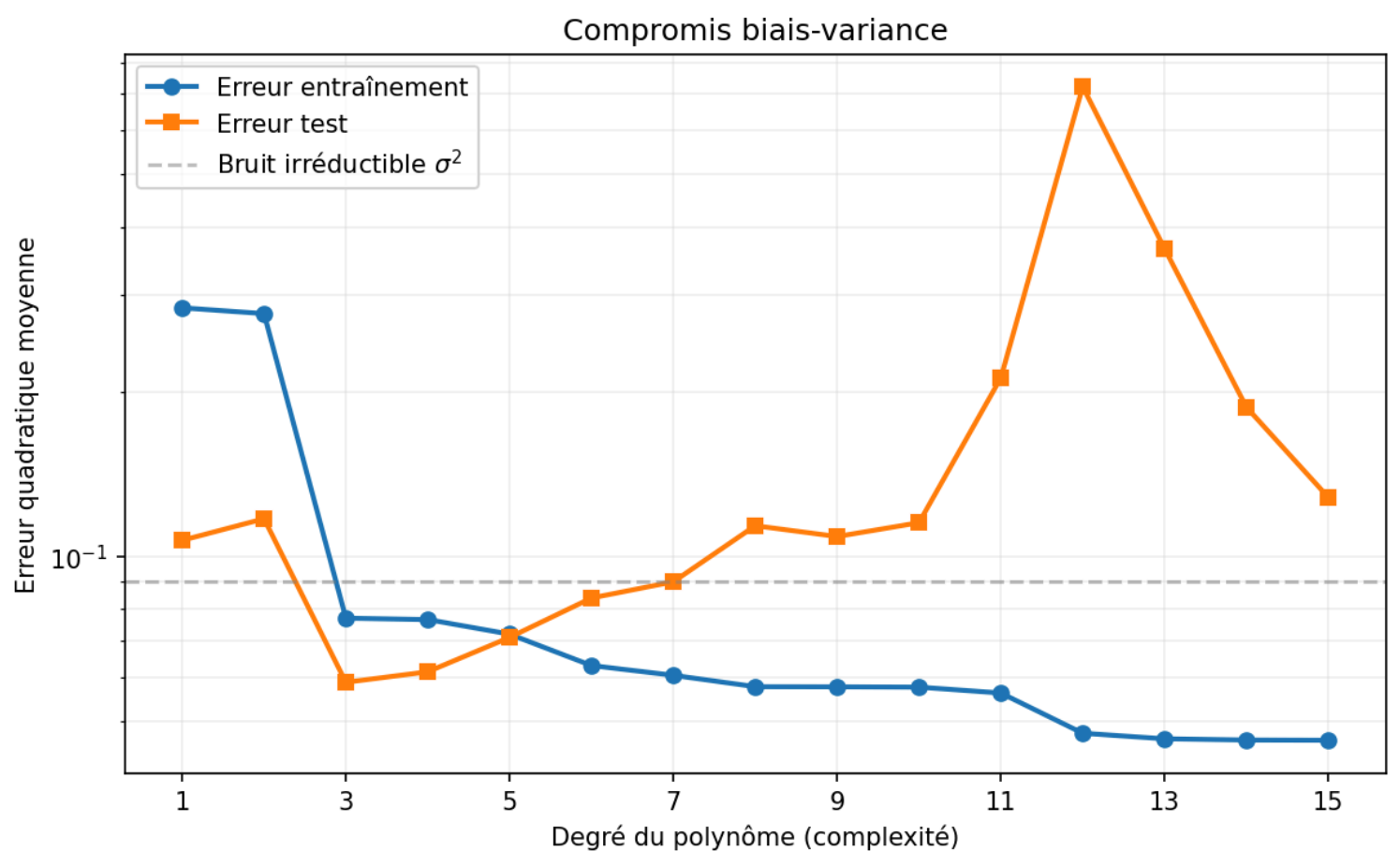
Diagnostic	Erreur entraînement	Erreur test	Problème
Sous-apprentissage	Élevée	Élevée	Modèle trop simple
Bon ajustement	Faible	Faible	Correct
Surapprentissage	**Très faible**	**Élevée**	**Modèle mémorise le bruit**

Le surapprentissage survient quand le modèle s'ajuste aux **particularités** de l'échantillon (y compris le bruit) plutôt qu'aux **régularités** sous-jacentes.

Illustration : régression polynomiale



Courbes d'erreur : le compromis en action



L'erreur d'entraînement ↓ avec la complexité. L'erreur de test forme un **U** : elle diminue puis augmente.

Décomposition biais-variance : dérivation

Rappel : le modèle génératif est $y = f^*(\mathbf{x}) + \epsilon$ avec $\mathbb{E}[\epsilon] = 0$, $\mathrm{Var}(\epsilon) = \sigma^2$.

Variables aléatoires :

- $f^*(\mathbf{x})$: **fixe** (vraie fonction, déterministe)
- ϵ : **aléatoire** (bruit d'observation)
- $\hat{f}(\mathbf{x})$: **aléatoire** via $\mathcal{D}$ (différents échantillons → différentes fonctions apprises → différentes prédictions)

L'espérance $\mathbb{E}[\hat{f}(\mathbf{x})]$ moyenne sur tous les $\mathcal{D}$ possibles. Décomposons l'erreur :

$$\mathbb{E}_{\mathcal{D},\epsilon}[(\hat{f}(\mathbf{x}) - y)^2] = \mathbb{E}[(\hat{f}(\mathbf{x}) - f^*(\mathbf{x}) - \epsilon)^2]$$

En développant et utilisant $\mathbb{E}[\epsilon] = 0$ et l'indépendance de ϵ et $\hat{f}$:

$$= \mathbb{E}[(\hat{f}(\mathbf{x}) - f^*(\mathbf{x}))^2] + \sigma^2$$

Décomposition biais-variance : suite

Ajoutons et retranchons $\mathbb{E}[\hat{f}(\mathbf{x})]$ dans le premier terme :

$$\mathbb{E}[(\hat{f} - f^*)^2] = \mathbb{E}[(\hat{f} - \mathbb{E}[\hat{f}] + \mathbb{E}[\hat{f}] - f^*)^2]$$

$$= \underbrace{\mathbb{E}[(\hat{f} - \mathbb{E}[\hat{f}])^2]}_{\mathrm{Var}(\hat{f})} + \underbrace{(\mathbb{E}[\hat{f}] - f^*)^2}_{\mathrm{Biais}^2} + 2\underbrace{\mathbb{E}[(\hat{f} - \mathbb{E}[\hat{f}])](\mathbb{E}[\hat{f}] - f^*)}_{=0}$$

$$\boxed{\mathrm{Erreur} = \mathrm{Biais}^2(\hat{f}) + \mathrm{Var}(\hat{f}) + \sigma^2}$$

Terme	Signification	Dépend de
Biais²	Écart systématique à f^*	Classe $\mathcal{H}$ (trop restrictive?)
Variance	Sensibilité à l'échantillon	Complexité et taille N
σ^2	Bruit irréductible	Données uniquement

Le compromis biais-variance

Complexité du modèle	Biais	Variance	Erreur totale
Trop simple	↑↑	↓	Élevée (sous-apprentissage)
Optimale	↓	↓	**Minimale**
Trop complexe	↓	↑↑	Élevée (surapprentissage)

Lien avec le prédicteur de Bayes : Si $f^* = \mathbb{E}[y|\mathbf{x}]$, le biais mesure à quel point notre classe $\mathcal{H}$ peut approcher cette fonction.

Lien avec la DVS : Les directions à petites valeurs singulières (faible variance des données) ont une grande **variance d'estimation**, amplifiant le bruit. Ridge cible précisément ces directions.

Besoin de régularisation

Problèmes avec MCO sans régularisation :

1. $\mathbf{X}^\top \mathbf{X}$ peut être **singulière** ou **mal conditionnée**
2. Coefficients **instables** quand les caractéristiques sont corrélées
3. **Surapprentissage** avec beaucoup de caractéristiques (d grand)

Solution, la régularisation : Pénaliser la « complexité » du modèle

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \left[\underbrace{\hat{\mathcal{R}}(\boldsymbol{\theta})}_{\text{Ajustement aux données}} + \underbrace{\lambda \cdot C(\boldsymbol{\theta})}_{\text{Pénalité de complexité}} \right]$$

Le paramètre $\lambda > 0$ contrôle le compromis biais-variance.

Objectif Ridge (régularisation L2)

Ajouter une pénalité sur la **norme L2** des paramètres :

$$\hat{\boldsymbol{\theta}}_{\text{ridge}} = \arg\min_{\boldsymbol{\theta}} \left[\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2 + \lambda \|\boldsymbol{\theta}\|_2^2 \right]$$

λ	**Effet**	**Résultat**
$\lambda = 0$	Pas de pénalité	Solution MCO
λ petit	Légère régularisation	Réduction de la variance
λ grand	Forte régularisation	Coefficients tendent vers 0
$\lambda \to \infty$	Pénalité dominante	$\boldsymbol{\theta} \to \mathbf{0}$

Interprétation : Nous cherchons un compromis entre bien ajuster les données et garder des coefficients raisonnables.

Solution analytique Ridge

Gradient de l'objectif :

$$\nabla_{\boldsymbol{\theta}} \left[\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|^2 + \lambda \|\boldsymbol{\theta}\|^2 \right] = -2\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) + 2\lambda\boldsymbol{\theta}$$

Condition d'optimalité (gradient nul) :

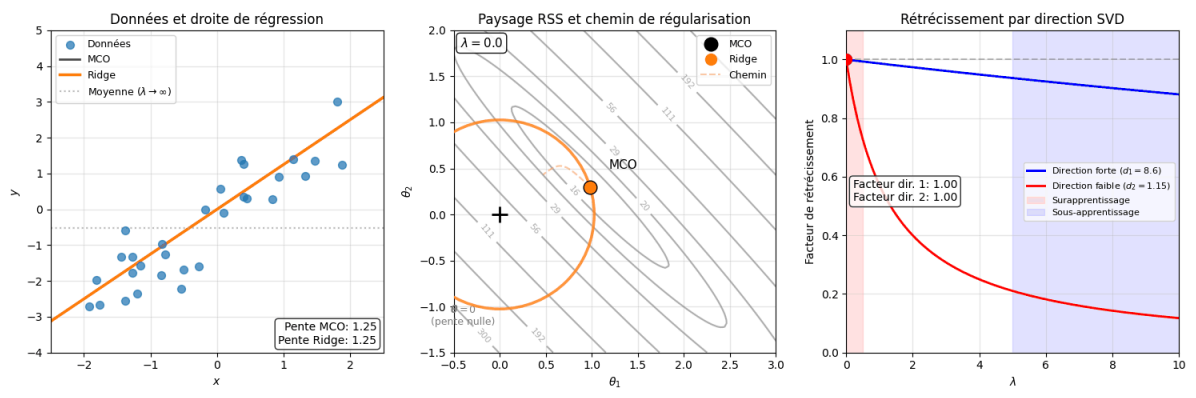
$$\mathbf{X}^\top\mathbf{X}\boldsymbol{\theta} + \lambda\boldsymbol{\theta} = \mathbf{X}^\top\mathbf{y}$$

$$(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})\boldsymbol{\theta} = \mathbf{X}^\top\mathbf{y}$$

$$\boxed{\hat{\boldsymbol{\theta}}_{\text{ridge}} = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y}}$$

L'ajout de $\lambda\mathbf{I}$ **garantit l'inversibilité** et améliore le conditionnement.

Géométrie de Ridge : formulation contrainte



La figure montre la formulation **contrainte** équivalente :

$$\min_{\boldsymbol{\theta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|^2 \quad \text{sous contrainte} \quad \|\boldsymbol{\theta}\|^2 \leq t$$

- **Ellipses** : Lignes de niveau de la SCR
- **Cercle** : Région admissible $\|\boldsymbol{\theta}\|^2 \leq t$
- **Solution** : Plus petite SCR compatible avec la contrainte

La formulation pénalisée $\text{SCR} + \lambda\|\boldsymbol{\theta}\|^2$ est le lagrangien; chaque λ correspond à un t.

Solution Ridge via DVS

Avec $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$, la solution Ridge devient :

$$\hat{\boldsymbol{\theta}}_{\text{ridge}} = \sum_{j=1}^{d} \underbrace{\frac{d_j^2}{d_j^2 + \lambda}}_{\text{facteur} \in [0,1]} \cdot \frac{\mathbf{u}_j^\top \mathbf{y}}{d_j} \cdot \mathbf{v}_j$$

Facteur de rétrécissement $\frac{d_j^2}{d_j^2+\lambda}$:

Valeur singulière	Facteur	Effet
d_j grand (signal fort)	≈ 1	Peu de rétrécissement
d_j petit (bruit)	≈ 0	Fort rétrécissement
$d_j = \sqrt{\lambda}$	$0{,}5$	Rétrécissement moyen

De Ridge à l'ACP : deux philosophies

Approche	Traitement des directions bruitées	Type
Ridge	Rétrécit (seuillage doux)	Continue : garde tout, pénalise
ACP	Élimine (seuillage dur)	Discrète : garde k, ignore le reste

Analyse en composantes principales (ACP) : Garder seulement les k premières directions :

$$\mathbf{z}_n = \mathbf{V}_k^\top (\mathbf{x}_n - \bar{\mathbf{x}}) \in \mathbb{R}^k$$

Ridge est appropriée pour la régression supervisée. L'ACP est appropriée pour la réduction de dimension non supervisée.

L'approche bayésienne

Plutôt que de choisir une perte arbitraire, **modélisons** explicitement la génération des données :

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{p(\boldsymbol{\theta}) \cdot p(\mathcal{D}|\boldsymbol{\theta})}{p(\mathcal{D})}$$

Terme	Nom	Signification
$p(\boldsymbol{\theta})$	**A priori**	Croyances avant de voir les données
$p(\mathcal{D}\|\boldsymbol{\theta})$	**Vraisemblance**	Probabilité des données pour un $\boldsymbol{\theta}$
$p(\boldsymbol{\theta}\|\mathcal{D})$	**A posteriori**	Croyances mises à jour
$p(\mathcal{D})$	**Évidence**	Constante de normalisation

L'a posteriori combine notre connaissance préalable avec l'information des données.

Distribution prédictive bayésienne

L'approche **complètement bayésienne** moyenne sur tous les paramètres possibles :

$$p(y|\mathbf{x}, \mathcal{D}) = \int p(y|\mathbf{x}, \boldsymbol{\theta}) \cdot p(\boldsymbol{\theta}|\mathcal{D})\, d\boldsymbol{\theta}$$

Cette **distribution prédictive a posteriori** :

- Intègre l'**incertitude sur les paramètres** dans la prédiction
- Ne s'engage pas sur une valeur unique de $\boldsymbol{\theta}$
- Donne des **intervalles de confiance** naturels

Avantages : Quantification de l'incertitude, robustesse, pas de surapprentissage.

Problème : l'intégrale est intraitable

$$p(y|\mathbf{x}, \mathcal{D}) = \int p(y|\mathbf{x}, \boldsymbol{\theta}) \cdot p(\boldsymbol{\theta}|\mathcal{D})\, d\boldsymbol{\theta}$$

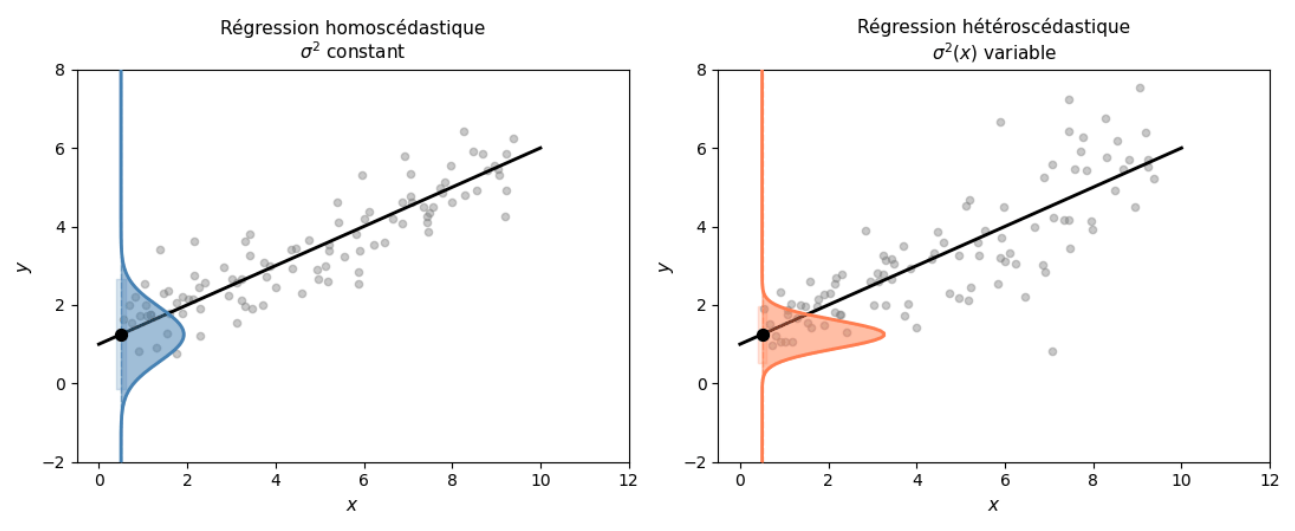
En haute dimension, cette intégrale est **impossible à calculer** analytiquement :

- Espace des paramètres de dimension d, intégration sur $\mathbb{R}^d$
- Pas de forme fermée en général

Solution pragmatique : Utiliser des **estimateurs ponctuels** :

Estimateur	Principe
EMV (maximum de vraisemblance)	$\hat{\boldsymbol{\theta}} = \arg\max p(\mathcal{D}\|\boldsymbol{\theta})$
MAP (maximum a posteriori)	$\hat{\boldsymbol{\theta}} = \arg\max p(\boldsymbol{\theta}\|\mathcal{D})$

Modèle probabiliste : bruit gaussien



Nous modélisons : $y = \boldsymbol{\theta}^\top \mathbf{x} + \epsilon$, où $\epsilon \sim \mathcal{N}(0, \sigma^2)$

Chaque observation y suit une gaussienne centrée sur la droite de régression.

Maximum de vraisemblance (EMV)

L'EMV trouve les paramètres qui **maximisent la probabilité d'observer les données** :

$$\hat{\boldsymbol{\theta}}_{\text{EMV}} = \arg\max_{\boldsymbol{\theta}} \prod_{i=1}^{N} p(y_i|\mathbf{x}_i;\boldsymbol{\theta})$$

En passant au log (transforme le produit en somme) :

$$\hat{\boldsymbol{\theta}}_{\text{EMV}} = \arg\min_{\boldsymbol{\theta}} \underbrace{-\sum_{i=1}^{N} \log p(y_i|\mathbf{x}_i;\boldsymbol{\theta})}_{\text{Log-vraisemblance négative}}$$

Sous bruit gaussien $p(y|\mathbf{x},\boldsymbol{\theta}) = \mathcal{N}(y|\boldsymbol{\theta}^\top\mathbf{x},\sigma^2)$:

$$\text{LVN} = \frac{N}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}\sum_i (y_i - \boldsymbol{\theta}^\top\mathbf{x}_i)^2 \propto \text{SCR}$$

EMV = MCO sous l'hypothèse de bruit gaussien.

Maximum a posteriori (MAP)

Le MAP trouve le **mode** de la distribution a posteriori :

$$\hat{\boldsymbol{\theta}}_{\text{MAP}} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\mathcal{D}) = \arg\max_{\boldsymbol{\theta}} p(\mathcal{D}|\boldsymbol{\theta}) \cdot p(\boldsymbol{\theta})$$

En passant au log :

$$\hat{\boldsymbol{\theta}}_{\text{MAP}} = \arg\min_{\boldsymbol{\theta}} \left[\underbrace{-\log p(\mathcal{D}|\boldsymbol{\theta})}_{\text{LVN (ajustement)}} + \underbrace{(-\log p(\boldsymbol{\theta}))}_{\text{Pénalité (a priori)}} \right]$$

L'a priori devient naturellement un **terme de régularisation**.

- A priori uniforme : $-\log p(\boldsymbol{\theta}) = \text{cst}$, donc MAP = EMV
- A priori gaussien : $-\log p(\boldsymbol{\theta}) \propto \|\boldsymbol{\theta}\|^2$, donc MAP = Ridge

A priori gaussien et régularisation L2

Supposons un a priori gaussien centré (isotrope) :

$$p(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}|\mathbf{0}, \tau^2\mathbf{I}) = \frac{1}{(2\pi\tau^2)^{d/2}}\exp\left(-\frac{\|\boldsymbol{\theta}\|^2}{2\tau^2}\right)$$

Le log de l'a priori est :

$$-\log p(\boldsymbol{\theta}) = \frac{d}{2}\log(2\pi\tau^2) + \frac{1}{2\tau^2}\|\boldsymbol{\theta}\|^2$$

La partie qui dépend de $\boldsymbol{\theta}$ est $\frac{1}{2\tau^2}\|\boldsymbol{\theta}\|^2$, une **pénalité L2**.

Interprétation : L'a priori gaussien encode notre croyance que les coefficients sont probablement « petits » (proches de 0).

MAP avec a priori gaussien = Ridge

Avec vraisemblance gaussienne (σ^2) et a priori gaussien (τ^2) :

$$\hat{\boldsymbol{\theta}}_{\text{MAP}} = \arg\min_{\boldsymbol{\theta}} \left[\frac{1}{2\sigma^2} \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|^2 + \frac{1}{2\tau^2} \|\boldsymbol{\theta}\|^2 \right]$$

En comparant avec Ridge $\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|^2 + \lambda \|\boldsymbol{\theta}\|^2$:

$$\boxed{\lambda = \frac{\sigma^2}{\tau^2}}$$

Paramètre	Interprétation	Effet sur λ
τ^2 grand	A priori large	λ petit, peu de régularisation
τ^2 petit	A priori concentré	λ grand, forte régularisation

Synthèse : deux langages, mêmes algorithmes

Perspective décisionnelle	Perspective probabiliste
Perte quadratique $(y - \hat{y})^2$	Bruit gaussien $\mathcal{N}(y\|\boldsymbol{\theta}^\top \mathbf{x}, \sigma^2)$
Minimiser SCR	Maximum de vraisemblance (EMV)
Solution MCO	**Solution MCO**
+ Régularisation L2 $\lambda\|\boldsymbol{\theta}\|^2$	+ A priori gaussien $\mathcal{N}(\mathbf{0}, \tau^2\mathbf{I})$
Régression Ridge	**Maximum a posteriori (MAP)**

Les deux perspectives sont **équivalentes** mathématiquement, mais offrent des éclairages complémentaires :

- **Décisionnelle** : Comment construire l'algorithme
- **Probabiliste** : Pourquoi ces choix sont raisonnables

Résumé

1. **MRE** : Minimiser l'erreur d'entraînement comme approximation du vrai risque
2. **Bayes optimal** : Pour la perte L2, le prédicteur optimal est $\mathbb{E}[y|\mathbf{x}]$
3. **MCO** : Solution analytique $(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$
4. **DVS** : Révèle l'instabilité (petites valeurs singulières amplifient le bruit)
5. **Généralisation** : Compromis biais-variance, surapprentissage
6. **Ridge** : $(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y}$
7. **Probabiliste** : EMV = MCO sous bruit gaussien
8. **MAP** : A priori gaussien + EMV = Ridge, avec $\lambda = \sigma^2/\tau^2$

Questions?

Exercices recommandés :

- Exercice 14 : Prédicteur de Bayes optimal
- Exercice 11 : MAP et régression Ridge
- Exercice 16 : DVS et facteurs de rétrécissement