

FTML 2025 — Exercice 3

Espérance du risque empirique en régression linéaire

1 Objectif

Ce troisième exercice du projet FTML a pour but d'étudier le comportement de l'espérance du risque empirique dans un cadre de régression linéaire, et de le comparer au *vrai risque* en fonction de la taille de l'échantillon n . Nous mettons en place une simulation contrôlée du modèle, puis analysons les deux risques sur un grand nombre d'expériences répétées.

2 Modèle génératif

Les données sont simulées selon le modèle gaussien linéaire suivant :

$$X \in \mathbb{R}^{n \times d} \sim \mathcal{N}(0, I_d), \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I_n), \quad Y = X\theta + \varepsilon$$

où :

- $d = 10$ est la dimension des variables d'entrée ;
- $\theta \in \mathbb{R}^d$ est un vecteur fixe aléatoire, généré une fois pour toute ;
- $\sigma^2 = 1.0$ est la variance du bruit additif.

3 Estimateur OLS

L'estimateur des moindres carrés ordinaires est défini par :

$$\hat{\theta} = (X^T X)^{-1} X^T Y$$

Lorsque X est mal conditionné (cas possible pour $n < d$), on utilise la pseudo-inverse de Moore–Penrose :

$$\hat{\theta} = \text{pinv}(X) \cdot Y$$

4 Définitions des risques

4.1 Risque empirique

Le risque empirique est défini comme l'erreur quadratique moyenne observée sur les données bruitées :

$$R_{\text{emp}} = \frac{1}{n} \sum_{i=1}^n (Y_i - X_i \hat{\theta})^2$$

Il dépend à la fois du bruit aléatoire et de l'approximation de $\hat{\theta}$.

4.2 Vrai risque

Le vrai risque mesure l'écart au prédicteur optimal $f^*(X) = X\theta$, sans bruit :

$$R_{\text{true}} = \frac{1}{n} \sum_{i=1}^n (X_i \hat{\theta} - X_i \theta)^2 = \frac{1}{n} \|X(\hat{\theta} - \theta)\|^2$$

5 Méthodologie expérimentale

Pour chaque valeur de n (de 10 à 500), nous :

1. Générons X , Y avec bruit, et $Y^* = X\theta$ (vrai signal) ;
2. Calculons $\hat{\theta}$ par OLS ;
3. Évaluons R_{emp} et R_{true} ;
4. Répétons $T = 50$ fois (tirages aléatoires indépendants) ;
5. Moyennons les deux risques sur les 50 expériences.

L'ensemble est accéléré via NumPy et Numba.

6 Résultats

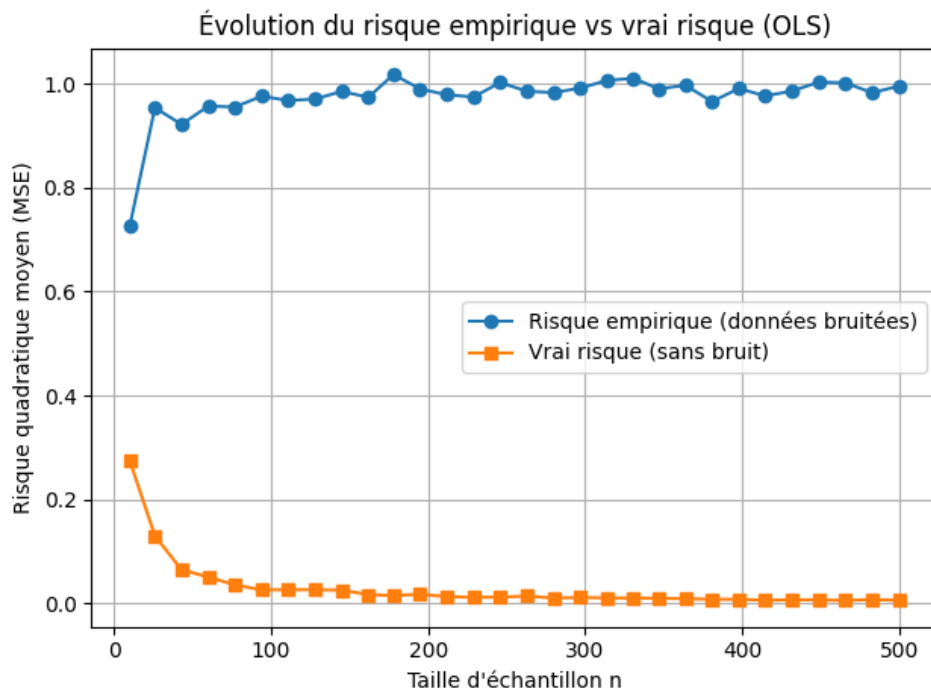


Figure 1: Évolution du risque empirique (en bleu) et du vrai risque (en orange) en fonction de la taille n

Quelques résultats numériques :

n	Risque empirique	Vrai risque
10	0.7258	0.2745
128	0.9700	0.0264
314	1.0058	0.0106
500	0.9947	0.0061

7 Analyse théorique

7.1 Pourquoi le risque empirique tend vers σ^2 ?

Par définition,

$$\mathbb{E}[R_{\text{emp}}] = \mathbb{E}[\|Y - X\hat{\theta}\|^2/n] = \text{Biais} + \text{Variance} + \sigma^2$$

Mais quand le modèle est bien spécifié, et n grand :

$$\mathbb{E}[R_{\text{emp}}] \rightarrow \sigma^2$$

car l'erreur d'approximation disparaît et seul le bruit reste. Cela montre que le risque empirique est un ****estimateur biaisé optimiste**** du vrai risque.

7.2 Pourquoi le vrai risque diminue ?

Le vrai risque mesure l'écart de $\hat{\theta}$ à θ . Par consistance de l'estimateur OLS :

$$\hat{\theta} \xrightarrow{p} \theta \quad \text{quand } n \rightarrow \infty$$

Ainsi, on a :

$$R_{\text{true}} = \frac{1}{n} \|X(\hat{\theta} - \theta)\|^2 \rightarrow 0$$

8 Discussion

- Le **risque empirique** reste borné autour de 1.0 pour tout n , ce qui reflète la variance du bruit.
- Le **vrai risque** diminue fortement, ce qui montre la convergence de l'estimateur vers le modèle génératif.
- L'écart entre les deux s'explique par la présence du bruit dans les données d'entraînement.

Ce phénomène illustre l'importance de valider les modèles sur un jeu de test hors bruit.

9 Conclusion

Cet exercice met en lumière :

- la distinction fondamentale entre risque empirique et vrai risque ;
- la convergence théorique et empirique de $\hat{\theta}$ vers θ avec n ;
- l'intérêt de la moyenne sur expériences pour évaluer l'espérance du risque.

Le cadre de simulation maîtrisé permet de valider rigoureusement les résultats attendus de la théorie du risque en apprentissage supervisé.