

Cours Magistral : Fondements de l'Apprentissage Statistique

Synthèse de Théorie de l'Estimation et de l'Approximation

Mars 2026

1 Cadre de l'Inférence Statistique

Soit $(\mathcal{X}, \mathcal{A})$ un espace mesurable. On considère un échantillon $\mathcal{D}_n = \{X_1, \dots, X_n\}$ de variables aléatoires (v.a.) indépendantes et identiquement distribuées (i.i.d.) selon une loi de probabilité inconnue P .

Définition 1.1 (Modèle Statistique). *Un modèle statistique est un triplet $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ où $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ est une famille de lois de probabilité sur l'espace des observations.*

- Si $\Theta \subseteq \mathbb{R}^d$ avec $d < \infty$, le modèle est dit **paramétrique**.
- Si Θ est de dimension infinie (ex : espace de fonctions), le modèle est **non-paramétrique**.

2 Estimation Paramétrique

On suppose ici que $P = P_\theta$ pour un certain $\theta \in \Theta \subseteq \mathbb{R}^d$.

Définition : Estimateur du Maximum de Vraisemblance (EMV)

Soit $L_n(\theta; X_1, \dots, X_n) = \prod_{i=1}^n p_\theta(X_i)$ la fonction de vraisemblance. L'estimateur du maximum de vraisemblance $\hat{\theta}_n$ est défini par :

$$\hat{\theta}_n \in \operatorname{argmax}_{\theta \in \Theta} \ell_n(\theta) \quad \text{où} \quad \ell_n(\theta) = \sum_{i=1}^n \log p_\theta(X_i)$$

Proposition 2.1 (Méthode des Moments). *Soit $m_k(\theta) = \mathbb{E}_\theta[X^k]$ le moment théorique d'ordre k . On définit le moment empirique par $\hat{m}_k = \frac{1}{n} \sum_{i=1}^n X_i^k$. L'estimateur des moments $\hat{\theta}_M$ est solution du système :*

$$\forall k \in \{1, \dots, d\}, \quad m_k(\hat{\theta}_M) = \hat{m}_k$$

3 Estimation de Densité Non-Paramétrique

On cherche à estimer une densité $f \in \mathcal{F}$ à partir de $X_1, \dots, X_n \sim f$.

3.1 Approche par Projection

Soit $\{e_k\}_{k \in \mathbb{Z}}$ une base orthonormée de $L^2([0, 1])$. On a $f = \sum_{k \in \mathbb{Z}} \alpha_k e_k$.

Définition : Estimateur par Projection

L'estimateur par projection tronqué à l'ordre M est :

$$\hat{f}_M(x) = \sum_{|k| \leq M} \hat{\alpha}_k e_k(x) \quad \text{où} \quad \hat{\alpha}_k = \frac{1}{n} \sum_{i=1}^n e_k(X_i)$$

Théorème 3.1 (Vitesse de convergence sur l'ellipsoïde de Sobolev). Soit $B(s, R) = \{f \in L^2([0, 1]) : \sum_k |\alpha_k|^2 (1 + |k|)^{2s} \leq R^2\}$. Pour $f \in B(s, R)$, le risque quadratique (MISE) vérifie :

$$\mathbb{E} \|\hat{f}_M - f\|^2 \leq \underbrace{\frac{R^2}{M^{2s}}}_{\text{Biais}^2} + \underbrace{\frac{2M+1}{n}}_{\text{Variance}}$$

Le choix optimal $M \asymp n^{\frac{1}{2s+1}}$ conduit à une vitesse de convergence de $n^{-\frac{2s}{2s+1}}$.

3.2 Approche par Noyau

Définition : Estimateur de Parzen-Rosenblatt

Soit $K : \mathbb{R} \rightarrow \mathbb{R}$ une fonction d'intégrale 1 (noyau) et $h > 0$ la fenêtre. L'estimateur à noyau est :

$$\hat{f}_{n,h}(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right)$$

Théorème 3.2 (Convergence en norme Hölderienne). Soit $f \in \Lambda(s, L)$ (Espace de Hölder) et K un noyau d'ordre $k = \lfloor s \rfloor$. Alors :

$$\sup_{f \in \Lambda(s, L)} \mathbb{E}[|\hat{f}_{n,h}(x) - f(x)|^2] \leq C \left(h^{2s} + \frac{1}{nh} \right)$$

L'équilibre est atteint pour $h^* \asymp n^{-\frac{1}{2s+1}}$.

4 Théorie de la Régression

On observe $(X_i, Y_i) \in \mathcal{X} \times \mathbb{R}$ i.i.d. On cherche à minimiser le risque $R(f) = \mathbb{E}[(Y - f(X))^2]$.

Proposition 4.1 (Caractérisation de la solution Bayes). La fonction minimisant le risque quadratique sur l'ensemble des fonctions mesurables est la fonction de régression :

$$m(x) = \mathbb{E}[Y|X = x]$$

4.1 Estimateur de Nadaraya-Watson

Il s'agit d'une version locale de la moyenne pondérée par un noyau K .

Théorème : Estimateur de Nadaraya-Watson

L'estimateur de la fonction de régression $m(x)$ est donné par :

$$\hat{m}_n(x) = \frac{\sum_{i=1}^n Y_i K\left(\frac{x - X_i}{h}\right)}{\sum_{j=1}^n K\left(\frac{x - X_j}{h}\right)}$$

4.2 Régularisation et Splines

Pour éviter l'overfitting dans un espace \mathcal{F} de dimension infinie, on minimise le risque empirique pénalisé.

Définition 4.2 (Splines de lissage). La spline de lissage cubique est la solution de :

$$\hat{f}_\lambda = \underset{f \in C^2([a, b])}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^n (Y_i - f(X_i))^2 + \lambda \int_a^b |f''(t)|^2 dt \right\}$$

La solution est une spline cubique naturelle dont les nœuds sont les X_i .

5 Limites de l'Approximation Linéaire

On s'intéresse à la classe $\mathcal{F}_C = \{f \mid \int_{\mathbb{R}^d} \|\omega\|_1 |F(\omega)| d\omega \leq C\}$, où F est la transformée de Fourier de f .

Théorème : Fléau de la dimension (Lower Bound)

Soit $w_N(\mathcal{F}_C)$ l'écart de Kolmogorov de dimension N . Il existe $\kappa > 0$ tel que :

$$\forall N \in \mathbb{N}^*, \forall d \in \mathbb{N}^*, \quad w_N(\mathcal{F}_C) \geq \kappa \frac{C}{d} N^{-1/d}$$

Remarque 5.1. *Ce résultat montre que pour les méthodes linéaires (polynômes, séries trigonométriques), l'erreur d'approximation se dégrade exponentiellement avec la dimension d . C'est ce qui justifie l'usage de modèles non-linéaires comme les **réseaux de neurones**, qui permettent de briser ce fléau sous certaines conditions de régularité.*