On the layered nearest neighbour estimate

Gérard Biau, Luc Devroye

Éloïse Berthier, Guillaume Dalle, Clément Mantoux

MAP585 - Théorie de l'apprentissage

06/03/18

1. Layered Nearest Neighbours
   - Définitions
   - Comportement asymptotique
   - Complexité
2. Estimation par LNN
   - Théorème de consistance
   - Preuve du théorème
3. Extensions et discussion
   - Random forests et bagging
   - A propos des LNN

Cadre du problème

Données : $(X_1, Y_1), ..., (X_n, Y_n)$ iid $\in \mathbf{R}^d \times \mathbf{R}$. On suppose que X a une densité f par rapport à la mesure de Lebesgue, et que $\mathbb{E}[|Y|] < \infty$.
But : estimer la fonction de régression $r(x) = \mathbb{E}[Y|X = x]$

Définition : Layered Nearest Neighbours

X_i est un LNN de $x \in \mathbf{R}^d$ si l'hyperrectangle $\mathcal{R}(x, X_i)$ ne contient aucune autre observation X_j.
Soient $\mathcal{L}_n(x) = \{X_i \text{ LNN de } x\}$ et $L_n(x) = |\mathcal{L}_n(x)|$

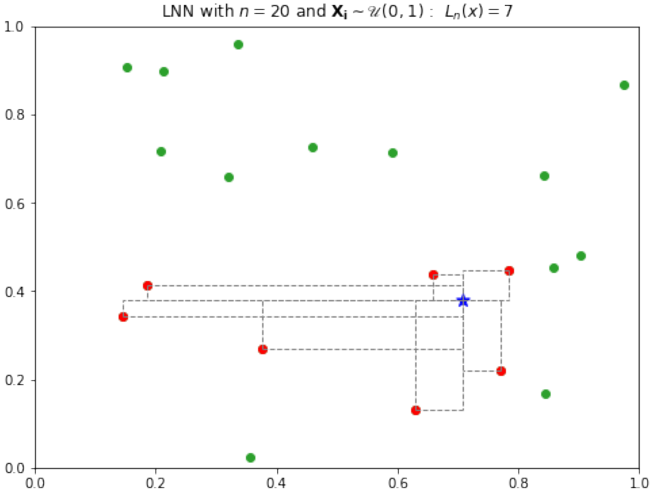


FIGURE – Exemple de LNN sur une loi uniforme

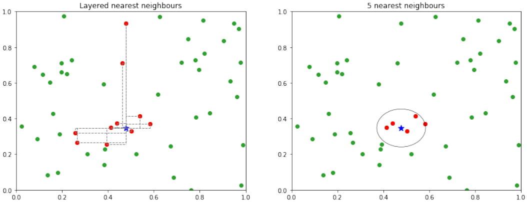


Figure – LNN / KNN en 2d

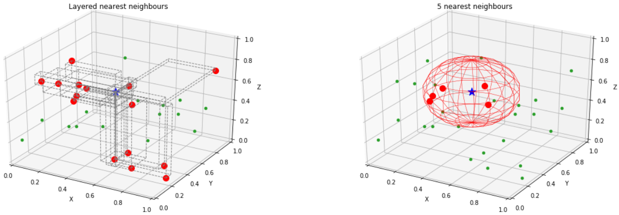


Figure – LNN / KNN en 3d

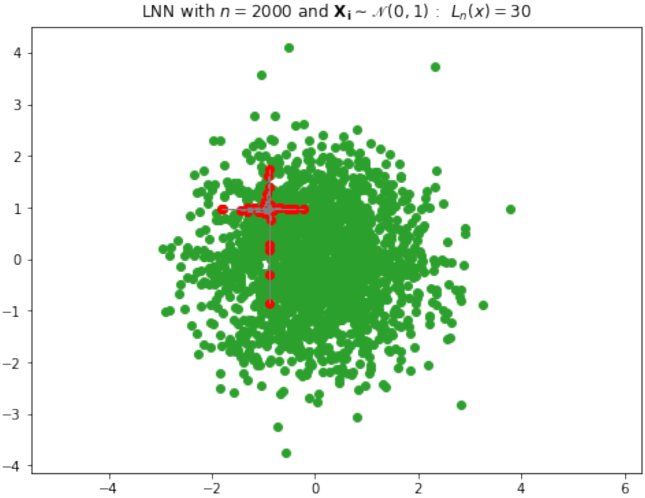


FIGURE – LNN : un voisinage étrange

Théorème 2.1 : Limite de $L_n(x)$

Pour $\mathbb{P}_X$-presque tout $x \in \mathbb{R}^d$, $L_n(x) \xrightarrow[n\to\infty]{\mathbb{P}} +\infty$

Théorème 2.1 : Limite de $L_n(x)$

Pour $\mathbb{P}_X$-presque tout $x \in \mathbb{R}^d$, $L_n(x) \xrightarrow[n\to\infty]{\mathbb{P}} +\infty$

Idées de la preuve

1. Bien choisir x et ε_n. Noter $\mathcal{R}_\varepsilon(x) = \prod [x_i, x_i + \epsilon]$.

Théorème 2.1 : Limite de $L_n(x)$

Pour $\mathbb{P}_X$-presque tout $x \in \mathbb{R}^d$, $L_n(x) \xrightarrow[n\to\infty]{\mathbb{P}} +\infty$

Idées de la preuve

1. Bien choisir x et ε_n. Noter $\mathcal{R}_\varepsilon(x) = \prod [x_i, x_i + \epsilon]$.
2. Construire $(W_1, ..., W_n)$ tel que W soit uniforme sur $\mathcal{R}_{\varepsilon_n}(x)$, et $\mathbb{P}(X_1^n \neq W_1^n) \to 0$

Théorème 2.1 : Limite de $L_n(x)$

Pour $\mathbb{P}_X$-presque tout $x \in \mathbb{R}^d$, $L_n(x) \xrightarrow[n\to\infty]{\mathbb{P}} +\infty$

Idées de la preuve

1. Bien choisir x et ε_n. Noter $\mathcal{R}_\varepsilon(x) = \prod [x_i, x_i + \epsilon]$.
2. Construire $(W_1, ..., W_n)$ tel que W soit uniforme sur $\mathcal{R}_{\varepsilon_n}(x)$, et $\mathbb{P}(X_1^n \neq W_1^n) \to 0$
3. Se ramener à étudier le nombre K_m de maxima de m VA iid uniformes : $\mathbb{E}[K_m] = \Omega((\log n)^{d-1})$ et $\mathbb{V}[K_m] = O((\log n)^{d-1})$

Théorème 2.1 : Limite de $L_n(x)$

Pour $\mathbb{P}_X$-presque tout $x \in \mathbb{R}^d$, $L_n(x) \xrightarrow[n\to\infty]{\mathbb{P}} +\infty$

Idées de la preuve

1. Bien choisir x et ε_n. Noter $\mathcal{R}_\varepsilon(x) = \prod [x_i, x_i + \epsilon]$.
2. Construire $(W_1, ..., W_n)$ tel que W soit uniforme sur $\mathcal{R}_{\varepsilon_n}(x)$, et $\mathbb{P}(X_1^n \neq W_1^n) \to 0$
3. Se ramener à étudier le nombre K_m de maxima de m VA iid uniformes : $\mathbb{E}[K_m] = \Omega((\log n)^{d-1})$ et $\mathbb{V}[K_m] = O((\log n)^{d-1})$
4. Faire que quand $n \to \infty$, $N_n = |\{X_i\} \cap \mathcal{R}_{\varepsilon_n}(x)| \to \infty$

Théorème 2.1 : Limite de $L_n(x)$

Pour $\mathbb{P}_X$-presque tout $x \in \mathbb{R}^d$, $L_n(x) \xrightarrow[n\to\infty]{\mathbb{P}} +\infty$

Idées de la preuve

1. Bien choisir x et ε_n. Noter $\mathcal{R}_\varepsilon(x) = \prod [x_i, x_i + \epsilon]$.
2. Construire $(W_1, ..., W_n)$ tel que W soit uniforme sur $\mathcal{R}_{\varepsilon_n}(x)$, et $\mathbb{P}(X_1^n \neq W_1^n) \to 0$
3. Se ramener à étudier le nombre K_m de maxima de m VA iid uniformes : $\mathbb{E}[K_m] = \Omega((\log n)^{d-1})$ et $\mathbb{V}[K_m] = O((\log n)^{d-1})$
4. Faire que quand $n \to \infty$, $N_n = |\{X_i\} \cap \mathcal{R}_{\varepsilon_n}(x)| \to \infty$
5. Pour N_n grand, $\mathbb{P}(K_{N_n} \geq A|N_n) \leq \mathbb{P}(K_{N_n} \geq \mathbb{E}[K_{N_n}|N_n]/2|N_n)$, et majorer par Tchebychev

Théorème 2.2 : Équivalent de $L_n(x)$

Si f est $\mathscr{C}^0$ presque partout, alors pour $\mathbb{P}_X$ presque tout $x \in \mathbb{R}^d$,

$$\mathbb{E}[L_n(x)] \underset{n\to\infty}{\sim} \frac{2^d(\log n)^{d-1}}{(d-1)!}$$

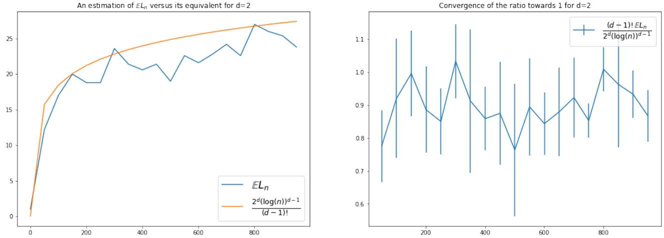


FIGURE – Equivalent de $L_n(x)$

Complexité KNN

- Calculer tous les $\|x - X_i\|$: $O(nd)$
- Trouver la k-ème plus petite distance par *quickselect* : $O(n)$
- Trouver les k éléments les plus proches : $O(n)$

Complexité moyenne : $O(nd)$

Complexité LNN

- Pour tout point X_i : parcourir les X_j et tester s'ils sont dans $\mathcal{R}(x, X_i)$. Passer au j suivant dès la première coordonnée tombant hors de l'intervalle $[x^k, X_i^k]$.
- Passer au i suivant dès qu'un test sur j est positif.
- Les X_i pour lesquels le test est négatif sont les LNN.

Complexité moyenne : $o(n^2 d)$

1. Layered Nearest Neighbours
   - Définitions
   - Comportement asymptotique
   - Complexité

2. Estimation par LNN
   - Théorème de consistance
   - Preuve du théorème

3. Extensions et discussion
   - Random forests et bagging
   - A propos des LNN

Définition : Estimateur LNN

$$r_n(x) = \frac{1}{L_n(x)} \sum_{i=1}^{n} Y_i \mathbf{1}_{X_i \in \mathcal{L}_n(x)}$$

Théorème 3.1 : Consistance ponctuelle de l'estimateur LNN

Supposons que :

1. $|Y| \leq \gamma < +\infty$ (p.s.)
2. r est continue (p.p.)

Alors pour tout $p \in \mathbb{N}^*$, pour $\mathbb{P}_X$-presque tout $x \in \mathbb{R}^d$,

$$\mathbb{E}|r(x) - r_n(x)|^p \xrightarrow[n \to +\infty]{} 0$$

- Plan de la preuve : séparations et contrôle individuel des termes.
- Pour simplifier, **on considère uniquement le premier quadrant** (sur 2^d au total).

- On sait que $|a+b|^p \leq A|a|^p + B|b|^p$

$$\begin{aligned}
\mathbb{E}|r_n(x) - r(x)|^p &= \mathbb{E}\left|\frac{1}{L_n(x)}\sum_{i=1}^{n}\mathbf{1}_{X_i \in \mathcal{L}_n(x)} Y_i - r(x)\right|^p \\
&= \mathbb{E}\left|\frac{1}{L_n(x)}\sum_{i=1}^{n}\mathbf{1}_{X_i \in \mathcal{L}_n(x)} (Y_i - r(x))\right|^p \\
&\leq A\mathbb{E}\left|\frac{1}{L_n(x)}\sum_{i=1}^{n}\mathbf{1}_{X_i \in \mathcal{L}_n(x)} (Y_i - {\color{red}r(X_i)})\right|^p \\
&+ B\mathbb{E}\left[\frac{1}{L_n(x)}\sum_{i=1}^{n}\mathbf{1}_{X_i \in \mathcal{L}_n(x)} |{\color{red}r(X_i)} - r(x)|^p\right]
\end{aligned}$$

- Pour le terme $(Y_i - r(X_i))$, on utilise un lemme d'analyse qui donne :

$$\text{Terme } (Y_i - r(X_i)) \leq \gamma^p C \mathbb{E}\left[\frac{1}{L_n(x)} \underbrace{\sum_{i=1}^{n} \frac{1}{L_n(x)} \mathbf{1}_{X_i \in \mathcal{L}_n(x)}}_{=1}\right]^{p/2}$$

$$= \gamma^p C \mathbb{E}\left[\frac{1}{L_n(x)}\right]^{p/2}$$

- Et, comme $L_n(x) \xrightarrow{\mathbb{P}} +\infty$, on a $\mathbb{E}\left[\frac{1}{L_n(x)}\right]^{p/2} \longrightarrow 0$

- On considère donc

$$\mathbb{E}\left[\frac{1}{L_n(x)}\sum_{i=1}^{n}\mathbf{1}_{X_i\in\mathcal{L}_n(x)}|r(X_i)-r(x)|^p\right]$$

- Soit $R_\varepsilon = [x_1, x_1+\varepsilon] \times ... \times [x_d, x_d+\varepsilon]$. On sépare avec $\mathbf{1}_{X_i\in R_\varepsilon}$:
  - Si $X_i \in R_\varepsilon$, alors

$$|r(X_i)-r(x)| \leq \sup_{z\in R_\varepsilon}|r(z)-r(x)|$$

  - Si $X_i \notin R_\varepsilon$, comme $r(x) = \mathbb{E}[Y|X=x]$ est bornée par γ,

$$|r(X_i)-r(x)| \leq 2\gamma$$

- Donc :

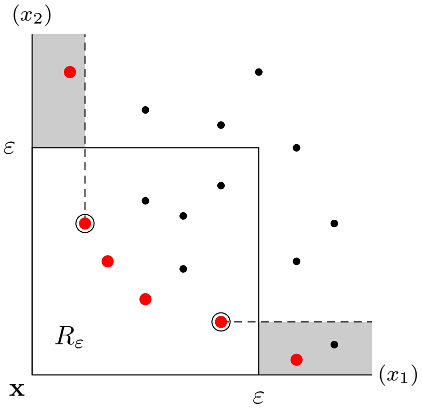
$$\mathbb{E}\left[\frac{1}{L_n(x)}\sum_{i=1}^{n}\mathbf{1}_{X_i\in\mathcal{L}_n(x)}|r(X_i)-r(x)|^p\right]$$
$$=\mathbb{E}\left[\frac{1}{L_n(x)}\sum_{i=1}^{n}\mathbf{1}_{X_i\in\mathcal{L}_n(x)}\mathbf{1}_{X_i\in R_\varepsilon}|r(X_i)-r(x)|^p\right]$$
$$+\,\mathbb{E}\left[\frac{1}{L_n(x)}\sum_{i=1}^{n}\mathbf{1}_{X_i\in\mathcal{L}_n(x)}\mathbf{1}_{X_i\notin R_\varepsilon}|r(X_i)-r(x)|^p\right]$$

- Donc :

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{L_n(x)}\sum_{i=1}^{n}\mathbf{1}_{X_i\in\mathcal{L}_n(x)}|r(X_i)-r(x)|^p\right] \\
&= \mathbb{E}\left[\frac{1}{L_n(x)}\sum_{i=1}^{n}\mathbf{1}_{X_i\in\mathcal{L}_n(x)}\mathbf{1}_{X_i\in R_\varepsilon}|r(X_i)-r(x)|^p\right] \\
&+ \mathbb{E}\left[\frac{1}{L_n(x)}\sum_{i=1}^{n}\mathbf{1}_{X_i\in\mathcal{L}_n(x)}\mathbf{1}_{X_i\notin R_\varepsilon}|r(X_i)-r(x)|^p\right] \\
&\leq \sup_{z\in R_\varepsilon}|r(z)-r(x)|^p + (2\gamma)^p\mathbb{E}\left[\frac{1}{L_n(x)}\sum_{i=1}^{n}\mathbf{1}_{X_i\in\mathcal{L}_n(x)}\mathbf{1}_{X_i\notin R_\varepsilon}\right]
\end{aligned}$$

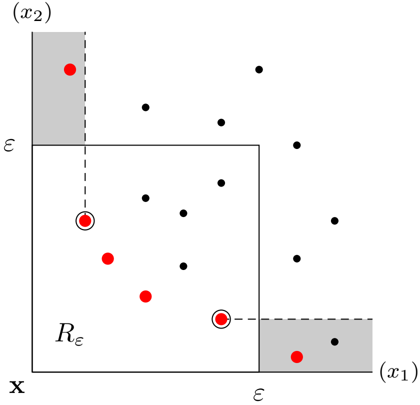
- r est **continue** donc le premier terme tend vers 0.

- On considère donc $\mathbb{E}\left[\frac{1}{L_n(x)}\sum_{i=1}^{n}\mathbf{1}_{X_i\in\mathcal{L}_n(x)}\mathbf{1}_{X_i\notin R_\varepsilon}\right]$
- Soit $\mathcal{P}_\varepsilon$ la zone grise :



- **Idée :** quand $n \to +\infty$, tous les LNN sont dans $R_\varepsilon \cup \mathcal{P}_\varepsilon$.

- On a un résultat technique : $|\{X_i \in \mathcal{P}_\varepsilon\}| = \mathcal{O}_\mathbb{P}(1)$



- **Rappel :** on considère $\mathbb{E}\left[\frac{1}{L_n(x)} \sum_{i=1}^{n} \mathbf{1}_{X_i \in \mathcal{L}_n(x)} \mathbf{1}_{X_i \notin R_\varepsilon}\right]$

- Et $\mathbf{1}_{X_i \in \mathcal{L}_n(x)} \mathbf{1}_{X_i \notin R_\varepsilon} \leq \mathbf{1}_{X_i \in \mathcal{P}_\varepsilon}$ pour $n \longrightarrow +\infty$

- On a donc

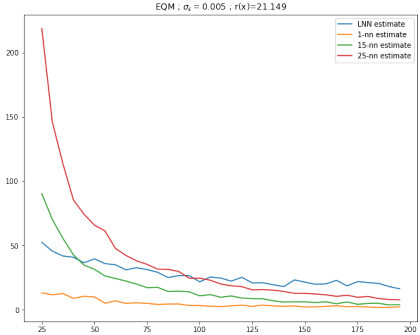
$$\begin{aligned} \mathbb{E}\left[\frac{1}{L_n(x)} \sum_{i=1}^{n} \mathbf{1}_{X_i \in \mathcal{L}_n(x)} \mathbf{1}_{X_i \notin R_\varepsilon}\right] &\leq \mathbb{E}\left[\frac{1}{L_n(x)} \sum_{i=1}^{n} \mathbf{1}_{X_i \in \mathcal{P}_\varepsilon}\right] \\ &\leq \mathbb{E}\left[\frac{|\{X_i \in \mathcal{P}_\varepsilon\}|}{L_n(x)}\right] \end{aligned}$$

- Et on a vu que $L_n(x) \xrightarrow{\mathbb{P}} +\infty$ et que $|\{X_i \in \mathcal{P}_\varepsilon\}| = \mathcal{O}_{\mathbb{P}}(1)$.

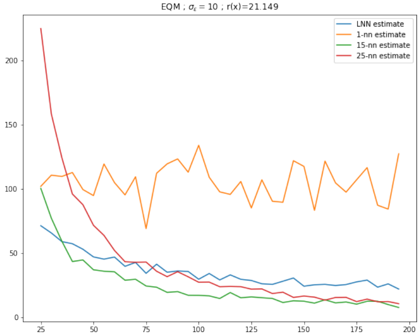
- On obtient donc finalement

$$\mathbb{E}\left[\frac{1}{L_n(x)}\sum_{i=1}^{n}\mathbf{1}_{X_i\in\mathcal{L}_n(x)}\mathbf{1}_{X_i\notin R_\varepsilon}\right] \longrightarrow 0$$

$\square$



(a) bruit faible



(b) bruit fort

FIGURE – Estimation par LNN vs kNN

1. Layered Nearest Neighbours
   - Définitions
   - Comportement asymptotique
   - Complexité
2. Estimation par LNN
   - Théorème de consistance
   - Preuve du théorème
3. Extensions et discussion
   - Random forests et bagging
   - A propos des LNN

Remarque : Lien entre LNN et RF

Une forêt aléatoire dont les arbres séparent en rectangles et jusqu'à ce qu'il reste au plus 1 point par cellule s'interprète comme un estimateur LNN pondéré.

Proposition 3.1 : Borne inférieure sur l'erreur

Supposons que $\sigma^2 = \mathbb{V}[Y|X = x]$ est indépendante de x. Alors :

$$\forall x \in \mathbf{R}^d,\ \mathbb{E}[|r_n(x) - r(x)|^2] \geq \frac{\sigma^2}{\mathbb{E}[L_n(x)]}$$

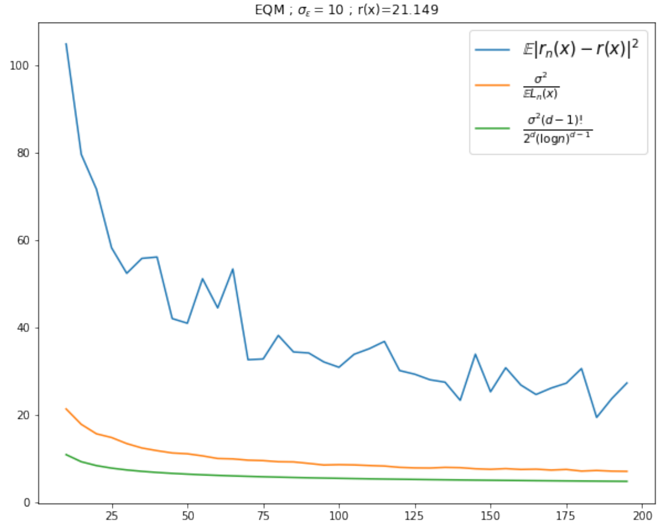


FIGURE – Borne inférieure sur l'erreur et erreur des LNN

Remarque : Conséquences sur l'estimateur LNN

L'erreur de l'estimateur LNN est en $\Omega\left(\frac{1}{(\log n)^{d-1}}\right)$: convergence lente.
Solution possible : bagging.

Proposition 4.1 : Consistance du 1NN + bagging

Dans l'échantillon de taille n, on fait m tirages d'un sous-ensemble de taille k. On prend comme estimateur final r_n^* la moyenne des estimateurs 1NN associés à ces m tirages.
Si $m \to \infty$, $k \to \infty$ et $k/n \to 0$, alors r_n^* est universellement L^p-consistant (pour $p > 1$).

Idées de la preuve

1. Remarquer que si $m = \infty$, r_n^* est un estimateur NN pondéré :

$$r_n^* = \sum_{i=1}^{n} V_i Y_{(i)}(x)$$

avec pour poids $V_i = \mathbb{P}$(le i-ème NN de x est *choisi* parmi une sélection de k points).

Idées de la preuve

1. Remarquer que si $m = \infty$, r_n^* est un estimateur NN pondéré :

$$r_n^* = \sum_{i=1}^{n} V_i Y_{(i)}(x)$$

   avec pour poids $V_i = \mathbb{P}(\text{le } i\text{-ème NN de } x \text{ est } \textit{choisi} \text{ parmi une sélection de } k \text{ points})$.

2. Calculer ces probabilités dans les cas des tirages avec et sans remise.

Idées de la preuve

1. Remarquer que si $m = \infty$, r_n^* est un estimateur NN pondéré :

$$r_n^* = \sum_{i=1}^{n} V_i Y_{(i)}(x)$$

   avec pour poids $V_i = \mathbb{P}($le i-ème NN de x est *choisi* parmi une sélection de k points$)$.
2. Calculer ces probabilités dans les cas des tirages avec et sans remise.
3. Montrer qu'ils vérifient les **conditions de Stone**.

Idées de la preuve

1. Remarquer que si $m = \infty$, r_n^* est un estimateur NN pondéré :

$$r_n^* = \sum_{i=1}^{n} V_i Y_{(i)}(x)$$

   avec pour poids $V_i = \mathbb{P}($le i-ème NN de x est *choisi* parmi une sélection de k points$)$.

2. Calculer ces probabilités dans les cas des tirages avec et sans remise.
3. Montrer qu'ils vérifient les **conditions de Stone**.
4. Le cas où $m < \infty, m \to \infty$ se ramène au précédent en remarquant que les poids sont des v.a. $(W_1, ..., W_n) \overset{\mathscr{L}}{=} \frac{\text{Multinomial}(m; V_1, ..., V_n)}{m}$, avec $\sum_{i=1}^{n} W_i = 1$.

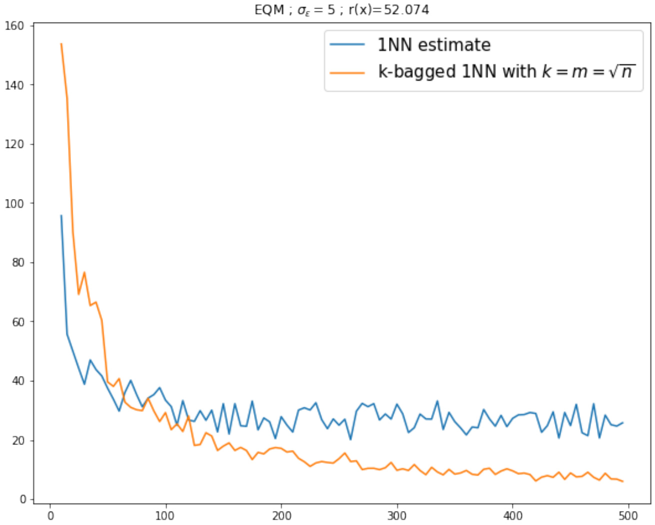


FIGURE – Erreur de l'estimateur des 1NN avec et sans bagging

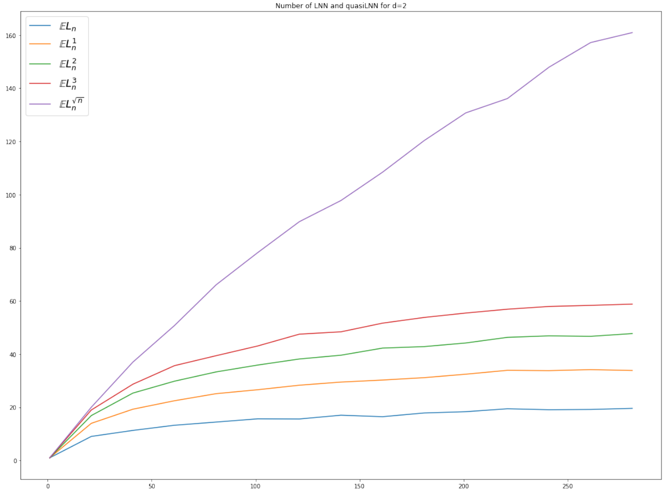


FIGURE – Généralisation du LNN : asymptotique

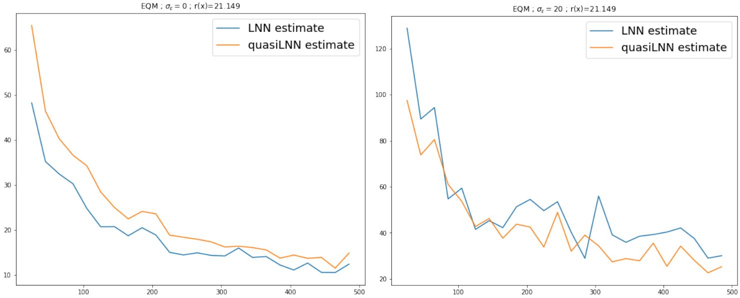


FIGURE – Généralisation du LNN : consistance