L'apprentissage face à la malédiction de la grande dimension

Collège de France

Cours 3: **Malédiction de la grande dimension**

de Stéphane Mallat

Notes de John Zarka et Stéphane Mallat

1 Erreur de généralisation

1.1 Retour sur le dilemme biais-complexité

Je vais d'abord rappeller les principaux résultats obtenus lors du cours précédent sur la majoration de l'erreur de généralisation, qui met en évidence le *dilemme biais-variance* ou *dilemme biais-complexité.*

Les n données d'entrainement sont notées $\{(x_i, y_i)\}_{i\leq n}$, où $x_i \in \Omega \subseteq \mathbb{R}^d$ sont les données et $y_i \in \mathcal{A}$ les réponses associées. Ce sont des réalisations de n variables aléatoires i.i.d $\{(X_i, Y_i)\}_{i\leq n}$ ayant la même distribution jointe que le couple (X, Y) représentant la distribution des données et des réponses. On note $\tilde{f}$ le minimiseur du risque *empirique* $\tilde{R}_e(\cdot)$ sur les données d'entraînement $\{(x_i, y_i)\}_{i\leq n}$:

$$\tilde{f} = \arg\min_{h\in\mathcal{H}} \tilde{R}_e(h)$$

où $\mathcal{H}$ est la classe de fonctions considérée par notre algorithme. On note $r : \mathcal{A}\times\mathcal{A} \to \mathbb{R}$ une mesure de risque entre la sortie y et sa prédiction $\tilde{y}$, et l'estimateur du risque moyen est :

$$\tilde{R}_e(h) = \frac{1}{n}\sum_{i=1}^{n} r(h(x_i), y_i).$$

On note également f_a le minimisateur du risque de *généralisation* $R(\cdot)$ défini en moyenne sur la distribution jointe de (X, Y) par :

$$f_a = \arg\min_{h\in\mathcal{H}} R(h) \tag{1}$$

avec :

$$R(h) = \mathop{\mathbb{E}}_{(X,Y)} \left[r\left(h(X), Y\right)\right].$$

Si la réponse y est définie de manière unique à partir de l'entrée x, on peut l'écrire $y = f(x)$ et donc

$$R(h) = \mathop{\mathbb{E}}_{X} \left[r\left(h(X), f(X)\right)\right].$$

Dans ce cas, f_a est la meilleure approximation de f au sein de la classe $\mathcal{H}$ pour la distance d définie par :

$$d(f, h) = \mathop{\mathbb{E}}_{X} \left[r\left(h(X), f(X)\right)\right].$$

Nous avons démontré le résultat suivant.

Proposition 1

$$R(f_a) \leq R(\tilde{f}) \leq \overbrace{R(f_a)}^{erreur\ de\ modèle} + 2 \overbrace{\max_{h \in \mathcal{H}} |R(h) - \tilde{R}_e(h)|}^{erreur\ de\ fluctuation} \tag{2}$$

Cette proposition montre que l'erreur de généralisation $R(\tilde{f})$ de notre minimisateur empirique $\tilde{f}$ peut-être majorée par un terme de biais $R(f_a)$ correspondant à une erreur de modèle et un terme de variance $\max_{h \in \mathcal{H}} |R(h) - \tilde{R}_e(h)|$ correspondant à une erreur de fluctuation.
L'erreur de modèle ou erreur d'approximation mesure la distance (en moyenne) entre f et sa meilleure approximation au sein de $\mathcal{H}$ pour la mesure de risque r. L'erreur de fluctuation mesure elle la différence maximale entre le risque empirique et le risque de généralisation. La borne uniforme est due au fait que le minimisateur du risque empirique $\tilde{f}$ est un prédicteur aléatoire pouvant se balader sur la classe de fonctions $\mathcal{H}$.

Ces deux erreurs évoluent (n étant fixé) dans des sens différents comme illustré dans la figure 1 : plus $\mathcal{H}$ est grand, meilleure sera la meilleure approximation f_a de f et plus l'erreur de biais sera faible tandis qu'au contraire l'erreur de fluctuation augmentera car $\tilde{f}$ aura tendance à surapprendre sur les données d'entrainement et mal généraliser. L'optimum se situe donc à n fixé pour un $\mathcal{H}$ tel que ces deux erreurs soient du même ordre.

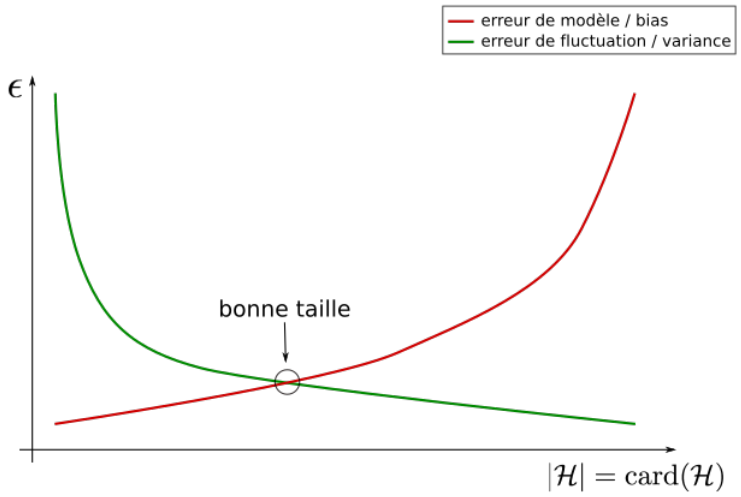

FIGURE 1 – Dilemme biais-fluctuation : le terme de biais décroit tandis que celui de fluctuation croit avec le cardinal de $\mathcal{H}$ à n fixé

Le théorème suivant, démontré à l'aide du lemme d'Hoeffding, donne une borne supérieure sur l'erreur de fluctuation.

Théorème 1 *Si $R(h) \in [0,1]\, \forall h$ et le cardinal $|\mathcal{H}|$ de $\mathcal{H}$ est fini, alors :*

$$\mathbb{P}\left(\max_{h \in \mathcal{H}} |R(h) - \tilde{R}_e(h)| \leq \epsilon\right) \geq 1 - \delta \quad si \quad \epsilon^2 = \frac{\log(|\mathcal{H}|) + \log(2/\delta)}{2n} \tag{3}$$

1.2 Algorithme du plus proches voisin

L'algorithme du plus proche voisin illustre le phénomène de non-contrôle de l'erreur de fluctuation, lorsque la classe de fonctions $\mathcal{H}$ est trop grande relativement aux nombres de données d'entraînement n. Cet algorithme produit une erreur empirique nulle mais il généralise mal car l'erreur de fluctuation est importante.

L'algorithme du plus proche voisin calcule le prédicteur $\tilde{f}$ à partir des données d'entraînement $\{(x_i, y_i)\}_{i \leq n}$ de la manière suivante :

$$\forall x \in \Omega \ , \quad \tilde{f}(x) = y_i \ \text{ si } \ \|x - x_i\| \leq \|x - x_i'\| \, \forall i' \neq i$$

où $\|\cdot\|$ est n'importe quelle norme sur $\mathbb{R}^d$. Cet algorithme découpe l'espace en "régions de Voronoï" $(V_i)_{1\leq i\leq n}$ définies par :

$$\Omega = \cup_{1\leq i\leq n} V_i \text{ avec } V_i = \{x \in \Omega \,/\, \|x - x_i\| \leq \|x - x_i'\| \, \forall i' \neq i\}$$

et

$$\tilde{f}(x) = \sum_{i=1}^{n} y_i \mathbf{1}_{\{x\in V_i\}} \,. \tag{4}$$

Comme $\tilde{f}(x_i) = y_i$, l'erreur empirique $\tilde{R}_e(\tilde{f})$ vaut 0. En revanche l'erreur de fluctuation est potentiellement importante pour tout n. La fonction $\tilde{f}$ appartient à une classe de fonctions $\mathcal{H}_n$ constantes par morceaux définies par (4) pour tous les $\{(x_i, y_i)\}_{i\leq n}$ possibles. Si les valeurs de x_i et y_i sont quantifiées sur N values alors le cardinal $|\mathcal{H}_n|$ est N^n. Comme $\log|\mathcal{H}_n| = n \log N$ on voit que la borne supérieure de l'erreur de fluctuation du théorème 1 ne diminue pas lorsque n augmente.

1.3 Décroissance de l'erreur de généralisation

La borne supérieure de l'erreur de fluctuation (en rouge sur le graphe) dépend à n fixé de $|\mathcal{H}|$ ou plus précisément $\log(|\mathcal{H}|)$. Pour minimiser la borne supérieure de la somme de l'erreur de modèle et de l'erreur de fluctuation, on veut trouver un ensemble $\mathcal{H}$ de fonctions telles que l'erreur d'approximation décroit le plus rapidement possible avec $\log(|\mathcal{H}|)$. Si on équilibre l'erreur d'approximation et l'erreur de fluctuation (intersection des courbes verte et rouge), alors on atteint le mininum à un facteur 2 prés.

Étant donnée la loi de décroissance de l'erreur d'approximation $R(f_a)$ d'un ensemble de classes $\mathcal{H}$ en fonction de $\log(|\mathcal{H}|)$, la proposition suivante en déduit une borne supérieure de l'erreur de généralisation pour un n nombre d'exemples suffisemment grand.

Proposition 2 *Si il existe $\beta > 0$ et $C > 0$ tels que $R(f_a) \leq C(\log(|\mathcal{H}|))^{-\beta}$ alors pour tout $\epsilon > 0$,*

$$\mathbb{P}\left(R(\tilde{f}) \leq 3\epsilon\right) \geq 1 - 2\mathrm{e}^{-C^{1/\beta}\epsilon^{-1/\beta}} \quad si \;\; n \geq C^{1/\beta}\epsilon^{-2-1/\beta}. \tag{5}$$

Démonstration : Nous allon appliquer l'inégalité (3) en ajustant δ et $\log(|\mathcal{H}|)$ en fonction de ϵ et n avec

$$\log(|\mathcal{H}|) = \log(2/\delta) = n\epsilon^2$$

et donc $\delta = 2e^{-n\epsilon^2}$. L'inégalité (3) implique que

$$\mathbb{P}\left(\max_{h\in\mathcal{H}} |R(h) - \tilde{R}_e(h)| \leq \epsilon\right) \geq 1 - \delta$$

Comme $R(f_a) \leq C(\log(|\mathcal{H}|))^{-\beta}$, on en déduit que $R(f_a) \leq \epsilon$ si

$$C(\log(|\mathcal{H}|))^{-\beta} \leq \epsilon$$

et donc si

$$n\epsilon^2 = \log(|\mathcal{H}|) \geq C^{1/\beta}\epsilon^{-1/\beta}. \tag{6}$$

En appliquant l'equation (2) on obtient (5). □

Ce théorème montre que plus l'erreur d'approximation de l'ensemble de classes $\mathcal{H}$ décroit rapidement avec $\log(|\mathcal{H}|)$ i.e plus β est grand, moins on a besoin d'exemples n afin de majorer l'erreur moyenne $R(\tilde{f})$ par une certaine erreur 3ϵ, et réciproquement étant donné un certain nombre d'exemples n plus cette erreur 3ϵ sera faible. Le problème consiste donc à pouvoir estimer la décroissance de l'erreur d'approximation en fonction de la taille $\log(|\mathcal{H}|)$ des classes de fonction $\mathcal{H}$ considérées.

1.4 Erreur d'approximation

Nous allons maintenant nous concentrer sur l'erreur d'approximation et calculer une borne supérieure qui ne dépend pas de la distribution de probabilité de X et Y.

On suppose que la réponse y est définie de manière unique à partir de la donnée x et donc $y = f(x)$. Dans ce cas

$$R(f_a) = \min_{h \in \mathcal{H}} d(f,h) \quad \text{avec} \quad d(f,h) = \mathop{\mathbb{E}}_{X} \left[r\left(h(X), f(X)\right) \right].$$

Par exemple $r(y, \tilde{y}) = (y - \tilde{y})^2$ définit une erreur quadratique moyenne.

Déterminer f_a et calculer $R(f_a)$ revient à résoudre un problème d'approximation de fonctions. Le problème est que le plus souvent on ne connait pas la distribution de probabilité de X et donc la distance d. On obtient un résultat uniforme quelle que soit la loi de X en majorant le terme $r\left(h(X), f(X)\right)$ par $\sup_{x \in \Omega} r\left(h(x), f(x)\right)$:

$$R(f_a) \leq \min_{h \in \mathcal{H}} \sup_{x \in \Omega} |h(x) - f(x)| = \min_{h \in \mathcal{H}} \|h - f\|_\infty.$$

en notant $\|f\|_\infty = \sup_{x \in \Omega} |f(x)|$.

On ne connait pas f mais on suppose que l'on a de l'information *à priori* qui permet de garantir que f appartient à une classe particulière de fonctions $\mathcal{F}$. Cet *à priori* sur la classe de fonctions $\mathcal{F}$ va guider le choix des algorithmes et des ensembles d'approximations $\mathcal{H}$. Pour majorer l'erreur pour tout f dans $\mathcal{F}$, nous allons controler l'erreur maxium :

$$\max_{f \in \mathcal{F}} R(f_a) \leq \max_{f \in \mathcal{F}} \min_{h \in \mathcal{H}} \|f - h\|_\infty,$$

et on va tenter de trouver des bornes d'approximation du type

$$\max_{f \in \mathcal{F}} \min_{h \in \mathcal{H}} \|f - h\|_\infty \leq C \log(|\mathcal{H}|)^{-\beta}$$

pour pouvoir appliquer la Proposition 2.

L'information *a priori* que l'on met sur la classe de fonctions $\mathcal{F}$ peut s'interpréter comme une forme de *régularité* qui explicite le fait que f n'est pas n'importe quel fonction. Plus une fonction f est "régulière" (dans un sens à définir) moins on aura besoin d'échantillons $(x_i, y_i = f(x_i))$ pour qu'une interpolation de ces échantillons l'approxime bien.

2 Malédiction de la grande dimension

Il y a beaucoup de façon différentes de spécifier la notion de régularité. Nous allons commencer par la régularité Lipschitzienne qui spécifie la régularité locale d'une fonction. Nous allons montrer comment la régularité Lipschitzienne se relie à la vitesse d'approximation des fonctions $f(x)$ pour $x \in \mathbb{R}^d$. Nous verrons que le nombre d'exemples nécessaires augmente exponentiellement en fonction de d, ce qui n'est pas faisable dés que d est grand. C'est la malédiction de la grande dimension. Pour éviter cela il faudra trouver d'autres types de régularité plus fortes.

2.1 Régularité Lipschitzienne

En dimension un, la dérivabilité quantifie localement l'amplitude des variations d'une fonction f. Elle se généralise en dimensions supérieures par la différentiabilité au sens de Fréchet ou de Gâteaux. Cependant, on utilise souvent plutôt la notion de régularité *Lipschitzienne* qui est légèrement plus faible mais plus facile à manipuler. On note $\|x\|^2 = \sum_{u=1}^d |x(u)|^2$.

Définition 1 *On dit que $f : \Omega \to \mathbb{R}$ est localement Lipschitz en $x \in \Omega$ si il existe $C_x > 0$ telle que*

$$\forall x' \in \Omega \ , \ |f(x) - f(x')| \leq C_x \|x - x'\|.$$

On dit que f est uniformément Lipschitz sur Ω si f est Lipschitz en tout $x \in \Omega$ et s'il existe $C > 0$ telle que $C_x \leq C$.

On peut vérifier que si f est différentiable et que la norme de son Jacobien est bornée en tout point $x \in \mathbb{R}^d$ i.e $\sup_{x \in \mathbb{R}^d} \|\nabla f(x)\| \leq C$ sur $\mathbb{R}^d$ alors f est uniformément Lipschitz sur $\mathbb{R}^d$ pour la constante C. En dimension 1, pour tout $(x, x') \in \mathbb{R}^2$ avec $x < x'$ on a :

$$|f(x) - f(x')| = |\int_{x'}^{x} f'(t) \mathrm{d}t| \leq \int_{x'}^{x} |f'(t)| \mathrm{d}t \leq \int_{x'}^{x} C \mathrm{d}t = C|x - x'|$$

La réciproque est plus subtile. Si $f : \mathbb{R} \to \mathbb{R}$ est uniformément Lipschitz alors on peut montrer qu'elle est dérivable presque partout. En dimension supérieure, la régularité lipschitzienne implique la différentiabilité presque partout au sens de Gâteaux. La différentiabilité au sens de Gâteaux correspondant à l'existence des dérivées directionnelles.

La régularité lipschitzienne est donc une mesure de régularité très similaire à la dérivabilité directionnelle mais légèrement plus souple. C'est une mesure de régularité locale qui spécifie les variations d'amplitude d'une fonction lorsque l'on bouge autour de points x. De même que l'on peut définir des dérivées d'ordre supérieur, on peut définir des notions de régularité Lipschitzienne d'ordre supérieur de la manière suivante.

Définition 2 *On dit que $f : \Omega \to \mathbb{R}$ est localement Lipschitz α en $x \in \Omega$ si il existe une constante $C_x > 0$ et un polynôme p_x de degré $q \leq \lfloor \alpha \rfloor$ tels que*

$$\forall x' \in \Omega \ , \ |f(x') - p_x(x')| \leq C_x \|x - x'\|^\alpha.$$

On dit que f est uniformément Lipschitz α sur Ω si f est Lipschitz α en tout $x \in \Omega$ et s'il existe $C > 0$ telle que $C_x \leq C$.

Cette régularité Lipschitzienne d'ordre supérieur mesure la vitesse de décroissance α du résidu entre $f(x')$ et sa meilleure approximation polynomiale p_x partant de x. Nous allons l'utiliser comme l'*à priori* en supposant que les fonctions f que l'on veut apprendre appartient à des classes de fonctions $\mathcal{F}_\alpha$:

$$\mathcal{F}_\alpha = \{f : \Omega \to \mathbb{R} \ : \ f \text{ est uniformément Lipschitz } \alpha\}$$

L'enjeu va alors être de contrôler l'erreur d'approximation uniforme $\max_{f \in \mathcal{F}} \min_{h \in \mathcal{H}} \|f - h\|_\infty$ des classes d'approximations $\mathcal{H}$, en fonction de $\log(|\mathcal{H}|)$. Pour cela nous allons calculer combien de paramètres et d'exemples il nous faut pour obtenir une erreur d'approximation ϵ.

2.2 La malédiction de l'approximation de fonctions Lipschitziennes

Nous allons considérer l'algorithme d'approximation du plus proche voisin pour approximer des fonctions qui sont uniformément Lipschitz ($\alpha = 1$). Cet algorithme ne permet pas de contrôler l'erreur de fluctuation mais il est à priori efficace pour réduire l'erreur d'approximation. En effet il calcule des approximations constantes par morceaux autour des examples, or la régularité Lipschitzienne contrôle précisément les amplitudes des variations locales.

Nous allons néanmoins voir qu'en grande dimension, l'erreur d'approximation de l'algorithme du plus proche voisin va décroitre très lentement avec le nombre d'exemples : c'est la *malédiction de la dimensionalité*. À partir de n exemples $(x_i, f(x_i))_i$, l'algorithme du plus proche voisin calcule

$$\tilde{f}(x) = f(x_i) \ \text{ pour } \ i = \arg\min_{i' \leq n} \|x - x_{i'}\|$$

Proposition 3 *Si $\mathcal{F}$ est la classe des fonctions uniformément Lipschitz ($\alpha = 1$) de constante C sur Ω compact, alors*

$$\sup_{f \in \mathcal{F}} \|f - \tilde{f}\|_\infty = C\epsilon \quad \textit{avec} \quad \epsilon = \sup_{x \in \Omega} \min_{i \leq n} \|x - x_i\|. \tag{7}$$

Démonstration. Puisque $f \in \mathcal{F}$ est uniformément Lipschitz, $\exists C > 0$ tel que $\forall i \leq n$, $\forall x \in \Omega$:

$$|f(x) - f(x_i)| \leq C\|x - x_i\|.$$

Or $\tilde{f}(x) = f(x_i)$ pour $i = \arg\min_{i' \leq n} \|x - x_{i'}\|$, donc

$$|f(x) - \tilde{f}(x)| \leq C \min_{i' \leq n} \|x - x_{i'}\|$$

ce qui implique que

$$\|f - \tilde{f}\|_\infty = \sup_{x \in \Omega} |f(x) - \tilde{f}(x)| \leq C \sup_{x \in \Omega} \min_{i \leq n} \|x - x_i\|.$$

Inversement, on va montrer que la borne supérieure est atteinte. Comme Ω est compact, il existe x_j un élément de $\{x_i\}_{i\leq n}$ et $x \in \Omega$ tel que $\|x-x_j\| \leq \|x-x_i\|$ for tout $i \leq n$ et $\|x-x_j\| = \epsilon$. La fonction $f(x) = C\,\|x-x_j\|$ est dans $\mathcal{F}$ et $\|f-\tilde{f}\|_\infty = C\,\epsilon$, ce qui montre que la borne supérieure est atteinte. □

Afin de contrôler l'erreur d'approximation il nous faut donc nous assurer que la distance par rapport à l'exemple le plus proche n'est jamais trop grande. On va calculer le nombre d'exemples n dont on a besoin en fonction de la dimension d. On suppose ici que $\Omega = [0,1]^d$.

On se place ici dans le cadre où les exemples sont idéalement répartis, afin d'avoir une borne inférieur sur l'erreur que l'on obtiendrait avec des exemples distribués aléatoirement. Pour ϵ fixé, on note $B_\epsilon(x_i)$ la boule de rayon ϵ autour de l'exemple x_i. On a $\sup_{x\in\Omega} \min_{i\leq n} \|x-x_i\| \leq \epsilon$ si et seulement si les $B_\epsilon(x_i)$ forment un recouvrement de Ω :

$$\Omega \subseteq \cup_{i=1}^n B_\epsilon(x_i).$$

Le recouvrement par un nombre minimum de sphère est un problème que l'on recontre souvent en théorie de l'information, notamment pour la compression par quantification vectorielle. Le rayon ϵ de la boule correspond à l'erreur maximum de quantification lorsque l'on approxime x par un x_i. On veut recouvrir l'espace avec un nombre minimal de boules afin de minimiser le nombre de bit nécessaire pour spécifier chacune des boules, pour un ϵ fixé. En effet, le nombre de bits nécessaire pour ce codage est égale au log du nombre de boules, qui est l'entropie du recouvrement. Nous allons maintenant voir que le nombre de boules nécessaires augmente exponentiellement avec d.

Proposition 4 *Le rayon minimum ϵ de n boules qui recouvrent $\Omega = [0,1]^d$ satisfait*

$$\frac{\sqrt{d}n^{-1/d}}{2} \geq \epsilon \geq \frac{\sqrt{d}n^{-1/d}}{2} \underbrace{\sqrt{\frac{2}{\pi \mathrm{e}}}}_{\sim 0.484} \left(1 + \underset{+\infty}{O}\left(\frac{\log d}{d}\right)\right). \tag{8}$$

Démonstration : La borne supérieure de la proposition 4 s'obtient avec un échantillonnage de chacune des d directions avec un pas Δ. Le nombre de boules est alors $n = \Delta^{-d}$. lons x_i. Par ailleurs $\forall x \in \Omega$:

$$\min_i \|x-x_i\|^2 = \sum_{u=1}^d |x(u)-x_i(u)|^2 \leq d\frac{\Delta^2}{4}.$$

On en déduit que :

$$\epsilon = \sup_{x\in\Omega} \min_i \|x-x_i\| \leq \frac{\sqrt{d}\Delta}{2} = \frac{\sqrt{d}n^{-1/d}}{2}$$

ce qui démontre la boune supérieure de ϵ.

On peut cependant faire mieux en optimisant la position des centres des boules x_i dans l'espace. La borne inférieure se démontre en observant que comme $\Omega \subseteq \cup_i B_\epsilon(x_i)$ et et que $Vol\,(\Omega) = 1$, on obtient :

$$\sum_{i=1}^n Vol\,(B_\epsilon(x_i)) = n\,Vol\,(B_\epsilon) \geq 1, \tag{9}$$

car les boules $B_\epsilon(x_i)$ ont le même volume que l'on note $Vol\,(B_\epsilon)$. Lorsque d est pair, le volume d'une boule B_ϵ de rayon ϵ en dimension d est :

$$Vol\,(B_\epsilon) = \frac{\pi^{d/2}\epsilon^d}{(d/2)!}$$

La formule de Stirling pour $n \in \mathbb{N}$:

$$n! = \left(\frac{n}{\mathrm{e}}\right)^n \sqrt{2\pi n}\left(1 + \underset{+\infty}{O}\left(\frac{1}{n}\right)\right)$$

implique pour d pair (valable aussi pour d impair) :

$$\begin{aligned} Vol\,(B_\epsilon) &= \frac{\pi^{d/2}\epsilon^d}{\left(\frac{d}{2\mathrm{e}}\right)^{d/2}\sqrt{\pi d}}\left(1 + \underset{+\infty}{O}\left(\frac{1}{d}\right)\right) \\ &= \left(\frac{d}{2\mathrm{e}\pi}\right)^{-d/2} \frac{\epsilon^d}{\sqrt{\pi d}}\left(1 + \underset{+\infty}{O}\left(\frac{1}{d}\right)\right) \end{aligned} \tag{10}$$

L'équation (9) implique que :

$$n \left(\frac{d}{2e\pi} \right)^{-d/2} \frac{\epsilon^d}{\sqrt{\pi d}} \left(1 + \underset{+\infty}{O} \left(\frac{1}{d} \right) \right) \geq 1 \ .$$

Un calcul directe permet alors de vérifier que

$$\epsilon \geq \frac{\sqrt{d} n^{-1/d}}{2} \underbrace{\sqrt{\frac{2}{\pi e}}}_{\sim 0.484} \left(1 + \underset{+\infty}{O} \left(\frac{\log d}{d} \right) \right)$$

ce qui complète la preuve. □

En insérant la borne inférieur (8) dans (7), on obtient le corollaire suivant.

Corollaire 1 *Si $\mathcal{F}$ est la classe des fonctions uniformément Lipschitz de constante C sur Ω compact, alors*

$$\sup_{f \in \mathcal{F}} \|f - \tilde{f}\|_\infty \geq C \frac{\sqrt{d} n^{-1/d}}{2} \underbrace{\sqrt{\frac{2}{\pi e}}}_{\sim 0.484} \left(1 + \underset{+\infty}{O} \left(\frac{\log d}{d} \right) \right) .$$

Pour atteindre une erreur $C\epsilon$ il faut donc un nombre d'exemples qui satisfait :

$$n \geq \frac{\epsilon^{-d} \, d^{d/2}}{(2\pi e)^{d/2}} .$$

La vitesse de décroissance de l'erreur d'approximation du classificateur du plus proche voisin est donc extrêmement lente et il faut un nombre d'exemples pire qu'exponentielle en la dimension, afin d'obtenir une erreur d'approximation uniforme inférieure à ϵ. Si on suppose uniquement que la fonction est uniformément Lipschitz, dès que l'on est en dimension 5 ou 6 le nombre d'exemples requis devient gigantesque. On ne va ainsi jamais avoir assez d'exemples en pratique pour s'assurer que l'erreur d'approximation soit petite.

On peut démontrer qu'il n'existe pas d'ensemble fini $\mathcal{H}$ de fonctions qui puissent battrer l'exposant $\beta = 1/d$ de décroissance de l'erreur sur la classe des fonctions Lipschitziennes, au sense où

$$\max_{f \in \mathcal{F}} \min_{\tilde{f} \in \mathcal{H}} \|f - \tilde{f}\|_\infty \leq C \left(\log(|\mathcal{H}|) \right)^{-\beta} \quad \Rightarrow \beta \leq 1/d \ .$$

On voit que si l'on veut réduire notre erreur d'approximation d'un facteur 2, on a besoin au moins de 2^d fois plus de points, ce qui est à nouveau l'expression de la *malédiction de la dimensionalité*.

Si on augmente la régularité de la fonction en considérant des fonctions uniformément Lipschitz $\alpha > 1$, on obtient le même type de résultat. On peut de même démontrer qu'il n'existe pas de classe $\mathcal{H}$ de fonctions qui puissent battrer l'exposant $\beta = \alpha/d$ de décroissance de l'erreur sur la classe des fonctions Lipschitziennes, au sense où

$$\max_{f \in \mathcal{F}} \min_{\tilde{f} \in \mathcal{H}} \|f - f\|_\infty \leq C \left(\log(|\mathcal{H}|) \right)^{-\beta} \quad \Rightarrow \beta \leq \alpha/d \ .$$

Comme l'exposant α est rarement plus grand que 2 ou 3, on ne gagne pas grand chose. Augmenter l'ordre de l'approximation locale n'est pas suffisant en grande dimension.