Estimation et contrôle des performances en généralisation des réseaux de neurones

Yann Guermeur, Olivier Teytaud

Chapitre 10

Estimation et contrôle des performances en généralisation des réseaux de neurones

Mots-clefs : consistance du processus d'apprentissage, vitesse de convergence du risque empirique, mesures de capacité, résultats asymptotiques.

10.1. Introduction

Le domaine de l'apprentissage automatique en général, et celui des réseaux de neurones en particulier, présente une forte dualité entre théorie et pratique. Le lien unissant ces deux aspects, pourtant parfaitement complémentaires, demeure souvent difficile à cerner. Ainsi, pendant plusieurs décennies, la théorie des bornes, représentée en premier lieu par les travaux de Vapnik [VAP 98], n'a pas eu d'influence notable sur la spécification des algorithmes d'apprentissage des méthodes connexionnistes. Cette situation a considérablement évolué avec l'introduction du principe inductif de minimisation structurelle du risque (MSR) [VAP 82] et sa mise en œuvre dans les machines à vecteurs support (SVM) [BOS 92, COR 95]. Ce progrès installe en effet les bornes sur le risque ou *les risques garantis* au cœur même du dispositif de l'apprentissage. De plus, il est possible de réévaluer a posteriori les algorithmes d'apprentissage standard à la lumière du principe MSR. Une bonne illustration en est donnée dans [BOT 97]. L'auteur souligne le paradoxe que présente l'algorithme de rétropropagation du gradient, le plus utilisé pour entraîner les perceptrons multi-couches (PMC) [BIS 96, ANT 99]. Les bonnes performances qu'il obtient sur les problèmes du monde réel résultent directement de l'inefficacité de la descente en gradient pour

Chapitre rédigé par Yann Guermeur et Olivier Teytaud.

optimiser un dispositif “grossièrement sur-paramétré”. En pratique, celle-ci crée une structure implicite sur l'ensemble des fonctions représentables par le réseau, structure qui est parcourue des classes les plus simples vers les classes les plus riches lorsque les poids sont initialisés avec des valeurs faibles. Le principe inductif MSR vient ainsi au secours de la rétro-propagation du gradient. De nos jours, l'utilité des bornes va donc bien au delà du calcul de la complexité en échantillon ou de l'évaluation des performances en généralisation (qui peut aussi se faire par validation croisée ou "hold-out", comme on l'aura vu au chapitre précédent). Celles-ci apparaissent comme l'outil central en sélection de fonction et sélection de modèle [BAR 99a]. Dans ce chapitre, nous abordons la question des performances en généralisation des réseaux de neurones dans cette perspective large. Le plan est le suivant. La section 10.2 introduit le problème de l'inférence empirique, à travers ses trois composantes que sont l'analyse discriminante, la régression et l'estimation de densité. Les théories statistiques sur lesquelles s'appuient les bornes sont ensuite présentées dans la section 10.3. Les différents paradigmes inférentiels font l'objet de la section 10.4. La section 10.5 présente des avancées récentes. Enfin, la section 10.6, tirant les conclusions de ce qui a été exposé précédemment, évoque les axes de recherche susceptibles de faire progresser significativement le domaine dans les années à venir.

Dans tous le chapitre, dans l'unique but de simplifier l'exposé, on supposera que tout ce qui a besoin d'être mesurable l'est ; on s'abstiendra en particulier de raisonner en "outer expectations" (en toute généralité, une supremum d'une quantité non-dénombrable de variables aléatoires n'est pas nécessairement mesurable). On fera aussi parfois l'abus de parler de l'argument minimum d'une fonction(nelle) même si l'existence et l'unicité ne sont pas systématiques (les résultats sont vrais pour tous les minima lorsqu'il y en a plusieurs, et en cas de non-existence - cas rare - ils sont vrais à ϵ près sur la fonction de coût lorsque l'on considère un argument minimum à ϵ près).

10.2. Position du problème

Dans son acception la plus générale, le problème de l'apprentissage (supervisé) à partir de données consiste à trouver un moyen d'associer à un objet “son” étiquette, ceci en s'appuyant uniquement sur l'information contenue dans un ensemble fini de couples objet-étiquette fourni initialement. Pour que ce problème ait un sens, il faut naturellement qu'il existe un lien entre les objets et les étiquettes. Ce lien est supposé être de nature probabiliste. En pratique, on fait l'hypothèse que les objets, ou plus précisément leurs *descriptions* x et les étiquettes y appartiennent respectivement à des espaces probabilisés $\mathcal{X}$ et $\mathcal{Y}$, et que l'espace produit, $\mathcal{X} \times \mathcal{Y}$, est muni d'une mesure de probabilité P, fixe mais inconnue. L'apprentissage s'appuie alors sur un échantillon de m paires aléatoires (X_i, Y_i) indépendamment et identiquement distribuées (i.i.d.) suivant P. Il consiste à rechercher, dans une famille $\mathcal{H}$ de fonctions donnée, une fonction h^* associant descriptions et étiquettes “avec la plus faible erreur possible”.

Cette reformulation du problème de l'apprentissage comme un problème de sélection de fonction, ou de manière équivalente un problème d'optimisation (précisément, un problème de M-estimation [GEE 00]), appelle naturellement des précisions quant à la nature du critère utilisé, la *fonction objectif*. La notion d'erreur, ou *risque*, n'est pas absolue. Il s'agit d'un critère que l'on se fixe a priori, en fonction du problème traité, et de la nature du co-domaine des fonctions de la classe $\mathcal{H}$, qui n'est pas nécessairement $\mathcal{Y}$. Pour le définir de manière générale, avant de spécifier les choses par type de problème dans les sous-sections suivantes, nous adoptons les conventions utilisées dans [BOU 02]. Celles-ci consistent à définir en premier lieu une *fonction de coût* et une *fonction de perte*. Pour des raisons de simplicité, nous considérons ces fonctions dans le cas particulier où le co-domaine des fonctions de $\mathcal{H}$ est $\mathcal{Y}$. La fonction de coût est une fonction c de $\mathcal{Y} \times \mathcal{Y}$ dans $\mathbb{R}_+$ mesurant, pour une fonction h et une paire (x, y) données, le coût $c(h(x), y)$ associé au fait de prévoir $h(x)$ alors que la véritable sortie est y. La fonction de perte correspondante, notée ℓ, associe h et la paire (x, y) au coût $c(h(x), y)$. c et ℓ sont donc liées par l'équation :

$$\ell(h, (x, y)) = c(h(x), y).$$

Ces définitions étant posées, le risque R se définit comme l'espérance de la fonction de perte :

$$R(h) = \mathbb{E}\left[\ell(h, (X, Y))\right],$$

l'espérance étant calculée par rapport à la distribution de probabilité P de la paire aléatoire (X, Y). Pour aller plus avant, en spécifiant en particulier les fonctions c et ℓ, nous devons à présent distinguer les trois grands types de problèmes que traite l'apprentissage à partir de données :

– la discrimination (à partir d'exemples $x_1, \ldots, x_m$ et de leurs étiquettes discrètes y_1,...,y_m, on cherche à étiqueter de nouveaux exemples x'_1, x'_2,...,x'_p) ;

– la régression (à partir d'exemples $x_1, \ldots, x_m$ et de leurs sorties réeelles $y_1, \ldots, y_m$, on cherche les sorties réelles associées à de nouveaux exemples x'_1, x'_2,...,x'_p) et

– l'estimation de densité (à partir d'exemples x_1,...,x_m on cherche à déterminer la probabilité (ou la densité de probabilité) de nouveaux exemples x'_1,...,x'_p).

10.2.1. *Discrimination*

En discrimination, l'ensemble $\mathcal{Y}$ est un ensemble fini de Q catégories. Le cas des dichotomies ($Q = 2$) est ordinairement traité à part, afin d'exploiter au mieux son caractère dégénéré. Deux situations doivent être distinguées, suivant que les fonctions

de $\mathcal{H}$ prennent leurs valeurs dans $\mathcal{Y}$ ou dans $\mathbb{R}^Q$. Le premier cas est le plus simple. La fonction de perte s'exprime alors généralement comme :

$$\ell\left(h,(X,Y)\right) = \mathbb{1}_{h(X)\neq Y}. \qquad [10.1]$$

Il s'agit tout simplement de la fonction de mauvais classement. On a donc par définition :

$$R(h) = \int_{\mathcal{X}\times\mathcal{Y}} \mathbb{1}_{h(x)\neq y}(x,y)dP(x,y).$$

Le séparateur optimal au sens de ce critère est le *classifieur de Bayes* [DUD 73]. C'est ce classifieur que l'on calculerait si P était connue. La seconde situation, pour laquelle les fonctions de $\mathcal{H}$ s'écrivent $h = (h_k)_{1\leq k\leq Q}$, les fonctions composantes h_k étant à valeurs réelles, est très commune. Elle correspond par exemple au cas où le classifieur considéré estime les probabilités a posteriori des classes. Ceci se produit en particulier avec le perceptron multi-couche, lorsque la fonction d'activation des unités de sortie et la fonction objectif sont convenablement choisies (voir par exemple [RIC 91, BIS 96]). Ce cas se ramène au précédant par application de la règle de décision de Bayes estimée, consistant à affecter à chaque description x la catégorie correspondant à la fonction composante dont la valeur en x est la plus élevée (voir la Figure 10.1 ; ici la classe 1 est choisie).

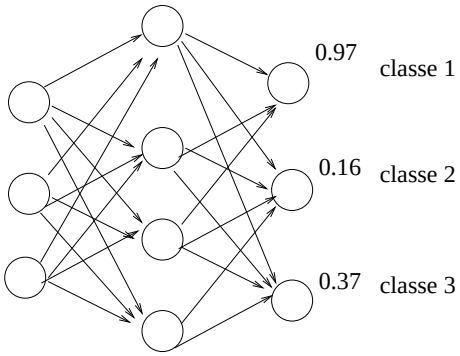

Figure 10.1. *Discrimination multi-classe par perceptron multi-couche.*

Notons que cette règle de décision est ordinairement employée même lorsque les sorties ne sont pas des estimations des probabilités a posteriori des classes. Un bon exemple est fourni par les SVM à catégories multiples, qui seront évoquées dans la suite. En se plaçant du simple point de vue de la polychotomie calculée, et du risque

associé, il n'y a donc pas de différence fondamentale entre les classes de fonctions prenant leurs valeurs dans $\mathcal{Y}$ et celles prenant leurs valeurs dans $\mathbb{R}^{\#\mathcal{Y}}$. Nous avons cependant dès à présent introduit cette distinction afin de préparer les sections suivantes (portant sur les mesures de capacité, le principe inductif MER, les bornes...), qui mettront en évidence le fait que les résultats sur les performances en généralisation sont quant à eux spécifiques.

10.2.2. *Régression*

En régression, l'ensemble $\mathcal{Y}$ est égal à $\mathbb{R}$, qui est également le co-domaine des fonctions de $\mathcal{H}$. La fonction de perte la plus utilisée est le coût quadratique :

$$\ell\left(h,(X,Y)\right) = \left(h(X) - Y\right)^2 .$$

Dans ce cas, la fonction optimale est la *fonction de régression* η définie par :

$$\forall x \in \mathcal{X},\ \eta(x) = \mathbb{E}\left[Y|X=x\right] .$$

De nouveau, il n'est pas possible d'obtenir cette fonction, puisque le calcul de l'espérance s'effectue par rapport à P (qui est inconnue). Il existe un lien direct entre discrimination et régression (multivariée, i.e. à plusieurs variables réelles de sortie) lorsqu'un classifieur estime les probabilités a posteriori des classes. Ce lien peut être mis en évidence lorsque les catégories sont représentées par leur codage binaire canonique, i.e. $\mathcal{Y} = \{y_k; 1 \leq k \leq Q\}$, avec $y_k = (\delta_{k,l})_{1 \leq l \leq Q}$. Dans ce cas,

$$\mathbb{E}\left[\|h(X) - Y\|^2\right] = \sum_{k=1}^{Q} \mathbb{E}\left[(h_k(X) - Y_k)^2\right] .$$

Chaque terme du membre de droite peut être minimisé indépendamment des autres. Cela revient à prendre pour fonctions composantes h_k les fonctions de régression η_k calculant les espérances conditionnelles i.e. $\eta_k(x) = \mathbb{E}\left[Y_k|X=x\right]$, ou de manière équivalente $\eta_k(x) = \mathbb{P}\left[Y_k = 1|X=x\right]$. On observe donc que, pour ce codage des catégories, le classifieur minimisant le coût quadratique est également celui dont le taux d'erreur (par application de la règle de Bayes) est minimal, c'est-à-dire le classifieur de Bayes. Naturellement, le critère des moindres carrés peut être utilisé pour entraîner un système discriminant même lorsque celui-ci ne contient pas le classifieur de Bayes associé au problème considéré. Cela se fera cependant alors au prix d'une *erreur d'approximation* (i.e. quelle que soit la fonction choisie dans la famille donnée, l'erreur ne sera jamais celle de Bayes).

On notera, comme bémol aux liens que l'on vient de citer entre la régression et la classification, le fait que la classification est un problème plus simple que celui de l'estimation de probabilités conditionnelles. Ce dernier, comme la régression, conduit en général (sauf cas très dégénérés) à des vitesses de convergence en $1/\sqrt{m}$; la classification, elle, selon les cas, conduit à des vitesses en $1/\sqrt{m}$ ou $1/m$. La détermination précise de la séparation entre ces deux catégories de classifications est centrale dans les avancées récentes. Si le cas d'une erreur optimale nulle est depuis longtemps connu comme entraînant une vitesse en $1/m$, des résultats plus récents placent de nombreux autres cas dans la même catégorie. Nous les évoquerons brièvement de manière synthétique en section 10.5.2. Nous allons maintenant nous intéresser à plus difficile encore, l'estimation de densité.

10.2.3. *Estimation de densité*

L'estimation de densité est certainement le plus difficile et le moins étudié des trois problèmes considérés par la théorie statistique de l'apprentissage. Il correspond au cas d'une dépendance fonctionnelle entre les x et les y donnée par $y = p(x)$ avec $\int_{\mathcal{X}} p(x)dx = 1$. Il est ordinairement traité, par exemple par Vapnik, dans le cadre restreint où la famille de fonctions considérée est une classe de densités $p(., \alpha)$, α étant un paramètre formel prenant ses valeurs dans un ensemble Λ, par application du principe de *maximum de vraisemblance*. Dans ce cas, le risque associé à la densité $p(., \alpha)$ est donné par :

$$R\left(p(., \alpha)\right) = -\int_{\mathcal{X}} \ln p(x, \alpha) dP(x) = -\int_{\mathcal{X}} \ln p(x, \alpha) p(x) dx.$$

Il "contient" les autres problèmes au sens où la connaissance de la densité de (x, y) détermine exactement la prévision p optimisant la moyenne $\mathbb{E}c(x, p, Y)|X = x$, conditionnelle à un x donné, de la fonction de coût.

D'autres méthodes ont été proposées. L'estimation de densité peut ainsi se faire :

– en approchant p par la convoluée de la distribution empirique (typiquement, convolution par une gaussienne) ; c'est-à-dire que l'on approche $p(x)$ par $\frac{1}{m}\sum_{i=1}^{m} \exp(-||x - x_i||^2/\sigma^2)$ où les x_i sont un échantillon de m points aléatoire simple de distribution p. On montre alors des propriétés de consistance pour peu que l'écart-type σ décroisse de manière appropriée en fonction du nombre d'exemples ;

– par des méthodes adaptées de l'algorithme des plus proches voisins (consistant à approcher la densité par le ratio k/v où v est le volume d'une boule incluant les k plus proches voisins) ou de la méthode des histogrammes ;

– par des méthodes de clustering (classification non-supervisée), segmentant le domaine, et proposer ainsi une partition du domaine en E_1,...,E_d et des estimations des $P(X \in E_i)$ (voir par exemple les cartes somatotopiques de Kohonen [KOH 89], présentées au chapitre intitulé "Cartes auto-organisatrices de Kohonen").

– en utilisant des hypothèses a priori, comme c'est le cas avec la méthode bayésienne naïve, consistant à remplacer $P(X = x)$ par le produit $P(X_1 = x_1) \times \ldots \times P(X_d = x_d)$, c'est-à-dire à considérer les différents attributs comme indépendants conditionnellement à la classe. D'autres façons de faire consistent à considérer des indépendances conditionnelles plus restreintes ; on aborde ainsi le très riche domaine des réseaux bayésiens [NAI 04].

On pourra consulter [DEV 87] pour plus d'informations sur ce domaine.

10.3. Les théories statistiques

Cette section comporte :

– une rubrique 10.3.1 "outils statistiques de base" présentant divers outils utiles pour la suite ;

– une rubrique 10.3.2 "mesures de capacité", présentant les critères classiques de mesure de la complexité d'une famille de fonctions ;

– une rubrique 10.3.3 expliquant comment évaluer ces mesures de capacité ;

– une rubrique 10.3.4 enfin exposant les résultats asymptotiques les plus classiques (classes de Glivenko-Cantelli et de Donsker, bootstrap).

10.3.1. *Les outils statistiques de base*

Nous verrons dans les sections qui suivent :

– les différentes formes de convergences statistiques ;

– quelques outils destinés à quantifier l'écart entre une moyenne et une espérance ;

– quelques éléments sur les processus gaussiens ;

– enfin quelques outils complémentaires.

10.3.1.1. *Convergences statistiques*

La notion de limite d'une suite est usuelle, la limite d'une suite de variables aléatoires est plus complexe. Nous définissons ici les types de convergences de telles suites dont nous aurons besoin plus tard.

DÉFINITION 1 (Convergence en probabilité) *Une suite de variables aléatoires réelles $(X_n)_{n\in\mathbb{N}}$ définies sur $(\Omega, \mathcal{B}, P)$ converge* en probabilité, *ou* en mesure *ou* stochastiquement *vers la variable aléatoire réelle X, définie sur $(\Omega, \mathcal{B}, P)$, si pour tout*

$\epsilon > 0$,

$$\lim_{n \longrightarrow \infty} P\{|X_n - X| \geq \epsilon\} = 0.$$

On le note $X_n \overset{P}{\rightarrow} X$.

DÉFINITION 2 (Convergence presque sûre) *Une suite de variables aléatoires réelles* $(X_n)_{n \in \mathbb{N}}$ *définies sur* $(\Omega, \mathcal{B}, P)$ *converge* presque sûrement (p.s.) *ou* en probabilité 1 *vers la variable aléatoire réelle* X*, définie sur* $(\Omega, \mathcal{B}, P)$*, si*

$$P\left\{\omega \in \Omega \ : \lim_{n \longrightarrow \infty} X_n(\omega) = X(\omega)\right\} = 1.$$

On le note $X_n \overset{p.s.}{\rightarrow} X$.

La convergence presque sûre implique la convergence en probabilité. Pour s'en convaincre, il suffit de reformuler la convergence presque sûre sous la forme équivalente suivante :

$$\forall \epsilon > 0, \quad \lim_{m \longrightarrow \infty} P\left\{\sup_{n \geq m} |X_n - X| \geq \epsilon\right\} = 0.$$

Une convergence plus faible et très utile est la convergence faible.

DÉFINITION 3 (Convergence en loi) *Une suite de variables aléatoires réelles* $(X_n)_{n \in \mathbb{N}}$ *définies sur* $(\Omega, \mathcal{B}, P)$ *converge* en loi *ou* faiblement *ou* en distribution *vers la variable aléatoire réelle* X*, définie sur* $(\Omega, \mathcal{B}, P)$*, si*

$$\lim_{n \longrightarrow \infty} F_{X_n}(t) = F_X(t)$$

en tout point de continuité t *de* F_X*. On le note* $X_n \overset{\mathcal{L}}{\rightarrow} X$ *ou* $X_n \overset{d}{\rightarrow} X$.

La section V.4.1. de [BAR 98a] liste plusieurs conditions équivalentes permettant de caractériser la convergence en loi. Parmi celles-ci, la suivante revêt une importance particulière.

PROPOSITION 1 (Caractérisation de la convergence en loi) *Une condition nécessaire et suffisante de convergence en loi est la suivante :*

$$\lim_{n \longrightarrow \infty} \int \phi(X_n) dP = \int \phi(X) dP$$

pour toute fonction continue bornée $\phi : \mathbb{R} \longrightarrow \mathbb{R}$.

La convergence en loi est le plus faible des trois types de convergence que nous avons définis ici. Elle est en effet impliquée par la convergence en probabilité (voir par exemple le chapitre V de [BAR 98a] pour une démonstration). La littérature considère encore le cas de la convergence dans L^p, c'est-à-dire la convergence topologique de l'espace $L^p(\Omega, \mathcal{B}, P)$ muni de la norme $\|.\|_p$. Cette convergence joue un rôle central dans le cas des variables aléatoires à valeurs dans des espaces de Banach [LED 91]. Elle implique la convergence en probabilité qui peut alors être considérée comme une convergence dans L^0.

10.3.1.2. *Ecart entre une moyenne et une espérance*

Quantifier l'écart entre une moyenne et une espérance est un problème très important de statistique comme on peut s'en douter si l'on voit l'apprentissage comme la minimisation (en h) de la moyenne $\hat{\mathbb{E}}\ell(h,(X,Y))$ avec pour objectif la minimisation de l'espérance $\mathbb{E}\ell(h,(X,Y))$, et nous commencerons donc par fournir une liste de résultats de ce type régulièrement utiles pour la théorie de l'apprentissage. Pour ce faire, nous nous appuyons en particulier sur [BAR 98a, LUG 04].

Tout d'abord, quelques inégalités classiques valables pour toute variable aléatoire réelle X (admettant les moments appropriés), qui permettent de dériver les inégalités suivantes, et sont de manière générale utiles en statistique. A partir de l'équation suivante, valable pour les variables aléatoires à valeurs positives :

$$\mathbb{E}X = \int_0^\infty P(X \geq t)dt$$

on déduit immédiatement l'inégalité de Markov, soit, pour toute variable aléatoire à valeurs positives et t supérieur à 0,

$$P(X \geq t) \leq \frac{\mathbb{E}X}{t}. \qquad [10.2]$$

En remplaçant dans l'inégalité de Markov X par $|X - \mathbb{E}X|^2$, on obtient l'inégalité de Bienaymé-Tchebytchev :

$$P(|X - \mathbb{E}X| \geq t) \leq \frac{Var(X)}{t^2}.$$

Cette inégalité se généralise, pour $q > 0$, en

$$P(|X - \mathbb{E}X| \geq t) \leq \frac{\mathbb{E}\left[|X - \mathbb{E}X|^q\right]}{t^q}.$$

Les inégalités qui suivent concernent, sauf mention explicite du contraire, une suite de variables aléatoires $X_1, \ldots, X_n$ indépendantes et d'espérance nulle. On s'intéresse au fait que $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ soit proche de son espérance. L'inégalité de Bienaymé-Tchebytchev donnée ci-dessous, conséquence directe de l'inégalité de Bienaymé (voir par exemple [BAR 98a]) et de 10.3, est bien connue et s'applique sans que les données soient identiquement distribuées.

THÉORÈME 1 (Inégalité de Bienaymé-Tchebytchev) *Si $X_1, \ldots, X_n$ sont deux-à-deux non corrélées, alors pour tout t strictement positif,*

$$P\left\{\left|\sum_{i=1}^{n}(X_i - \mathbb{E}X_i)\right| \geq t\right\} \leq \frac{1}{t^2}\sum_{i=1}^{n} Var(X_i). \qquad [10.3]$$

Cette simple inégalité est à la base de la loi faible des grands nombres, due à Bernoulli :

THÉORÈME 2 (Loi faible des grands nombres) *Soit X une variable aléatoire à valeurs réelles vérifiant $\mathbb{E}|X| < \infty$. Soit μ son espérance. Soit $X_1, \ldots, X_n$ un ensemble de n copies indépendantes de X. Alors*

$$\frac{1}{n}\sum_{i=1}^{n} X_i \overset{P}{\rightarrow} \mu.$$

Le même résultat est vrai pour la convergence dans L^1. Il est à noter que l'inégalité de Bienaymé-Tchebytchev 10.3 n'entraîne rapidement la loi faible des grands nombres que dans le cas où X possède une variance finie. Nous avons présenté ici une version plus générale, dont le lecteur pourra trouver une démonstration dans [TOU 99].

Après Laplace et Poisson [POI 35], Kolmogorov et Khintchine [KHI 28] ont obtenu une version plus forte de la loi des grands nombres.

THÉORÈME 3 (Loi forte des grands nombres) *Soit X une variable aléatoire à valeurs réelles. Soit $X_1, \ldots, X_n$ un ensemble de n copies indépendantes de X. Alors*

$$\frac{1}{n}\sum_{i=1}^{n} X_i \text{ converge p.s.} \Longleftrightarrow \mathbb{E}|X| < \infty.$$

Lorsqu'il y a convergence, la limite est $\mathbb{E}X$.

Cette convergence est donnée sans indication de vitesse. Pour en obtenir une, nous faisons appel au célèbre théorème de la limite centrale, établi ou raffiné sous différentes formes par De Moivre, Stirling, Gauss, Laplace, Tchebytchev, Markov et Lyapunov.

THÉORÈME 4 (Théorème de la limite centrale multivarié [VAA 98]) *Soit X un vecteur aléatoire à valeurs dans $\mathbb{R}^d$ admettant des moments d'ordre 1 et 2. Notons μ son espérance et $\Sigma = \mathbb{E}\left\{(X-\mu)(X-\mu)^T\right\}$ sa matrice de variance-covariance. Soit $X_1, \ldots, X_n$ un ensemble de n copies indépendantes de X. Alors*

$$\frac{1}{\sqrt{n}}\left(\sum_{i=1}^{n} X_i - n\mu\right) \overset{\mathcal{L}}{\rightarrow} \mathcal{N}(0,\Sigma)$$

où $\mathcal{N}(0,\Sigma)$ est la loi normale centrée de matrice de variance-covariance Σ.

En se restreignant au cas des variables aléatoires à valeurs réelles, la vitesse de convergence est caractérisée par la formule suivante

$$P\left\{\sqrt{\frac{n}{\sigma^2}}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right) \geq t\right\} \rightarrow \frac{1}{\sqrt{2\pi}}\int_t^{+\infty} e^{-\frac{u^2}{2}}du$$

$$\leq \frac{1}{\sqrt{2\pi}}\frac{e^{-\frac{t^2}{2}}}{t}. \qquad [10.4]$$

pour $n \rightarrow \infty$. L'inégalité de droite (borne supérieure sur la "fonction Q") s'obtient facilement par intégration par parties. La loi du logarithme itéré (voir par exemple [GIR 01]) vient compléter cette borne.

THÉORÈME 5 (Loi du logarithme itéré) *Soit X une variable aléatoire à valeurs réelles de carré intégrable et σ^2 sa variance. Soit $(X_i)_{i\in\mathbb{N}}$ une suite de copies indépendantes de X. Alors*

$$\limsup_{n\longrightarrow\infty} \frac{\sqrt{\frac{n}{\sigma^2}}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right)}{\sqrt{2\log\log(n)}} = 1 \qquad [10.5]$$

et

$$\liminf_{n\longrightarrow\infty} \frac{\sqrt{\frac{n}{\sigma^2}}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right)}{\sqrt{2\log\log(n)}} = -1. \qquad [10.6]$$

On verra avec le théorème 12 une version uniforme directement applicable en apprentissage de la loi du logarithme itéré.

En ayant à l'esprit 10.4, on peut rapidement se convaincre que le taux de convergence donné par l'inégalité de Bienaymé-Tchebytchev (théorème 1) n'est pas satisfaisant. Ce taux s'exprime en effet comme $P(\frac{1}{n}\left|\sum_{i=1}^{n}\left(X_i - \mathbb{E}X_i\right)\right| > t) = O(1/t^2)$, alors que 10.4 promet une décroissance exponentielle. Afin d'obtenir des bornes de meilleure qualité, on peut s'appuyer sur la méthode de Chernoff, reposant sur l'inégalité suivante.

THÉORÈME 6 (Inégalité de Chernoff) *Soit X une variable aléatoire à valeurs réelles et s et t deux réels strictement positifs. Alors,*

$$P\left\{X \geq t\right\} = P\left\{e^{sX} \geq e^{st}\right\} \leq \frac{\mathbb{E}e^{sX}}{e^{st}}. \qquad [10.7]$$

L'inégalité de Chernoff permet en particulier de dériver l'inégalité de Hoeffding (voir par exemple [POL 84, LUG 04]).

THÉORÈME 7 (Inégalité de Hoeffding [HOE 63]) *Soit $X_1, \ldots, X_n$ un ensemble de n variables aléatoires à valeurs réelles, indépendantes et centrées, prenant leurs valeurs dans $[a_i, b_i]$. Alors, pour tout $t > 0$,*

$$P\left\{\sum_{i=1}^{n} X_i > t\right\} \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right). \qquad [10.8]$$

Lorsque l'on dispose d'une information sur les variances des variables aléatoires considérées, il est possible d'utiliser d'autres résultats, en particulier les inégalités de Bennett et de Bernstein.

THÉORÈME 8 (Inégalité de Bennett [BEN 62]) *Soit $X_1, \ldots, X_n$ un ensemble de n variables aléatoires à valeurs réelles, indépendantes et centrées. On fait l'hypothèse que $|X_i| \leq c$ avec une probabilité 1 et l'on note $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} Var(X_i)$. Alors, pour tout $t > 0$,*

$$P\left\{\sum_{i=1}^{n} X_i > t\right\} \leq \exp\left(-\frac{n\sigma^2}{c^2} h\left(\frac{ct}{n\sigma^2}\right)\right), \qquad [10.9]$$

où la fonction h est donnée par $h(u) = (1+u)\log(1+u) - u$ pour u positif ou nul.

THÉORÈME 9 (Inégalité de Bernstein [BER 46]) *Sous les hypothèses de l'inégalité de Bennett, on a, pour tout $t > 0$,*

$$P\left\{\frac{1}{n}\sum_{i=1}^{n} X_i > t\right\} \leq \exp\left(-\frac{nt^2}{2\sigma^2 + 2ct/3}\right). \qquad [10.10]$$

Il convient de noter que, si nous avons donné de l'inégalité de Bernstein la référence standard, Pollard suggère dans l'annexe B de [POL 84] que ce résultat pourrait en fait être bien antérieur. Les inégalités de Bennett et de Bernstein permettent notamment de dériver des inégalités beaucoup plus rapides (erreur en $O(1/m)$) lorsque l'erreur minimale est nulle (alors que l'on obtient des bornes de l'ordre $O(1/\sqrt{m})$ dans le cas général). Ces résultats d'accélération (convergence de l'erreur en $1/m$ au lieu de $1/\sqrt{m}$), que l'on pourra trouver en particulier dans [DEV 96, Problem 12.9] sont meilleurs que ceux obtenus directement (voir par exemple [VAP 98, section 4.2]).

L'inégalité de concentration jouant actuellement le rôle le plus important dans la dérivation des bornes est sans doute l'inégalité des différences bornées.

THÉORÈME 10 (Inégalité des différences bornées [MCD 89]) *Soit g une fonction de n variables de $\mathcal{X}^n$ dans $\mathbb{R}$ tel qu'il existe des constantes $c_1, \ldots, c_n$ positives ou nulles satisfaisant :*

$$\sup_{(x_i)_{1\leq i\leq n}\in\mathcal{X}^n, x_i'\in\mathcal{X}} |g(x_1, \ldots, x_n) - g(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n)| \leq c_i, \qquad [10.11]$$

pour $1 \leq i \leq n$. Alors, si $X_1, \ldots, X_n$ sont des variables aléatoires à valeurs dans $\mathcal{X}$ indépendantes, la variable aléatoire $Z = g(X_1, \ldots, X_n)$ satisfait :

$$P\{|Z - \mathbb{E}Z| > t\} \leq 2\exp\left(-\frac{2t^2}{C}\right), \qquad [10.12]$$

où $C = \sum_{i=1}^{n} c_i^2$.

Dans cette sous-section et la précédente, afin de simplifier l'exposé, nous avons principalement considéré le cas de variables aléatoires à valeurs réelles. Cependant, l'essentiel du contenu s'étend à des cas plus généraux, en particulier celui des variables à valeurs dans des espaces de Banach, traité en profondeur dans [LED 91].

10.3.1.3. *Processus gaussiens*

DÉFINITION 4 (Processus gaussien (voir par exemple [VAA 96])) *Un* processus gaussien *est l'extension naturelle d'une gaussienne en dimension infinie. Précisément, un processus stochastique X à valeurs dans $\mathbb{R}^T$ pour un certain ensemble T est dit* gaussien *si chacune de ses lois marginales de dimension finie, $(X_{t_1}, \ldots, X_{t_k})$, suit une loi normale multivariée sur un espace euclidien.*

L'étude des processus gaussiens est fondamentale car le vecteur (infini) $(\hat{\mathbb{E}}f - \mathbb{E}f)_{f\in\mathcal{H}}$ tend faiblement vers un processus gaussien dans un grand nombre de cas. Le résultat suivant quant au supremum en $t \in T$ d'un processus gaussien a été prouvé initialement par Borell et indépendamment par Cirelson, Ibragimov, Sudakov ; voir [ADL 90].

PROPOSITION 2 (Inégalité de Borell) *Soit $(Y_t)_{t\in T}$ un processus gaussien borné presque sûrement et $\sigma(Y)^2 = \sup_{t\in T} Var(Y_t)$, alors*

$$\lim_{s\to\infty} \frac{\log P(\sup_t Y_t > s)}{s^2} = -1/(2\sigma(Y)^2)$$

et pour tout $\epsilon > 0$, pour s suffisamment grand, on a

$$P(\sup_t Y_t > s) \leq \exp\left(\epsilon s^2 - \frac{1}{2}s^2/\sigma(Y)^2\right).$$

Interprétation : asymptotiquement, la plus grande déviation par rapport à 0 dans un processus gaussien a lieu essentiellement pour le t conduisant à la variance marginale la plus forte, et la déviation maximale est du même ordre que la déviation que l'on obtient avec juste Y_t pour t maximisant la variance de Y_t. Ce point est très surprenant, et ne doit pas cacher la réalité des faits : cela n'est vrai que de manière asymptotique (on regarde la probabilité de dépasser une déviation donnée, à la limite d'une très grande déviation). Cette impression selon laquelle l'essentiel est la variance maximale du processus gaussien est en fait clairement fausse pour des critères plus réalistes. En particulier, en anticipant un peu sur la suite de ce chapitre, il est clair que la déviation maximale obtenue sur une famille de fonctions est nettement plus grande dans l'immense majorité des cas que la déviation maximale obtenue pour une fonction, quelle que soit cette fonction.

10.3.1.4. *Divers*

Un outil important en théorie de l'apprentissage est le lemme de Borel-Cantelli, qui fournit des conditions sous lesquelles la convergence en probabilité implique la convergence presque sûre.

LEMME 1 (Lemme de Borel-Cantelli) *Soit* $(X_n)_{n\in\mathbb{N}}$ *une suite de variables aléatoires réelles définies sur* $(\Omega, \mathcal{B}, P)$.

1) Si pour tout $\epsilon > 0$, $\sum_{n\in\mathbb{N}} P\{|X_n - X| \geq \epsilon\} < \infty$, *alors* $X_n \overset{p.s.}{\to} X$.

2) Si les $(X_n)_{n\in\mathbb{N}}$ *sont mutuellement indépendantes, alors* $X_n \overset{p.s.}{\to} 0$ *si et seulement si* $\sum_{n\in\mathbb{N}} P\{|X_n| \geq \epsilon\} < \infty$ *pour tout* $\epsilon > 0$.

Interprétation : si une convergence en probabilité est "suffisamment" rapide, alors elle est en fait presque sûre.

Différentes lois du logarithme itéré trouveront une utilité dans la suite.

Soit L la fonction définie par $L(t) = \max(1, \log(t))$, pour t réel positif ou nul, et LL, la fonction du logarithme itéré, définie, toujours pour $t \geq 0$, par $LL(t) = L(L(t))$.

THÉORÈME 11 (Loi du logarithme itéré de Kolmogorov [LED 91]) *Soit* $(X_i)_{i\in\mathbb{N}}$ *une suite de variables aléatoires réelles indépendantes et d'espérance nulle, telle que l'on ait, pour tout entier naturel* i, $\mathbb{E}X_i^2 < \infty$. *Posons, pour tout* n, $s_n = \left(\sum_{i=1}^n \mathbb{E}X_i^2\right)^{1/2}$. *On fait l'hypothèse que la suite* $(s_n)_{n\in\mathbb{N}}$ *croît vers l'infini et qu'il existe une suite* $(\eta_i)_{i\in\mathbb{N}}$ *de réels strictement positifs tendant vers 0 telle que, pour tout* i,

$$\|X_i\|_\infty \leq \frac{\eta_i s_i}{\left(LL\left(s_i^2\right)\right)^{1/2}}.$$

Alors, la loi du logarithme itéré de Kolmogorov *exprime le fait que, avec une probabilité* 1,

$$\limsup_{n\longrightarrow\infty} \frac{\sum_{i=1}^n X_i}{\left(2s_n^2 LL\left(s_n^2\right)\right)^{1/2}} = 1.$$

Cette loi s'étend au cas uniforme sur un espace de fonctions, comme les classes de Glivenko-Cantelli généralisent la loi des grands nombres et comme les classes de Donsker généralisent le théorème de la limite centrale.

THÉORÈME 12 (Loi du logarithme itéré [CHU 49]) *Soit* $x_1, \ldots, x_m$ *i.i.d. de loi uniforme dans* $[0,1]^d$. *Alors*

$$\limsup \sqrt{2m} \frac{D_m^*}{\sqrt{\log\log(m)}} = 1 \text{ p.s.} \qquad [10.13]$$

où $\limsup u_n = \lim_{n\to\infty} \sup_{m\geq n} u_m$ *et où* D_m^* *est la* discrépance *de l'ensemble de points* $x_1, \ldots, x_m$*, définie par*

$$D_m^* = \sup_{r\in[0,1]^d} \mid |\{i/x_i < r\}|/m - r_1 \times r_2 \times \ldots \times r_d \mid .$$

Une version non-asymptotique existe aussi, que l'on trouvera dans [KIE 61] :

THÉORÈME 13 (Borne non-asymptotique sur la discrépance) *Soit* $x_1, \ldots, x_m$ *i.i.d. de loi uniforme dans* $[0,1]^d$*. Alors pour tout* $\epsilon > 0$*, il existe une constante* $c > 0$ *dépendant de* ϵ *et de* d *seulement telle que :*

$$u \geq 0 \Rightarrow P(\sqrt{m} D_m^* \leq u) \geq 1 - c\exp(-(2-\epsilon)u^2).$$

Interprétation : la proportion de points dans chaque hyper-rectangle tend, assez vite (à une précision $O(1/\sqrt{m})$ près), vers la surface associée. Il s'agit donc là d'un théorème similaire aux résultats sur les classes de Donsker, à ceci près que l'on travaille en $limsup$ plutôt qu'en convergence faible. Examinons le cas de l'apprentissage de relation déterministe : supposons x de loi P uniforme sur $[0,1]$, $y = g(x)$ avec g déterministe et redéfinissons par léger abus de notation $L_h(x) = L_h(x, g(x))$, $L_{\mathcal{H}} = \{L_h; h \in \mathcal{H}\}$. On peut alors raisonner comme suit :

– *la loi du logarithme itéré* permet de majorer D_m^* (asymptotiquement) en fonction de m ;

– *l'inégalité de Koksma-Hlawka* (voir le chapitre 2 de [NIE 92]) permet de borner l'écart entre l'espérance et la moyenne de L_h en fonction du produit de D_m^* et de $V(L_h)$ la variation totale au sens de Hardy et Krause[1] de L_h :

$$|\hat{\mathbb{E}} L_h - \mathbb{E} L_h| \leq V(L_h) D_m^*$$

(pour tout L_h et dans tous les cas ; il n'y a pas ici de probabilité !) cf e.g. [NIE 92] pour plus d'informations.

Ce résultat est un résultat en soi ; il permet de borner l'erreur en généralisation, à partir d'hypothèses sur la variation de Hardy et Krause ; et même sans chiffrer la variation de Hardy et Krause, ce résultat fournit une vitesse de convergence. Il permet

1. La définition de la variation totale est longue et complexe ; elle généralise le cas de la dimension 1 $V(h) = \sup_{p\in\mathbb{N}} \sup_{x_1\leq x_2\leq\ldots\leq x_p} \sum_{i=1}^{p-1} |h(x_{i+1}) - h(x_i)|$; faire l'hypothèse de la finitude de cette variation n'est ni irréaliste ni anodin.

aussi de quantifier la qualité d'un jeu de point, la discrépance apparaissant comme un critère naturel de qualité d'un ensemble de points. Autre chose, plus étonnant, est possible. Si l'on positionne les points de manière non-aléatoire (i.e. les x_i sont maintenant déterministes), alors D_m^* peut être meilleur qu'avec une suite i.i.d. En particulier, on peut utiliser comme x_i une *suite à faible discrépance*, garantissant D_m^* majoré par $c(d)\log(m)^d/m$ où $c(d)$ dépend de la dimension seulement. Ces techniques sont une version formalisée des techniques consistant à bien choisir le positionnement des points : plan d'expérience, apprentissage actif. Cette méthode présente la limitation d'être restreinte au cas hors-ligne ; réduire la discrépance est indépendant des valeurs de la fonction en les points déjà considérés.

On passe ainsi de l'équation suivante dans le cas de points x_i aléatoires simples :

$$|\hat{\mathbb{E}}L_h - \mathbb{E}L_h| \leq V(L_h)D_m^*(x_1, \ldots, x_m)$$

dont le comportement asymptotique est majoré (au sens de la $\limsup$, via la loi du logarithme itéré) par

$$V(L_h)\sqrt{\log(\log(m))/2m}$$

à l'équation suivante pour des x_i placés adéquatement (suite à faible discrépance) :

$$|\hat{\mathbb{E}}L_h - \mathbb{E}L_h| \leq V(L_h)D_m^*(x_1, \ldots, x_m)$$

qui est majorée, via un meilleur majorant de D_m^*, par

$$V(L_h)c(d)\log(m)^d/m$$

où $c(d)$ dépend de la dimension seulement. Le lecteur intéressé est renvoyé à [NIE 92, CER 04].

Interprétation : On voit ici la vitesse accrue consécutive au remplacement d'un ensemble aléatoire de points par un ensemble déterministe lorsque la relation est déterministe. On peut être déçu par le caractère exponentiel en la dimension de la discrépance des suites classiques précédemment citées (i.e., le terme $\log(m)^d$). Les résultats existants sont tout à fait étonnants : des preuves d'existence de suites sans le terme exponentiel existent, et il ne reste plus que le terme en $1/m^\alpha$. Malheureusement, là où les suites aléatoires atteignent $\alpha = \frac{1}{2}$, ces suites n'atteignent que α de l'ordre de 0.4 ; des résultats existent avec α de l'ordre de 0.7 mais malheureusement ils sont non-constructifs et la suite réalisant cette performance n'est pas connue ([WAS 97]).

10.3.2. *Mesures de capacité*

Nous verrons dans les sections suivantes que la théorie de l'apprentissage définit trois critères pour caractériser le succès de la mise en œuvre d'un principe inductif : les convergences faible et forte, ainsi que la propriété de consistance non triviale. Sous

les hypothèses standard concernant la nature du problème traité, l'existence de ces convergences repose entièrement sur la finitude de certaines *mesures de capacité* de la classe de fonctions utilisée. La notion de mesure de capacité des familles de fonctions joue donc en apprentissage un rôle central. Non seulement la finitude d'une mesure de capacité apparaît comme une condition suffisante (et bien souvent nécessaire) pour pouvoir apprendre, mais la valeur de cette mesure établit également la vitesse (au sens de la complexité en échantillon) avec laquelle l'apprentissage se fera. Cette section est consacrée aux principales mesures de capacité proposées dans la littérature et possédant un intérêt pratique (pouvant être majorées). Nous commençons avec celles ayant joué historiquement le rôle le plus important, en apparaissant au cœur du premier résultat de convergence uniforme du risque empirique vers le vrai risque [VAP 71], les *mesures de capacité globales*. Nous citons en partie 10.5 des travaux récents tirant plus parti de l'échantillon pour borner l'erreur en généralisation.

Les mesures de capacité globales sont des mesures faisant intervenir l'ensemble des fonctions de la classe considérée, ceci sans tenir compte du processus d'apprentissage. Parmi ces mesures, celle sur laquelle s'appuient les résultats les plus fondamentaux est la dimension de Vapnik-Chervonenkis. Afin de la définir, nous devons introduire au préalable la notion de fonction de croissance.

DÉFINITION 5 (Fonction de croissance [VAP 71]) *Soit $\mathcal{H}$ une famille de fonctions à valeurs binaires. Soit $\mathcal{H}|_{\mathcal{X}_m}$ sa restriction à un sous-ensemble $\mathcal{X}_m$ de $\mathcal{X}$ de cardinalité m. Soit $\Pi_{\mathcal{H}}(\mathcal{X}_m)$ le cardinal de $\mathcal{H}|_{\mathcal{X}_m}$. La* fonction de croissance *de $\mathcal{H}$ est la fonction de $\mathbb{N}^*$ dans $\mathbb{N}$ définie par :*

$$\forall m \in \mathbb{N}^*,\ \Pi_{\mathcal{H}}(m) = \max_{\mathcal{X}_m \in \mathcal{X}^m} \Pi_{\mathcal{H}}(\mathcal{X}_m). \qquad [10.14]$$

Nous donnons cette définition, avec son abus de notation, comme étant la plus communément employée. Il convient de remarquer que dans ses écrits les plus récents, Vapnik définit la fonction de croissance comme étant le logarithme de la fonction $\Pi_{\mathcal{H}}$.

DÉFINITION 6 (Dimension de Vapnik-Chervonenkis [VAP 71]) *Soit $\mathcal{H}$ une famille de fonctions à valeurs binaires. La* dimension de Vapnik-Chervonenkis *de $\mathcal{H}$, VC-dim($\mathcal{H}$), est la taille du plus grand sous-ensemble de $\mathcal{X}$* pulvérisé *par $\mathcal{H}$, c'est-à-dire sur lequel $\mathcal{H}$ calcule toutes les dichotomies possibles, si cette taille est finie, l'infini dans le cas contraire. Lorsque la dimension VC est finie, on a donc, en posant $d =$ VC-dim($\mathcal{H}$) :*

$$\Pi_{\mathcal{H}}(d) = 2^d,\ \Pi_{\mathcal{H}}(d+1) < 2^{d+1}. \qquad [10.15]$$

Si la fonction de croissance apparaît dans les bornes de manière plus directe que la dimension VC, cette dernière peut généralement être plus aisément bornée. C'est la raison pour laquelle on fait souvent usage d'un lien simple entre les deux mesures, lien connu sous le nom de "lemme de Sauer".

LEMME 2 (Lemme de Sauer [VAP 71, SAU 72, SHE 72]) *Soit $\mathcal{H}$ une famille de fonctions à valeurs binaires de dimension VC finie $d \geq 1$. Alors, pour tout entier $m \geq 1$, on a :*

$$\Pi_{\mathcal{H}}(m) \leq \sum_{i=0}^{d} C_m^i \qquad [10.16]$$

avec de plus, pour $m \geq d$

$$\sum_{i=0}^{d} C_m^i < \left(\frac{em}{d}\right)^d . \qquad [10.17]$$

Interprétation : si la dimension VC est finie, alors au-delà de cette valeur, la fonction de croissance cesse de croître exponentiellement avec le nombre de points, pour ne plus croître que polynomialement.

La littérature propose deux types d'extensions de cette mesure de capacité. Les unes concernent les classes de fonctions à valeurs réelles, les autres les classes de fonctions à valeurs dans un ensemble fini. Nous les introduisons à présent en respectant une progression dans l'adéquation au calcul de bornes.

DÉFINITION 7 (Dimension de Vapnik [VAP 89]) *Soit $\mathcal{H}$ une famille de fonctions d'un domaine $\mathcal{X}$ dans $\mathbb{R}$. Un sous-ensemble $\mathcal{X}_m = \{x_i\}$, $(1 \leq i \leq m)$ de $\mathcal{X}$ est dit être V-*pulvérisé *par $\mathcal{H}$ s'il existe un scalaire b tel que, pour tout vecteur binaire $v_y = (y_i) \in \{-1, 1\}^m$, il existe une fonction $h_y \in \mathcal{H}$ satisfaisant*

$$\forall i \in \{1, \ldots, m\}, \begin{cases} h_y(x_i) - b \geq 0 & \textit{si} \quad y_i = 1 \\ h_y(x_i) - b < 0 & \textit{si} \quad y_i = -1 \end{cases}$$

La dimension de Vapnik *de $\mathcal{H}$, V-dim$(\mathcal{H})$, est le cardinal du plus grand des sous-ensembles de $\mathcal{X}$ V-pulvérisé par $\mathcal{H}$, si ce cardinal est fini, l'infini sinon.*

La dimension de Vapnik est une version uniforme de la pseudo-dimension de Pollard.

DÉFINITION 8 (Pseudo-dimension de Pollard [POL 90, HAU 92]) *Soit $\mathcal{H}$ une famille de fonctions d'un domaine $\mathcal{X}$ dans $\mathbb{R}$. Un sous-ensemble $\mathcal{X}_m = \{x_i\}$, $(1 \leq i \leq m)$ de $\mathcal{X}$ est dit être* P-pulvérisé *par $\mathcal{H}$ s'il existe un vecteur $v_b = (b_i) \in \mathbb{R}^m$ tel que, pour tout vecteur binaire $v_y = (y_i) \in \{-1, 1\}^m$, il existe une fonction $h_y \in \mathcal{H}$ satisfaisant*

$$\forall i \in \{1, \ldots, m\}, \begin{cases} h_y(x_i) - b_i \geq 0 & \textit{si} \quad y_i = 1 \\ h_y(x_i) - b_i < 0 & \textit{si} \quad y_i = -1 \end{cases}$$

La pseudo-dimension *de $\mathcal{H}$, P-dim($\mathcal{H}$), est le cardinal du plus grand des sous-ensembles de $\mathcal{X}$ P-pulvérisé par $\mathcal{H}$, si ce cardinal est fini, l'infini sinon.*

La dimension V_γ est une variante à marge de la dimension de Vapnik.

DÉFINITION 9 (Dimension V_γ [ALO 97, GUR 01]) *Soit $\mathcal{H}$ une famille de fonctions d'un domaine $\mathcal{X}$ dans $\mathbb{R}$. Pour $\gamma > 0$, un sous-ensemble $\mathcal{X}_m = \{x_i\}$, $(1 \leq i \leq m)$ de $\mathcal{X}$ est dit être* V_γ-pulvérisé *par $\mathcal{H}$ s'il existe un scalaire b tel que, pour tout vecteur binaire $v_y = (y_i) \in \{-1, 1\}^m$, il existe une fonction $h_y \in \mathcal{H}$ satisfaisant*

$$(h_y(x_i) - b)\, y_i \geq \gamma, \ (1 \leq i \leq m). \qquad [10.18]$$

La dimension *V_γ de $\mathcal{H}$, V_γ-dim($\mathcal{H}$), est le cardinal du plus grand des sous-ensembles de $\mathcal{X}$ V_γ-pulvérisé par $\mathcal{H}$, si ce cardinal est fini, l'infini sinon.*

De même que la dimension de Vapnik peut être considérée comme une variante uniforme de la pseudo-dimension, la dimension V_γ peut être vue comme une variante uniforme de la dimension "fat-shattering".

DÉFINITION 10 (Dimension fat-shattering [KEA 90, KEA 94]) *Soit $\mathcal{H}$ une famille de fonctions d'un domaine $\mathcal{X}$ dans $\mathbb{R}$. Pour $\gamma > 0$, un sous-ensemble $\mathcal{X}_m = \{x_i\}$, $(1 \leq i \leq m)$ de $\mathcal{X}$ est dit être* γ-pulvérisé *par $\mathcal{H}$ s'il existe un vecteur $v_b = (b_i) \in \mathbb{R}^m$ tel que, pour tout vecteur binaire $v_y = (y_i) \in \{-1, 1\}^m$, il existe une fonction $h_y \in \mathcal{H}$ satisfaisant*

$$(h_y(x_i) - b_i)\, y_i \geq \gamma, \ (1 \leq i \leq m). \qquad [10.19]$$

La dimension fat-shattering à marge *γ de $\mathcal{H}$, P_γ-dim($\mathcal{H}$), est le cardinal du plus grand des sous-ensembles de $\mathcal{X}$ γ-pulvérisé par $\mathcal{H}$, si ce cardinal est fini, l'infini sinon.*

On trouvera dans [ALO 97] un lemme de Sauer étendu faisant intervenir la dimension fat-shattering. Si les différentes notions de dimension VC dédiées aux classes de fonctions à valeurs dans des ensembles finis, et par extension à la discrimination multi-classe, ont été introduites progressivement dans la littérature, une théorie unificatrice en est proposée dans [BEN 95]. Celle-ci repose sur la notion de Ψ-dimension, que nous définissons donc en premier lieu.

DÉFINITION 11 (Ψ-dimensions [BEN 95]) *Soit $\mathcal{H}$ une famille de fonctions d'un domaine $\mathcal{X}$ dans un ensemble de cardinalité finie $\{1, \ldots, Q\}$. Soit Ψ un ensemble d'applications ψ de $\{1, \ldots, Q\}$ dans $\{-1, 1, *\}$, où $*$ représente l'élément par défaut. Un sous-ensemble $\mathcal{X}_m = \{x_i\}$, $(1 \leq i \leq m)$ de $\mathcal{X}$ est dit être* Ψ-pulvérisé *par $\mathcal{H}$ s'il existe une application $\psi^m = \left(\psi^{(1)}, \ldots, \psi^{(i)}, \psi^{(m)}\right)$ dans Ψ^m telle que, pour tout vecteur v_y dans $\{-1, 1\}^m$, il existe une fonction $h_y \in \mathcal{H}$ satisfaisant*

$$\left(\psi^{(1)} \circ f_y(x_1), \ldots, \psi^{(i)} \circ f_y(x_i), \ldots, \psi^{(m)} \circ f_y(x_m)\right) = v_y. \qquad [10.20]$$

La Ψ-dimension de $\mathcal{H}$, Ψ-dim$(\mathcal{H})$, est le cardinal du plus grand des sous-ensembles de $\mathcal{X}$ Ψ-pulvérisé par $\mathcal{H}$, si ce cardinal est fini, l'infini sinon.

Deux Ψ-dimensions sont à distinguer, dans la mesure où elles ont fait l'objet d'études spécifiques, la dimension graphique et la dimension de Natarajan.

DÉFINITION 12 (Dimension graphique [DUD 87, NAT 89]) *Soit $\mathcal{H}$ une famille de fonctions d'un domaine $\mathcal{X}$ dans un ensemble dénombrable $\mathcal{Y}$. Pour toute fonction h de $\mathcal{H}$, le* graphe *de h est la fonction $\mathcal{G}(h)$ de $\mathcal{X} \times \mathcal{Y}$ dans $\{0, 1\}$ définie par :*

$$\forall (x, y) \in \mathcal{X} \times \mathcal{Y}, \;\; \mathcal{G}(h)(x, y) = 1 \Longleftrightarrow h(x) = y. \qquad [10.21]$$

*L'*espace des graphes *de $\mathcal{H}$ est $\mathcal{G}(\mathcal{H}) = \{\mathcal{G}(h) \;/\; h \in \mathcal{H}\}$. Ces définitions étant données, la* dimension graphique *de $\mathcal{H}$, G-dim$(\mathcal{H})$, est la dimension VC de l'espace $\mathcal{G}(\mathcal{H})$.*

Lorsque les fonctions de $\mathcal{H}$ prennent leurs valeurs dans un ensemble fini, la reformulation de cette définition comme étant celle d'une Ψ-dimension est la suivante.

DÉFINITION 13 (Dimension graphique) *Soit $\mathcal{H}$ une famille de fonctions d'un domaine $\mathcal{X}$ dans $\{1, \ldots, Q\}$. La* dimension graphique *de $\mathcal{H}$ est la Ψ-dimension de $\mathcal{H}$ lorsque l'ensemble Ψ est constitué des applications ψ_k, $(1 \leq k \leq Q)$, telles que ψ_k prend la valeur 1 si son argument est égal à k et la valeur -1 dans le cas contraire. Reformulé dans le contexte de la discrimination multi-classe, les fonctions ψ_k sont les indicatrices des catégories.*

DÉFINITION 14 (Dimension de Natarajan [NAT 89]) *Soit $\mathcal{H}$ une famille de fonctions d'un domaine $\mathcal{X}$ dans $\{1, \ldots, Q\}$. La* dimension de Natarajan *de $\mathcal{H}$, N-dim($\mathcal{H}$) est la Ψ-dimension de $\mathcal{H}$ lorsque l'ensemble Ψ est constitué des C_Q^2 applications $\psi_{k,l}$, $(1 \leq k < l \leq Q)$, telles que $\psi_{k,l}$ prend la valeur 1 si son argument est égal à k, la valeur -1 si son argument est égal à l et la valeur $*$ dans tous les autres cas.*

De même qu'il existe un lemme de Sauer généralisé faisant intervenir la dimension fat-shattering, il en existe qui sont dédiés aux Ψ-dimensions. La principale référence dans le domaine est [HAU 95b]. En résumé, en discrimination, la dimension VC est utilisée pour les modèles bi-classes à valeurs binaires. Ses extensions à marge sont utilisées pour les systèmes à valeurs réelles calculant des dichotomies, tandis que le cas de la discrimination multi-classe par des modèles prenant leurs valeurs dans l'ensemble des catégories est traité au moyen des Ψ-dimensions. La théorie n'a pas encore abordé en profondeur le cas le plus général, celui des classifieurs à marge multi-classe. Ce manque est significatif, dans la mesure où il laisse sans solution satisfaisante un cas aussi classique que celui du PMC utilisant des unités de sortie munies des fonctions d'activation standard, sigmoïde ou softmax (voir la section 10.3.3). On ne sait actuellement le traiter qu'en appliquant des méthodes de décomposition. A notre connaissance, les seuls travaux proposant une extension à marge des Ψ-dimensions sont décrits dans [GUE 04]. Cette référence établit en particulier un lemme de Sauer pour l'extension à marge de la dimension de Natarajan.

Avec les dimensions VC, les autres mesures de capacité apparaissant dans les bornes "à la Vapnik" sont principalement des nombres de couverture. Ces nombres sont des outils de l'analyse fonctionnelle qui trouvent de nombreuses applications, par exemple pour l'étude de la compacité des opérateurs [CAR 90]. Ils reposent sur la notion d'ϵ-couverture et d'ϵ-réseau.

DÉFINITION 15 (pseudo-métrique) *On appelle* pseudo-métrique *sur un ensemble E une application ρ de $E \times E$ dans $\mathbb{R}_+$ vérifiant les propriétés d'une métrique à l'exception de la séparation. On a dans ce cas $x = x' \Longrightarrow \rho(x, x') = 0$. A l'inverse, $\rho(x, x') = 0 \not\Longrightarrow x = x'$. Un* espace pseudo-métrique *est un espace muni d'une pseudo-métrique.*

DÉFINITION 16 (ϵ-couverture et ϵ-réseau) *Soit (E, ρ) un espace (pseudo-)métrique et E' un sous-ensemble de E. Une* ϵ-couverture *de E' est un recouvrement de E' par des boules de rayon ϵ dont les centres appartiennent à E. Ces centres forment un* ϵ-réseau *de E'. On parle parfois aussi d'*ϵ-squelette.

DÉFINITION 17 (Nombres de couverture) *Soit (E, ρ) un espace (pseudo-)métrique. Si $E' \subset E$ possède un ϵ-réseau de cardinalité finie, alors son* nombre de couverture *$N(\epsilon, E', \rho)$ est la plus petite cardinalité de ses ϵ-réseaux. S'il n'existe pas de tel ϵ-réseau, alors $N(\epsilon, E', \rho) = \infty$.*

Dans le cas des classes de fonctions à valeurs binaires, il existe un lien simple entre ces nombres et la fonction de croissance. Soit en effet $\mathcal{F}$ une telle classe de fonctions, définie sur $\mathcal{X}$, et $\mathcal{X}_m = (x_i)_{1 \leq i \leq m}$ un élément de $\mathcal{X}^m$. En munissant $\mathcal{F}$ de la pseudo-métrique suivante :

$$\forall (f, f') \in \mathcal{F}^2, \;\; d_m(f, f') = \max_{\mathcal{X}_m \in \mathcal{X}^m} \max_{x_i \in \mathcal{X}_m} |f(x_i) - f'(x_i)|$$

on obtient immédiatement

$$N(\epsilon, \mathcal{F}, d_m) = \Pi_{\mathcal{F}}(m)$$

dès lors que $\epsilon \in]0, 1/2[$.

DÉFINITION 18 (Nombres de couverture avec crochets) *Soit $\mathcal{F}$ un espace de fonctions inclus dans $\mathbb{R}^{\mathcal{X}}$. Alors, un* ϵ-réseau avec crochets *de $\mathcal{F}$ (on parle parfois aussi d'*ϵ-squelette avec crochets*) est un ensemble $(f_1, g_1), \ldots, (f_N, g_N)$ de couples d'éléments de $\mathbb{R}^{\mathcal{X}}$ tel que*

$$\forall i \in [[1, N]], \;\; ||g_i - f_i|| \leq \epsilon$$

$$\text{et } \forall f \in \mathcal{F}, \;\; \exists i \in [[1, N]] \;/\; f_i \leq f \leq g_i.$$

L'ϵ-nombre de couverture avec crochets *de $\mathcal{F}$, noté $N_{[\,]}(\epsilon, \mathcal{F}, ||.||)$, dépendant de la norme, est le cardinal minimal d'un ϵ-réseau avec crochets de $\mathcal{F}$, quand un tel réseau existe, et ∞ sinon.*

10.3.3. *Bornes sur les mesures de capacité*

Le calcul de la dimension VC, première mesure de capacité apparaissant dans les bornes, a fait l'objet de nombreux travaux. Dans le domaine du connexionnisme, la littérature la plus abondante porte sur le cas des PMC. Nous en proposons ici les résultats principaux, en nous appuyant en particulier sur des exposés de synthèse, comme [ANT 97, SON 98, ANT 99]. Le premier cas considéré, le plus simple et le plus classique, est celui du perceptron.

PROPOSITION 3 *La dimension VC d'un séparateur affine opérant dans $\mathbb{R}^d$ est égale à $d + 1$.*

La preuve est très simple. La résolution d'un système linéaire nous donne VC-dim$(\mathcal{H}) \geq d+1$, tandis que le théorème de Radon [GRü 67] permet de conclure que VC-dim$(\mathcal{H}) \leq d+1$. Le cas suivant est celui des PMC à unités à seuil utilisés pour le calcul des dichotomies (réseaux ne possédant qu'une unité de sortie). Nous considérons ici des architectures à propagation avant, ou "feedforward", c'est-à-dire sans cycle. La référence de base sur le sujet est [BAU 89]. Le résultat fondamental qu'elle contient s'appuie sur le lemme suivant.

LEMME 3 (Composition des fonctions de croissance) *Soit $\mathcal{H}$ la classe de fonctions calculées par un réseau à propagation avant constitué de N* unités de calcul *(unités autres que les unités d'entrée). Soit $\mathcal{H}_i$ la classe de fonctions calculée par sa i-ième unité (de calcul). Alors pour tout entier positif m*

$$\Pi_{\mathcal{H}}(m) < \Pi_{i=1}^{N} \Pi_{\mathcal{H}i}(m). \qquad [10.22]$$

En combinant ce lemme et le lemme de Sauer, on obtient immédiatement le théorème suivant.

THÉORÈME 14 (Corollaire 3 dans [BAU 89]) *Soit $\mathcal{H}$ la classe de fonctions calculée par un réseau à propagation avant constitué de N unités de calcul à seuil, dont une unique unité de sortie, et W connexions. Alors*

$$\textit{VC-dim}(\mathcal{H}) \leq 2W \log_2 (eN). \qquad [10.23]$$

Ce théorème a été étendu par Shawe-Taylor et Anthony au cas des réseaux à unités à seuil à sorties multiples (permettant le calcul de polychotomies).

THÉORÈME 15 (Corollaire 4.2 dans [SHA 91]) *Soit $\mathcal{H}$ la classe de fonctions calculée par un réseau à propagation avant constitué de N unités de calcul à seuil et W connexions. Alors*

$$\textit{G-dim}(\mathcal{H}) \leq 2W \log_2 (eN). \qquad [10.24]$$

Il convient de remarquer que ce résultat est contre-intuitif, dans la mesure où il ne fait pas intervenir le nombre d'unités de sortie et par conséquent ne fait pas intervenir le nombre de catégories. Sous des hypothèses plus fortes concernant l'architecture du réseau, des améliorations au résultat de Baum et Haussler furent apportées par Sakurai.

Les deux théorèmes qui suivent concernent des réseaux à propagation avant entièrement connectés, i.e. pour lesquels chaque unité d'une couche est connectée à chaque unité de la couche suivante. Les fonctions d'activation sont des fonctions à seuil.

THÉORÈME 16 (Théorème 3.3 dans [SAK 93]) *Soit $\mathcal{H}$ la classe de fonctions calculée par un réseau à propagation avant constitué d'une couche d'entrée de n unités, d'une couche cachée de h unités et une couche de sortie d'une seule unité, la fonction d'activation étant une fonction à seuil. Alors, pour $n \geq 32$ et $h \geq 256$*

$$\textit{VC-dim}(\mathcal{H}) \leq nh \log_2(h) \left(1 + 2\frac{\log_2(\log_2(h))}{\log_2(h)}\right). \qquad [10.25]$$

[SAK 93] montre aussi la borne inférieure suivante :

THÉORÈME 17 (Borne inférieure) *Soit $\mathcal{H}$ la classe de fonctions calculée par un réseau à propagation avant constitué d'une couche d'entrée de n unités, d'une couche cachée de h unités et une couche de sortie d'une seule unité, la fonction d'activation étant une fonction à seuil. Alors, pour $n \geq 3$ et $h \leq 2^{n/2-2}$,*

$$\textit{VC-dim}(\mathcal{H}) \geq nh \log_2(h/4)/8 \geq W \log_2(h/4)/32$$

où $W = nh + 2h + 1$ est le nombre total de paramètres (poids et seuils).

La situation devient nettement plus compliquée lorsque l'étude ne se restreint plus au cas des unités à seuil. La classe des fonctions d'activation la plus utilisée est celle des sigmoïdes.

DÉFINITION 19 (Fonction sigmoïde) *Une fonction f, de $\mathbb{R}$ dans $\mathbb{R}$, sera appelée* sigmoïde *si elle vérifie les deux propriétés suivantes :*

1) les limites de f en $-\infty$ et $+\infty$ existent en sont distinctes ;

2) il existe un réel t_0 tel que f est différentiable en t_0 et $f'(t_0) \neq 0$.

Nous qualifierons de standard la fonction sigmoïde suivante :

$$f(t) = (1 + \exp(-t))^{-1}.$$

Il existe de nombreuses autres fonctions d'activation, parmi lesquelles on peut encore citer les fonctions à base radiale (r.b.f.) ou les exponentielles normalisées (fonctions softmax). Celles-ci sont introduites ailleurs dans ce livre (voir en particulier le chapitre intitulé "Fonctions de base radiales"). Des neurones nettement plus différents, plus proches du biologique, existent aussi ; il s'agit des réseaux à trains d'excitations ("spiking neurons") [GER 02]. Dans ce qui suit, nous restreignons l'étude au cas des sigmoïdes (avec une unité de sortie pouvant être à seuil). Le résultat de base est que dans ce cas, la dimension VC d'un PMC peut être infinie. Il est illustré en particulier par des travaux de Sontag.

THÉORÈME 18 (Lemme 8.2 dans [SON 92]) *Soit $\mathcal{H}$ la classe de fonctions calculée par un réseau à propagation avant constitué d'une couche d'entrée de 2 unités ($x \in \mathbb{R}^2$), deux unités cachées de fonction d'activation :*

$$f(t) = \frac{1}{\pi}\arctan(t) + \frac{\cos(t)}{\alpha\,(1+t^2)} + \frac{1}{2}$$

où α est un réel supérieur à 2π et d'une unité de sortie à seuil.

$$\textit{VC-dim}(\mathcal{H}) = +\infty. \qquad [10.26]$$

Ce résultat négatif est tempéré par un résultat positif, dans le cas particulier où la fonction d'activation est la fonction sigmoïde standard (on pourra consulter [ANT 99] pour le cas de fonctions d'activation polynomiales par morceau).

THÉORÈME 19 (Théorème 1 dans [MAC 93]) *Soit $\mathcal{H}$ la classe de fonctions calculée par un réseau à propagation avant dont les unités ont une fonction d'activation sigmoïdale standard. Alors*

$$\textit{VC-dim}(\mathcal{H}) < +\infty. \qquad [10.27]$$

Deux remarques doivent être faites concernant ce travail. Tout d'abord, les auteurs établissent en fait le théorème 19 pour une famille de fonctions d'activation incluant la sigmoïde standard, la famille des fonctions *exp-RA définissables*. De plus, si ce résultat ne fournit pas explicitement de borne supérieure, la preuve proposée est constructive. Elle peut donc être utilisée pour en dériver une. Des bornes inférieures et supérieures sur la dimension VC, ou la dimension de Vapnik dans le cas d'unités de sortie à valeurs dans $\mathbb{R}$, sont fournies par un ensemble de références. Nous commençons par une borne inférieure établie dans [KOI 97] (voir aussi [GOL 95]).

THÉORÈME 20 (Théorème 4 dans [KOI 97]) *Pour une certaine constante universelle c, pour tout entier W, il existe une classe $\mathcal{H}$ de fonctions calculée par un réseau à propagation avant dont les unités de calcul ont une fonction d'activation sigmoïdale (au sens de la définition 19), dont le nombre des connexions est $\leq cW$, telle que $\mathcal{H}$ peut V-pulvériser un sous-ensemble de $\mathbb{R}$ de cardinalité W^2.*

La dimension de Vapnik de certains réseaux à unités sigmoïdales est donc au moins quadratique en le nombre de connexions. Une borne supérieure est proposée dans [KAR 95].

THÉORÈME 21 (Section 3 dans [KAR 95]) *Soit $\mathcal{H}$ la classe de fonctions calculée par un réseau à propagation avant dont les unités de calcul ont une fonction d'activation sigmoïdale. Alors, avec les mêmes notations que précédemment*

$$\textit{VC-dim}(\mathcal{H}) = O(N^2W^2). \qquad [10.28]$$

Le théorème 8.13 de [ANT 99], utilisant une technique développée par les auteurs du théorème 21 dans [KAR 97], fournit une constante en $1 + o(1)$ dans ce $O(.)$:

$$\text{VC-dim}(\mathcal{H}) \leq N^2W^2 + 11WN\log_2(18WN^2).$$

Le théorème suivant, améliorant un résultat de [BAR 96], précise les choses dans le cas où le "fan-in", c'est-à-dire le nombre maximal de neurones fournissant une entrée à un neurone donné, est borné, et lorsque le domaine est discret.

THÉORÈME 22 (D'après les théorèmes 8.11 et 8.12 dans [ANT 99]) *Soit $D, n \in \mathbb{N}$. Considérons des neurones ayant pour fonction d'activation la sigmoïde standard. Soit un réseau feedforward à n entrées, k neurones connectés exclusivement aux entrées, et une sortie connectée exclusivement à ces k neurones, avec un fan-in des neurones de la couche cachée majoré par N, sur le domaine $\{-D, -D+1, \ldots, D\}^n$. Soit $\mathcal{H}$ l'ensemble des fonctions que peut calculer ce réseau sur ce domaine. Alors*

$$\textit{VC-dim}(\mathcal{H}) \leq 2W\log_2(60ND)$$

où W est le nombre de paramètres. Si le domaine est réduit à $\{0,1\}^n$, alors

$$\textit{VC-dim}(\mathcal{H}) \leq 2W\log_2(60N)$$

et pour une certaine constante universelle $c > 0$ un réseau feedforward ainsi structuré atteint une VC-dimension $\geq cW$ avec W paramètres.

Jusqu'à présent, nous avons principalement considéré des réseaux à valeurs binaires (sorties appartenant à $\{0,1\}$ ou $\{0,1\}^Q$), dont les unités de sortie étaient à seuil. Cependant, comme nous l'avons vu dans la section 10.2, ou plus récemment avec le théorème 20, il est également possible d'utiliser en discrimination des modèles à valeurs réelles (appartenant à $\mathbb{R}$ ou $\mathbb{R}^Q$). Dans ce cas, la mesure de capacité à borner n'est plus la dimension VC, mais une de ses extensions ou un nombre de couverture.

Nous présentons tout d'abord le résultat suivant, utilisant la pseudo-dimension. Il s'agit du théorème 14.2 dans [ANT 99], lui-même découlant de résultats de [KAR 97] ; il découle en fait de bornes sur la VC-dimension citées plus haut et de liens entre la VC-dimension et la pseudo-dimension qui ont leur origine dans [VID 97].

THÉORÈME 23 *Soit $\mathcal{H}$ la famille de fonctions calculée par un réseau de neurones à W paramètres, n neurones de calcul, et ayant la fonction d'activation sigmoïde standard sur toutes ses unités (même la sortie). Alors*

$$\textit{P-dim}(\mathcal{H}) \leq ((W+2)n)^2 + 11(W+2)n\log_2(18(W+2)n^2).$$

Toutefois, la référence de base concernant les performances en généralisation des MLP à valeurs réelles utilisés pour le calcul des dichotomies est [BAR 98b]. Ce travail s'appuie sur la fat-shattering dimension. Pour l'introduire, nous commençons, comme dans le cas de la dimension VC, par considérer le cas d'un simple séparateur linéaire.

THÉORÈME 24 (Théorème 4.6 dans [BAR 99b]) *Soit $\mathcal{H}$ la classe des fonctions représentables par un séparateur linéaire calculant sur $\mathcal{X}$ les fonctions h_w paramétrées par le vecteur w et définies par*

$$h_w(x) = \langle w, x\rangle. \qquad [10.29]$$

Alors, sous les hypothèses que les éléments de $\mathcal{X}$ sont inclus dans la boule de rayon $\Lambda_{\mathcal{X}}$ et que le paramètre w satisfait $\|w\| \leq \Lambda_w$, on a, pour toute valeur positive de γ,

$$P_\gamma\textit{-dim}(\mathcal{H}) \leq \left(\frac{\Lambda_w\Lambda_{\mathcal{X}}}{\gamma}\right)^2. \qquad [10.30]$$

Une autre référence sur le même sujet est [GUR 01]. La transition vers le cas des PMC à unités sigmoïdales est fournie par [GUR 95, GUR 97]. Bartlett formule le résultat central de ces articles sous la forme suivante.

THÉORÈME 25 (Proposition 16 dans [BAR 98b]) *Soit n un entier strictement positif et $\mathcal{F}$ la famille de fonctions f_w, de $\mathbb{R}^n$ dans $\{-1, 1\}$, telles que f_w associe à x sgn $(\langle w, x\rangle)$. Soit $\mathcal{H}$ la famille de fonctions calculée par un PMC possédant une couche cachée de taille arbitraire N dont les unités (à seuil) calculent une fonction de $\mathcal{F}$. Sous l'hypothèse supplémentaire que les valeurs des poids de la couche haute sont bornées, ce qui se traduit par*

$$\mathcal{H} = \left\{ \sum_{i=1}^{N} \alpha_i f_{w_i} \;\; / \;\; N \in \mathbb{N}, f_{w_i} \in \mathcal{F}, \sum_{i=1}^{N} |\alpha_i| \leq A \right\}, \qquad [10.31]$$

alors

$$P_\gamma\textit{-dim}(\mathcal{H}) = O\left(\frac{A^2n^2}{\gamma^2}\log\left(\frac{n}{\gamma}\right)\right). \qquad [10.32]$$

En s'appuyant sur ce résultat, il établit le théorème suivant, relatif cette fois aux réseaux dont les unités cachées sont à valeurs réelles.

THÉORÈME 26 (Théorème 17 dans [BAR 98b]) *Soit $\mathcal{F}$ une famille non vide de fonctions de $\mathcal{X}$ dans $[-M/2, M/2]$. Pour $A \geq 0$, définissons la classe $\mathcal{H}$ des réseaux à une couche cachée dont les unités cachées appartiennent à $\mathcal{F}$ de la manière suivante :*

$$\mathcal{H} = \left\{ \sum_{i=1}^{N} w_i f_i \ / \ N \in \mathbb{N}, f_i \in \mathcal{F}, \sum_{i=1}^{N} |w_i| \leq A \right\}. \quad [10.33]$$

Sous l'hypothèse que $\gamma > 0$ satisfait $d = P_{\frac{\gamma}{32A}}$-dim$(\mathcal{F}) \geq 1$, alors

$$P_\gamma\text{-dim}(\mathcal{H}) \leq \frac{cM^2A^2d}{\gamma^2} \left(\log \left(\frac{MAd}{\gamma} \right) \right)^2 \quad [10.34]$$

où c est une constante universelle.

Un résultat équivalent, indépendant du nombre de neurones et du nombre de poids, existe pour des réseaux à grands nombres de couches :

THÉORÈME 27 (Théorème 14.19 dans [ANT 99]) *Soit un PMC sans cycle [2] dont toutes les unités ont la même fonction d'activation, croissante, L-lipschitzienne, à valeurs dans $[-b, b]$. On fait l'hypothèse que la somme des valeurs absolues des paramètres des neurones situés sur une couche donnée est majorée par V. En notant $\mathcal{H}$ la famille de fonctions calculée par ce PMC avec cette restriction sur les poids, si $b \geq 1$, si $V \geq 1/(2L)$, si $\epsilon < 16VbL$, si le nombre de couches est l et si $\mathcal{X} \subset \mathbb{R}^n$, alors*

$$P_\gamma\text{-dim}(\mathcal{H}) \leq 4(32b/\epsilon)^{2l}(2VL)^{l(l+1)} \log(2n+2).$$

Ce résultat revêt une importance particulière car il est indépendant du nombre de neurones. On peut donc avoir des fonctions très complexes, pourvu qu'elles n'aient pas des poids trop grands ; cela n'est pas sans évoquer les résultats basés sur les coefficients de Lipschitz ou les normes Hölderiennes (voir théorème 30). En particulier,

2. Mais éventuellement doté de connexions qui "sautent" des couches.

ce résultat illustre beaucoup mieux que ceux basés sur le nombre de connexions le fait que des réseaux immenses soient efficaces (e.g., réseaux biologiques).

On sait que, lorsque c'est possible, il convient de travailler directement avec la fonction de croissance plutôt qu'avec la dimension VC. De la même façon, lorsque l'intervalle de confiance de la borne fait apparaître un nombre de couverture, il est plus intéressant de tenter de le borner directement plutôt que de passer par un lemme de Sauer étendu et une dimension VC généralisée.

On peut ainsi utiliser la borne suivante sur les nombres de couverture pour la norme L^∞. Ces nombres de couverture peuvent être utilisés en classification (voir section 10.4.2.3) mais aussi en régression directement (voir [VID 97] ou [ANT 99]) ; voir aussi le théorème 44.

THÉORÈME 28 (Théorèmes 14.5 et 14.9 dans [ANT 99]) *Soit un réseau de neurones à l couches, avec des connexions exclusivement entre couches adjacentes et W paramètres ; supposons que toutes les fonctions d'activation sont identiques, à valeurs dans $[-b, b]$, L-Lipschitzienne et croissantes ; supposons en outre que pour $V > 0$ et $L > 1/V$, dans toutes les couches sauf la première, le vecteur w de poids associé à une unité donnée est de norme L^1 majorée par V. Alors, si $\epsilon < 2b$, et si $\mathcal{H}$ est la famille de fonctions calculée par ce réseau, pour tout E de cardinal $\leq m$,*

$$N(\epsilon, \mathcal{H}_{|E}, L^\infty) \leq \left(\frac{4embW(LV)^l}{\epsilon(LV-1)} \right)^W$$

et

$$P_\gamma\text{-}dim(\mathcal{H}) \leq 16W \left(l \log(LV) + 2\log(32W) + \log\left(\frac{b}{\epsilon(LV-1)} \right) \right).$$

Dans [WIL 00, WIL 01], les auteurs proposent une méthode pour borner directement les nombres de couverture des machines à noyau [SCH 02]. Cette méthode relie le calcul de cette borne à celui d'une borne sur les nombres d'entropie d'un opérateur linéaire. Les résultats principaux sur lesquels se fonde cette démarche sont résumés ci-dessous. Dans un but de simplification, l'exposé se limite au cas des SVM. L'algorithme d'apprentissage de ces machines étant présenté dans le chapitre intitulé "Fonctions de base radiales", nous nous bornons à rappeler ce qui nous sera utile ici, c'est-à-dire l'architecture sur laquelle elles s'appuient.

DÉFINITION 20 (Pseudo-métrique) *Soit $\mathcal{H}$ une classe de fonctions sur $\mathcal{X}$ à valeurs réelles. Pour un ensemble s d'éléments de $\mathcal{X}$ en nombre fini, on définit la pseudo-métrique d_s sur $\mathcal{H}$ comme :*

$$\forall (h, h') \in \mathcal{H}^2,\ d_s(h, h') = \max_{x \in s} |h(x) - h'(x)| . \qquad [10.35]$$

On note : $N_\infty(\epsilon, \mathcal{H}, m) = \sup_{s_m \in \mathcal{X}^m} N(\epsilon, \mathcal{H}, d_{s_m})$.

DÉFINITION 21 (Nombres d'entropie) *Soit E un espace de Banach et $\tilde{E}$ un sous-ensemble de E. Pour n entier positif, le* n-ième nombre d'entropie *de $\tilde{E}$ est le plus petit réel positif ϵ tel qu'il existe un ϵ-réseau de $\tilde{E}$ (inclus dans E) de cardinalité au plus n. Soit E et F deux espaces de Banach et U_E la boule unité fermée de E. Le* n-ième nombre d'entropie *de S appartenant à $\mathfrak{L}(E, F)$ est donné par*

$$\epsilon_n(S) = \epsilon_n\left(S(U_E)\right). \qquad [10.36]$$

Le n-ième nombre d'entropie diadique *de S est défini par*

$$e_n(S) = \epsilon_{2^{n-1}}(S). \qquad [10.37]$$

DÉFINITION 22 (SVM) *Soit κ un noyau défini positif sur $\mathcal{X}$ et $E_{\Phi(\mathcal{X})}$ un espace de Hilbert à noyau reproduisant (RKHS [ARO 50, WAH 90, WAH 99]) correspondant, où la fonction Φ vérifie :*

$$\forall (x, x') \in \mathcal{X}^2,\ \kappa(x, x') = \langle \Phi(x), \Phi(x') \rangle. \qquad [10.38]$$

La classe de fonction $\mathcal{H} = \{h\}$ calculée par la SVM de noyau κ est l'ensemble des applications affines

$$h(x) = \langle w, \Phi(x) \rangle + b, \qquad [10.39]$$

où w est un vecteur de $E_{\Phi(\mathcal{X})}$ et b un scalaire.

DÉFINITION 23 (Opérateur d'évaluation) *Soit s_m un élément quelconque de $\mathcal{X}^m$. S_{s_m} est l'opérateur linéaire défini par :*

$$\begin{array}{rccc} S_{s_m} : & E_{\Phi(\mathcal{X})} & \longrightarrow & \ell_\infty^m \\ & w & \mapsto & S_{s_m}(w) = \left(\langle w, \Phi(x_i) \rangle\right)_{1 \le i \le m} \end{array}$$

Ces définitions posées, le lien entre les nombres de couverture d'une SVM linéaire (i.e. sans biais), et les nombres d'entropie de l'opérateur d'évaluation défini ci-dessus est donné par la proposition suivante.

PROPOSITION 4 ([WIL 01], page 2523) *Si pour tout $s_m \in \mathcal{X}^m$, $\epsilon_n(S_{s_m}) \leq \epsilon$, alors $N_\infty(\epsilon, \mathcal{U}, m) \leq n$, où $\mathcal{U}$ est la famille des SVM satisfaisant les contraintes $\|w\|_{E_{\Phi(\mathcal{X})}} \leq 1$ et $b = 0$.*

Un résultat similaire, de portée plus générale, est fourni par le Lemme 11 de [WIL 00]. L'intérêt d'introduire des nombres d'entropie d'opérateurs réside naturellement dans le fait qu'il existe des résultats permettant de les borner. Nous citons ici les deux principaux.

PROPOSITION 5 (Proposition 1.3.1 dans [CAR 90]) *Soient E et F deux espaces de Banach et S un opérateur de E dans F ($S \in \mathfrak{L}(E, F)$) de rang r. Alors pour tont entier strictement positif n,*

$$\epsilon_n(S) \leq 4\|S\| n^{-1/r}. \qquad [10.40]$$

THÉORÈME 29 (Théorème de Maurey-Carl, lemme 6.4.1 dans [CAR 90]) *Soit H un espace de Hilbert, m un entier strictement positif et $S \in \mathfrak{L}(H, \ell_\infty^m)$ ou $S \in \mathfrak{L}(\ell_1^m, H)$. Alors, pour $1 \leq n \leq m$,*

$$\epsilon_{2^{n-1}}(S) \leq c\|S\| \left(\frac{1}{n} \log_2 \left(1 + \frac{m}{n} \right) \right)^{1/2}, \qquad [10.41]$$

où c est une constante universelle.

Si la proposition 5 apparaît d'un usage plus simple que le théorème 29, dans la mesure en particulier où elle ne fait pas intervenir de constante universelle, son domaine d'application est aussi plus limité. On se convaincra en effet facilement du fait que seul le théorème 29 s'avère utile lorsque l'espace de représentation est de dimension infinie. On trouvera dans [GUE 05a] une extension de ces travaux au cas des SVM multi-classes. Cette extension s'appuie en particulier sur une généralisation du théorème de Maurey-Carl démontrée dans [GUE 05b]. Des bornes plus fines sur les nombres de couverture des SVM sont décrites dans [GUO 02].

On considère ci-dessous des espaces de fonctions régulières au sens de Hölder ([KOL 61, LOR 66, BIR 67, DUD 84, VAA 96]), pour leur généralité (en particulier, les réseaux de neurones avec activation sigmoïde ou les SVM sont des séparateurs hölderiens) :

THÉORÈME 30 (Nombres de couverture pour des espaces de fonctions régulières) *Considérons $\mathcal{F}_{B,\alpha,d}$, où $\alpha > 0$, la famille des applications f de $[0,1]^{d-1}$ dans $[0,1]$ telles que $|||f|||_\alpha$ est bien défini et borné par B, avec*

$$|||f|||_\alpha = \max_{\sum k_i \leq \lfloor \alpha \rfloor} \sup_x \left| \frac{\partial^{\sum k_i} f(x)}{\partial x_1^{k_1} \partial x_2^{k_2} \dots \partial x_d^{k_d}} \right| \quad [10.42]$$

$$+ \max_{\sum k_i = \lfloor \alpha \rfloor} \sup_{x \neq y} \frac{\left| \frac{\partial^{\sum k_i} f(x)}{\partial x_1^{k_1} \partial x_2^{k_2} \dots \partial x_d^{k_d}} - \frac{\partial^{\sum k_i} f(y)}{\partial x_1^{k_1} \partial x_2^{k_2} \dots \partial x_d^{k_d}} \right|}{|x-y|^{\alpha - \lfloor \alpha \rfloor}} \quad [10.43]$$

où $\lfloor \alpha \rfloor$ est ici le plus grand entier strictement inférieur *à α (égal à $\alpha - 1$ si α est entier). Cet espace de fonctions est appelé un espace de* Hölder.

Notons $\mathcal{C}_{B,\alpha,d}$ l'ensemble des ensembles de la forme $\{(x,t)/f(x) < t\}$ avec $f \in \mathcal{F}_{B,\alpha,d}$.

Alors

$$\log N_{[\,]}(\epsilon, \mathcal{F}_{1,\alpha,d}, L^2) \leq K(\alpha, d) \left(\frac{1}{\epsilon}\right)^{(d-1)/\alpha}$$

$$\log N_{[\,]}(\epsilon, \mathcal{C}_{1,\alpha,d}, L^2) \leq K'(\alpha, d) \left(\frac{1}{\epsilon}\right)^{2(d-1)/\alpha}$$

pour certaines applications K et K'.

On peut noter que cette famille de fonctions a une VC-dimension infinie. On peut voir [DEV 96, chapitre 28] pour plus d'informations sur les nombres de couverture pour la norme L^∞ au lieu de la norme L^1. Le résultat suivant provient de [BIR 67, DUD 66].

THÉORÈME 31 (Nombres de couverture pour des espaces d'ensembles convexes) *Considérons $\mathcal{F}$ l'ensemble des ensembles convexes de $[0,1]^d$, où $d \geq 2$, alors*

$$\log N_{[\,]}(\epsilon, \mathcal{F}, L^r) \leq K \left(\frac{1}{\epsilon}\right)^{r(d-1)/2}.$$

Ce résultat ne conduit à une classe de Donsker que pour $d = 2$. Toutefois, même en dimension 2, la VC-dimension est infinie - ainsi on a un avantage ici à utiliser les classes de Donsker au lieu des classes VC. On verra en outre que même des nombres de couverture exponentiels au-delà de la limite Donsker peuvent conduire à des vitesses de convergence (e.g. section 10.4.2.3 ; on perd par contre la vitesse $1/\sqrt{m}$).

REMARQUE 1 (Fonctions de coût et nombres de couverture) *Enfin, il convient de souligner que les nombres de couverture (comme la dimension VC) de $\mathcal{H}$ sont très liés à ceux de $L_{\mathcal{H}}$. En effet, si $c(y, y')$ est lipschitzien (ce qui est le cas de $c(y, y') = (y - y')^2$ si $\mathcal{Y} \subset [-1, 1]$), alors les nombres de couverture (respectivement avec crochets) de $L_{\mathcal{H}}$ sont majorés par ceux de $\mathcal{H}$.*

10.3.4. *Résultats asymptotiques*

Les résultats classiques comme la loi forte des grands nombres et le théorème central limite s'étendent aux classes de fonctions : la loi des grands nombres ($\frac{1}{m}\sum_{i=1}^{m} L_h(z_i) \rightarrow \mathbb{E}\ L_h$), dans sa version *uniforme* en $h \in \mathcal{H}$, est vraie lorsque $L_{\mathcal{H}}$ est une classe de Glivenko-Cantelli (section 10.3.4.1) et le théorème central limite ($\frac{1}{\sqrt{m}}\sum_{i=1}^{m} L_h(z_i)$ asymptotiquement gaussien) est vrai dans sa version *uniforme* en $h \in \mathcal{H}$ lorsque $L_{\mathcal{H}}$ est une classe de Donsker (section 10.3.4.2). Le bootstrap est une application très riche des classes de Donsker, permettant d'évaluer de manière asymptotiquement consistante différentes quantités dont le supremum sur $h \in \mathcal{H}$ des $|\hat{\mathbb{E}}L_h - \mathbb{E}L_h|$.

10.3.4.1. *Classes de Glivenko-Cantelli*

DÉFINITION 24 (Classes de Glivenko-Cantelli [DEH 71] ; voir aussi [DUD 91]) *Soit $L_{\mathcal{H}}$ une famille de fonctions de $\mathcal{X}$ dans $\mathbb{R}$ [3] et P une mesure de probabilité. Alors, on dit que $L_{\mathcal{H}}$ est une P-classe de Glivenko-Cantelli si pour toute suite i.i.d. de variables aléatoires de même loi P on a $\sup_{L_h \in L_{\mathcal{H}}} |\hat{\mathbb{E}}L_h - \mathbb{E}L_h| \overset{d}{\rightarrow} 0$, et on dit que $L_{\mathcal{H}}$ est une P-classe de Glivenko-Cantelli pour la convergence presque sûre si pour toute suite i.i.d. de variables aléatoires de même loi P on a $\sup_{L_h \in L_{\mathcal{H}}} |\hat{\mathbb{E}}L_h - \mathbb{E}L_h| \overset{p.s.}{\rightarrow} 0$. Si $\mathcal{P}$ est une famille de distributions de probabilité, on dit que $\mathcal{H}$ est une $\mathcal{P}$-classe de Glivenko-Cantelli (respectivement pour la convergence presque sûre) si $\mathcal{H}$ est une P-classe de Glivenko-Cantelli (respectivement pour la convergence presque sûre) pour tout $P \in \mathcal{P}$.*

Le nom de "classes de Glivenko-Cantelli" fait référence au théorème de Glivenko-Cantelli (voir par exemple [DEV 96, VAP 98]), première loi (forte) des grands nombres uniforme démontrée. Ce théorème établit, pour une variable aléatoire à valeurs réelles, la convergence presque sûre de la fonction de répartition empirique vers la vraie fonction de répartition, ceci uniformément sur $\mathbb{R}$. Il peut donc également être vu comme un résultat de convergence d'un processus empirique indexé par une famille de fonctions, ces fonctions étant très simples, puisqu'il s'agit de raies ($t \mapsto 1$

3. Il est à noter que $L_{\mathcal{H}}$ peut tout à fait être une famille de fonctions autre qu'une famille de fonctions de coût, même si nous adoptons cette notation dans un souci de cohérence avec le reste de ce chapitre.

si $t > a$, 0 sinon). Les classes de Glivenko-Cantelli sont donc des familles de fonctions pour lesquelles on dispose d'un théorème de Glivenko-Cantelli étendu. Un outil efficace pour prouver le caractère Glivenko-Cantelli d'une classe est le suivant ([BLU 55, DEH 71, DUD 84]) :

THÉORÈME 32 (Couverture avec crochets et classes de Glivenko-Cantelli) *Si $L_{\mathcal{H}}$ est une classe de fonctions telle que $N_{[\,]}(\epsilon, L_{\mathcal{H}}, L^1(P)) < \infty$ pour tout $\epsilon > 0$, alors $L_{\mathcal{H}}$ est P-Glivenko-Cantelli pour la convergence presque sûre.*

Les nombres de couverture en jeu sont donc "avec crochets" et concernent la distance L^1 associée à la distribution P. On trouvera une démonstration (simple) de ce théorème dans [VAA 96, chap. 2.4]. Un second théorème, d'énoncé et de démonstration plus complexes, trouvera un grand nombre d'applications notamment en raison de son lien avec la VC-théorie ; on va ici considérer les nombres de couverture pour la distance $L^1(\hat{P}_n)$.

THÉORÈME 33 (Couverture et classes de Glivenko-Cantelli) *Si $L_{\mathcal{H}}$ est borné* [4] *et si $\log N(\epsilon, L_{\mathcal{H}}, L^1(\hat{P}_n)) = o_P(n)$ pour tout $\epsilon > 0$, alors $L_{\mathcal{H}}$ est P-Glivenko-Cantelli pour la convergence presque sûre.*

Very related results can be found as early as [VAP 71, VAP 81].

Ce théorème permet donc de traiter des nombres de couvertures super-polynomiaux en $1/\epsilon$, donc très au délà de la VC-dimension finie. En effet, les nombres de couverture pour $L^1(\hat{P}_n)$ sont polynomiaux quand la VC-dimension est finie ; ils le sont même pour la distance $L^1(Q)$ pour toute distribution de probabilité discrète finie Q.

THÉORÈME 34 (Couverture et VC-dimension) *Soit $r \geq 1$. Si $L_{\mathcal{H}}$ est une famille de fonctions à valeurs dans $\{0, 1\}$, alors*

$$N(\epsilon, L_{\mathcal{H}}, L_r(Q)) \leq K(V+1)(4e)^{V+1}(1/\epsilon)^{rV}$$

pour toute distribution de probabilité discrète et finie Q, en notant V la VC-dimension de $L_{\mathcal{H}}$ et K est une constante universelle. Si $L_{\mathcal{H}}$ est une famille de fonctions à valeurs dans $[0, 1]$, alors

$$N(\epsilon, L_{\mathcal{H}}, L_r(Q)) \leq K'(V+1)(16e)^{V+1}(1/\epsilon)^{rV}$$

pour toute distribution de probabilité discrète et finie Q, V désignant de nouveau la VC-dimension de $L_{\mathcal{H}}$ et K' une constante universelle.

4. Il est en fait suffisant, sous certaines conditions, que $L_{\mathcal{H}}$ soit d'enveloppe L^1 pour P.

On pourra trouver une preuve de ces résultats dans [VAA 96, théorèmes 2.6.4 et 2.6.7] qui crédite [HAU 95a, DUD 78] de la preuve originale de 2.6.4 ; le passage à 2.6.7 (extension aux fonctions réelles) provient de [POL 84]. Par conséquent, toute famille de fonctions de VC-dimension finie ou de P-dimension finie est universellement Glivenko-Cantelli.

On verra en section 10.4.1 des applications des classes de Glivenko-Cantelli.

10.3.4.2. *Classes de Donsker*

Nous commencerons par définir l'intégrale d'entropie et la condition d'entropie.

DÉFINITION 25 *L'*intégrale d'entropie *est définie comme suit pour $\mathcal{F}$ un ensemble de fonctions d'enveloppe bornée :*

$$I_{eu} = \int_0^\infty \sup_Q \sqrt{\log N(\epsilon, \mathcal{F}, L^2(Q))} d\epsilon \qquad [10.44]$$

où le supremum porte sur l'ensemble des distributions de probabilité discrètes finies Q. La condition d'entropie uniforme *est $I_{eu} < \infty$. L'*entropie *est de manière générale le logarithme des nombres de couverture.*

La condition d'entropie uniforme est en particulier vérifiée lorsque la VC-dimension est finie (c.f. théorème 34).

Interprétation : D'une manière ou d'une autre on essaiera en général de se débarasser du supremum sur Q, peu commode à manier. Il s'agira donc, pour que la condition soit vérifiée, d'avoir des nombres de couverture qui ne croissent pas trop vite (le problème dans la convergence de l'intégrale est en 0 et non en ∞).

Des versions raffinées où $\mathcal{F}$ n'est pas d'enveloppe bornée existent ; voir [VAA 96, chapitre 2.5].

DÉFINITION 26 (Classes de Donsker, voir par exemple [VAA 96], page 81) *On considère une suite $(x_1, y_1), \ldots, (x_m, y_m)$ de variables aléatoires i.i.d. de loi P.*

Une famille de fonctions $L_\mathcal{H}$ est P-Donsker si $\sqrt{m}(\hat{\mathbb{E}}L_f - \mathbb{E}L_f)_{L_f \in L_\mathcal{H}}$ converge faiblement dans $L^\infty(L_\mathcal{H})$ vers un processus aléatoire tendu [5].

5. Un processus aléatoire est tendu si pour tout $\epsilon > 0$ il existe un quasi-compact de mesure $\geq 1 - \epsilon$.

La définition est un peu abstraite, mais à bien la regarder elle nous stipule simplement que l'on a bien convergence faible uniforme en $1/\sqrt{m}$. Elle laisse ouverte la définition de ce processus aléatoire limite, mais la proposition suivante comble ce manque :

PROPOSITION 6 *Si la condition de la définition ci-dessus est réalisée, alors le processus aléatoire limite est un processus gaussien centré, entièrement défini par ses variances/covariances qui sont les variances/covariances marginales ; i.e. avec G le processus limite de $\sqrt{m}(\hat{\mathbb{E}}L_f - \mathbb{E}L_f)$, l'espérance $\mathbb{E}(Gf) \times (Gg)$ est égale à $\mathbb{E}(fg) - (\mathbb{E}f)(\mathbb{E}g)$.*

Cette proposition n'est pas immédiate ; après avoir montré qu'il s'agit bien d'un processus gaussien et que ses covariances sont bien celles-ci, il faut établir qu'il est uniquement déterminé par cela. Le lecteur intéressé est renvoyé à [VAA 96, sections 1.5 et 2.1].

Le théorème ci-dessous est central en théorie du processus empirique et généralise le théorème central limite.

THÉORÈME 35 (Couverture et classes de Donsker [DUD 78, KOL 81, POL 82]) *Si $L_{\mathcal{H}}$ est d'enveloppe bornée et si la condition d'entropie uniforme est vérifiée, alors $L_{\mathcal{H}}$ est P-Donsker.*

Interprétation : Si la condition d'entropie uniforme est vérifiée, alors les déviations, indexées par les fonctions, et multipliées par $\sqrt{m}$, tendent vers un processus gaussien. En particulier, l'inégalité de Borell (proposition 2) s'applique ; on a par ailleurs convergence faible en $1/\sqrt{m}$ du supremum des déviations, ce qui explique donc les vitesses constatées.

Il est à noter que ces convergences en $1/\sqrt{m}$ ont donc lieu même à dimension VC infinie, tant que les nombres de couverture ne croissent pas trop vite (c.f. condition d'intégrale d'entropie uniforme plus haut : les nombres de couverture peuvent être légèrement exponentiels). On verra plus bas qu'en outre le caractère Donsker permet l'utilisation du boostrap.

Un théorème équivalent existe avec les nombres de couverture avec crochets ([DUD 78, DUD 84, OSS 87, AND 88]) :

THÉORÈME 36 (Nombres de couverture avec crochets et classes de Donsker) *Si $L_{\mathcal{H}}$ est à valeurs dans $[0,1]$ et si*

$$\int_0^\infty \sqrt{\log N_{[\,]}(\epsilon, L_{\mathcal{H}}, L^2(P))} d\epsilon < \infty$$

alors $L_{\mathcal{H}}$ est une classe de Donsker.

On pourra trouver une démonstration dans le chapitre 2.5 de [VAA 96].

Exemples

– pour toute distribution de probabilité P sur $\mathbb{R}$, $\{]-\infty, x]; x \in \mathbb{R}\}$ est P-Donsker (statistique de Kolmogorov-Smirnov);

– l'ensemble des fonctions α-hölderiennes dans un compact de $\mathbb{R}^d$ est Donsker pour une loi de probabilité de densité bornée dès lors que $\alpha > 2(d-1)$.

On verra en partie 10.4.1 des applications des classes de Donsker.

10.3.4.3. *Le Bootstrap*

Le Bootstrap a été initialement défini par Efron [EFR 79] et des versions uniformes sont apparues avec le travail de Giné et Zinn [GIN 90]. Ces travaux et beaucoup d'autres sont résumés dans [VAA 96] (voir aussi [HAS 02]).

On appelle rééchantillonnage bootstrap d'un échantillon D de taille m un tirage avec remise de m exemples parmi D. Le fait que le tirage soit avec remise fait que ces rééchantillonnages ne retombent pas nécessairement sur D.

On note dans la suite $G_f = \sqrt{m}(\hat{\mathbb{E}}L_f - \mathbb{E}L_f)$; G est une famille de réels indexée par $f \in \mathcal{H}$. Implicitement, G dépend de m. On note $G'_f = \sqrt{m}(\hat{\hat{\mathbb{E}}}L_f - \hat{\mathbb{E}}L_f)$, où $\hat{\hat{\mathbb{E}}}L_f$ est la moyenne des $L_f(Z_{i_k})$, où les i_k, pour $k \in [[1,m]]$, sont i.i.d. parmi $[[1,m]]$ (i.e., $\hat{\hat{\mathbb{E}}}$ est la distribution correspondant à un échantillon de m points tirés avec remise dans l'échantillon conduisant à $\hat{\mathbb{E}}$). G' dépend en fait à la fois du premier échantillon $z_1, \ldots, z_m$ et du rééchantillonnage opéré.

THÉORÈME 37 (Classes de Donsker et bootstrap) *Soit $L_{\mathcal{H}}$ une classes de Donsker de fonctions à valeurs dans $\mathbb{R}$, telle que $L_{\mathcal{H}}$ ait une fonction enveloppe de norme L^2 finie. Alors, G' tend vers G en convergence faible conditionnelle* [6].

6. Cette convergence est à formuler proprement en terme de constante de Lipschitz :

$$\sup_{h \in BL(1)} |\mathbb{E}_{bs} h(G') - h(G)| \overset{p.s.}{\to} 0$$

En outre, le résultat demeure vrai si on remplace G' par son estimateur empirique défini comme la moyenne de B échantillons bootstrap indépendants pourvu que $B \to \infty$ comme $m \to \infty$.

Interprétation : On peut approcher G (inconnu) par G' (simulable) du point de vue de leur supremum dans $L^\infty(L_\mathcal{H})$.

Le résultat se transmet à l'approximation de $h(G)$ par $h(G')$ si h est suffisamment régulière. En particulier, cela est vrai lorsque $h(.)$ est Hadamard-différentiable en G ([VAA 96, chap. 3.9]).

Le bootstrap fournit donc des informations sur le processus limite, dans un cadre très général. Il peut ainsi être utilisé de différentes manières, que ce soit la construction d'intervalles de confiance ou la suppression de biais. Le lecteur est renvoyé à [VAA 96, section 3] ou à des livres standard de statistique ([SAP 90]) ; on pourra consulter aussi [DEV 96] pour les limites de la méthode.

Il est à noter que d'autres méthodes que le bootstrap permettent d'étudier le processus limite, en évitant son caractère asymptotique. En première approximation, on voit que le bootstrap permet d'approcher la loi de $\sup(\mathbb{E}L_f - \hat{\mathbb{E}}L_f)$:

– par une quantité que l'on peut évaluer par des rééchantillonnages ;

– à une quantité asympotiquement négligeable près ;

– d'une manière qui dépend des données (les rééchantillonnages dépendent de $\hat{P}$).

On verra en section 10.5.1 une méthode (méthode de Rademacher) permettant d'approcher ce même supremum :

– par une quantité que l'on peut évaluer par des rééchantillonnages, plus simples car il s'agit de rééchantillonner des variables binaires, mais coûtant l'évaluation d'un supremum bien délicat à chaque fois,

– à une quantité asymptotiquement négligeable près là-aussi, qui est en outre bornée non-asymptotiquement ;

– de manière dépendant des données (les rééchantillonnages sont binaires, mais le calcul de $\sup$ qui en découle utilise les données).

Enfin, on peut citer des applications étonnantes des méthodes de rééchantillonnage. Les améliorations de précision précédemment citées (qui sont en fait valables

où $BL(1) = \{h \in [0,1]^{L^\infty(L_\mathcal{H})}; h \text{ est } 1-\text{lipschitzien}\}$ et bs est la mesure de probabilité associée au rééchantillonnage bootstrap. Ceci est équivalent à la convergence faible, à ceci près que G' ne peut à proprement parler être défini en termes de convergence faible.

dans un cadre très général d'où l'immense succès du bootstrap dans la réduction de biais), ont encouragé les gens à tester les rééchantillonnages pour apprendre et pas seulement pour valider, et cela a donné naissance au bagging [BRE 96], au boosting [SCH 98], et de manière générale aux méthodes d'ensemble. L'idée est d'avoir un grand nombre de rééchantillonnages des données, d'apprendre sur ces rééchantillonnages, et de combiner les fonctions obtenues.

10.4. Les paradigmes

La solution la plus naturelle pour choisir une fonction dans $\mathcal{H}$ consiste à choisir celle qui a la plus faible erreur moyenne sur les exemples (section 10.4.1). Nous verrons que cela fonctionne notamment en dimension VC finie ; notons cependant que si l'on n'exige pas l'uniformité en la distribution cela fonctionne aussi avec des nombres de couverture légèrement exponentiels, cf 10.3.4.1. Le cas de la dimension VC finie représente toutefois un cadre restrictif 10.4.2.2, et la section 10.4.2 fournit donc un paradigme plus élaboré, garantissant notamment la *consistance universelle* 10.4.2.1. Nous verrons enfin en section 10.4.3 la *minimisation structurelle du risque*, qui justifie notamment les techniques de régularisation.

10.4.1. *Minimisation empirique du risque*

On sait que toute la difficulté (statistique) de l'apprentissage provient du fait que l'expression analytique de la probabilité jointe caractérisant le problème à traiter est inconnue, et que l'on a accès à cette mesure qu'à travers un m-échantillon. Sur cet échantillon, il est possible de calculer une estimation empirique du risque, donnée par :

$$R_m(h) = \frac{1}{m} \sum_{i=1}^{m} \ell\left(h, (x_i, y_i)\right). \qquad [10.45]$$

Le principe inductif de *minimisation empirique du risque* (MER) [VAP 82] consiste à utiliser le risque empirique comme critère de sélection de fonction. Pour que la mise en œuvre de ce principe soit justifiée, il faut disposer d'un ensemble de résultats que nous allons à présent évoquer. Pour ce faire, nous nous appuyons de manière privilégiée sur le chapitre 3 de [VAP 98] et sur [GAS 97]. En premier lieu, nous faisons l'hypothèse que, pour un m-échantillon donné, il existe une fonction h_m vérifiant $R_m(h_m) = inf_{h \in \mathcal{H}} R_m(h)$, fonction que nous nommons *fonction empirique*. Si la théorie peut encore être développée sous des hypothèses plus légères (voir par exemple [BOU 02]), cette hypothèse présente l'avantage de simplifier l'exposé.

Le risque empirique est l'estimateur de resubstitution du risque. Si l'on s'intéresse plus particulièrement à son minimum, il s'agit naturellement d'un estimateur fortement biaisé donnant une vue optimiste des performances réelles.

Il n'est donc pas évident, même si $R_m(h_m)$ est faible, que $R(h_m)$ soit faible. Une solution simple et naturelle, pour le garantir, est d'utiliser une convergence uniforme.

REMARQUE 2 *Si* $\sup_{h\in\mathcal{H}} |R_m(h) - R(h)| < \epsilon$ *et* $h_m = \arg\min R_m(.)$, *alors*

$$R(h_m) \leq R_m(h_m) + \epsilon \leq \inf_h R_m(h) + \epsilon \leq \inf_{h\in\mathcal{H}} R(h) + 2\epsilon.$$

Nous allons ici utiliser cette simple observation pour nous éloigner brièvement de ce qui se fait usuellement en théorie statistique de l'apprentissage, et passer par les classes de Glivenko-Cantelli et les classes de Donsker.

Si $L_\mathcal{H}$ est Glivenko-Cantelli pour la convergence presque sûre, alors $R(\arg\min R_m) = \mathbb{E}L_{\arg\min f\mapsto \hat{E}L_f} \to \inf_{L_f\in L_\mathcal{H}} \mathbb{E}L_f$ presque sûrement. On a ainsi une forme de consistance de la minimisation empirique du risque dans $\mathcal{H}$. Les classes de VC-dimension finie étant universellement Glivenko-Cantelli (théorème 33), nous avons ainsi un résultat de consistance indépendant de la distribution.

Par ailleurs, si $L_\mathcal{H}$ est Donsker, alors $\sup_{h\in\mathcal{H}} |R_m(h) - R(h)|$ converge faiblement vers 0 en $1/\sqrt{m}$. Par conséquent, $R(\arg\min R_m) = \mathbb{E}L_{\arg\min f\mapsto \hat{E}L_f}$ converge faiblement vers $\inf R(.) = \inf_{L_f\in L_\mathcal{H}} \mathbb{E}L_f$ en $1/\sqrt{m}$. Les classes de VC-dimension finie étant universellement Donsker (voir section 10.3.4.2), cette propriété est vraie pour toute distribution, pour toute classe de VC-dimension finie.

Nous venons donc de montrer que la finitude de la VC-dimension permet d'obtenir des convergences saines, valables pour toute distribution, et que des hypothèses plus faibles permettent des résultats dépendants de la distribution, soit avec soit sans vitesse en $1/\sqrt{m}$, même avec VC-dimension finie. L'approche ci-dessus est notamment développée dans [VAA 96]. Son inconvénient principal est de ne pas exhiber de bornes non-asymptotiques ; on ne peut énoncer un résultat sous la forme « avec probabilité au moins $1-\delta$, $R(h_m) \leq \inf_{h\in\mathcal{H}} R(h) + \epsilon$ ». Nous allons maintenant suivre une démarche plus classique en théorie statistique de l'apprentissage, qui nous permettra d'expliciter les ϵ et les δ ci-dessus souhaités, en fonction de m.

Nous avons vu qu'il semble souhaitable, pour que l'application du principe inductif MER ait un sens, qu'il y ait une convergence de $R_m(h_m)$ vers $R(h_m)$; ainsi, a minima, on saura estimer le risque $R(h_m)$ à partir du risque constaté $R_m(h_m)$. Nous nous intéressons ici à une convergence en probabilité. On souhaite donc établir :

$$\lim_{m\longrightarrow\infty} P^m \{|R(h_m) - R_m(h_m)| \geq \epsilon\} = 0, \qquad [10.46]$$

où la probabilité produit P^m est relative au m-échantillon d'apprentissage (tiré suivant la loi P). Lorsque la propriété 10.46 est vérifiée, on parle de *convergence faible* du processus d'apprentissage. On peut également caractériser l'*erreur d'estimation*, c'est-à-dire l'écart entre $R(h_m)$ et $\inf_{h\in\mathcal{H}} R(h)$. La convergence :

$$\lim_{m\longrightarrow\infty} P^m \left\{ R(h_m) - \inf_{h\in\mathcal{H}} R(h) \geq \epsilon \right\} = 0, \qquad [10.47]$$

caractérise la *convergence forte* du processus d'apprentissage. Remarquons qu'ici, nous ne nous soucions pas de l'existence d'une fonction minimisant le risque sur $\mathcal{H}$. Le problème fondamental de la théorie du principe inductif MER réside dans la démonstration de la *consistance*. Celle-ci est définie de la manière suivante.

DÉFINITION 27 (Consistance du principe inductif MER) *Le principe inductif MER est dit* consistant *pour la famille de fonctions $L_{\mathcal{H}}$, et pour le problème d'apprentissage caractérisé par la loi jointe P, si l'on dispose des deux résultats de convergence en probabilité :*

$$\begin{cases} \lim_{m\longrightarrow\infty} P^m \left\{ R(h_m) - \inf_{h\in\mathcal{H}} R(h) \geq \epsilon \right\} = 0 \\ \lim_{m\longrightarrow\infty} P^m \left\{ |R_m(h_m) - \inf_{h\in\mathcal{H}} R(h)| \geq \epsilon \right\} = 0 \end{cases} .$$

On observe que la première condition correspond à la convergence forte du processus d'apprentissage, tandis que la seconde correspond à la convergence du risque empirique minimal vers le plus petit des risques atteignable sur $\mathcal{H}$. Olivier Gascuel [GAS 97] donne une signification différente à la consistance, en la définissant comme la convergence, toujours en probabilité, du risque de la fonction minimisant le risque empirique vers le plus petit risque possible, c'est-à-dire le risque de Bayes. Cette vision ambitieuse des choses requiert de la part de la famille de fonctions $\mathcal{H}$ d'être un approximateur universel (ou de supposer l'appartenance d'un classifieur de Bayes à $\mathcal{H}$). De fait, des résultats de ce type ont été établis pour les PMC (voir par exemple [CYB 89], théorème 2 ou [HOR 89, BAR 93]) et les machines à noyau [STE 01]. Cependant, dans le cas des PMC, ils supposent que la taille de la couche cachée puisse être choisie arbitrairement grande. Chercher à faire converger le risque de la fonction empirique vers le risque de Bayes n'a donc généralement pas beaucoup de sens en pratique, sauf si l'on s'autorise à travailler sur des classes emboîtées de fonctions de complexités croissantes. Ce cadre n'est cependant plus celui du principe inductif MER, mais celui du principe inductif de minimisation structurelle du risque, traité plus bas. Pour cette raison, nous retenons ici la définition plus limitée de Vapnik. La définition de la consistance étant donnée, se pose à présent la question de la caractérisation

de conditions sous lesquelles elle est obtenue. On souhaite que ces conditions portent sur la famille de fonctions $\mathcal{H}$, ou plus précisément la famille de fonctions $L_{\mathcal{H}}$. Déterminer de telles conditions n'est possible qu'en considérant une notion de consistance légèrement différente.

DÉFINITION 28 (Consistance stricte du principe inductif MER) *Le principe inductif MER est dit* strictement consistant, *ou* consistant de manière non triviale *pour la famille de fonctions $L_{\mathcal{H}}$, et pour le problème d'apprentissage caractérisé par la loi jointe P, si pour tout sous ensemble non vide $\mathcal{H}_c$ de $\mathcal{H}$, indexé par un réel c et défini par*

$$\mathcal{H}_c = \left\{ h \in \mathcal{H} \ / \ R(h) = \int_{\mathcal{X}\times\mathcal{Y}} \ell\left(h,(x,y)\right) dP(x,y) \geq c \right\}$$

on dispose de la convergence en probabilité :

$$\lim_{m\longrightarrow\infty} P^m \left\{ \left| \inf_{h\in\mathcal{H}_c} R_m(h) - \inf_{h\in\mathcal{H}_c} R(h) \right| \geq \epsilon \right\} = 0. \qquad [10.48]$$

Vapnik a démontré qu'une condition nécessaire et suffisante de consistance stricte du principe inductif MER est la propriété de convergence uniforme sur $\mathcal{H}$, "d'un seul côté", des moyennes vers leurs espérances. La convergence uniforme se définit sous deux formes.

DÉFINITION 29 (Convergence unifome) *Pour un problème d'apprentissage et une famille de fonctions donnés, on parle de* convergence uniforme des moyennes vers leurs espérances ("des deux côtés") *si l'on dispose de la convergence en probabilité suitante :*

$$\lim_{m\longrightarrow\infty} P^m \left\{ \sup_{h\in\mathcal{H}} |R(h) - R_m(h)| \geq \epsilon \right\} = 0. \qquad [10.49]$$

DÉFINITION 30 (Convergence unifome "d'un seul côté") *Pour un problème d'apprentissage et une famille de fonctions donnés, on parle de* convergence uniforme des moyennes vers leurs espérances "d'un seul côté" *si l'on dispose de la convergence en probabilité suitante :*

$$\lim_{m\longrightarrow\infty} P^m \left\{ \sup_{h\in\mathcal{H}} \left(R(h) - R_m(h)\right) \geq \epsilon \right\} = 0. \qquad [10.50]$$

THÉORÈME 38 (Conditions de consistance stricte du principe inductif MER) *Pour que l'on dispose de la consistance stricte du principe inductif MER, il est nécessaire et suffisant que la convergence uniforme "d'un seul côté" ait lieu.*

La raison pour laquelle la consistance stricte est caractérisée par une condition dissymétrique, la convergence uniforme "d'un seul côté", résulte du fait que nous nous intéressons seulement à la minimisation empirique du risque, et non à la maximisation empirique du risque.

Il résulte du théorème 38 que le problème de la caractérisation de la consistance stricte se réduit à celui de la caractérisation de la convergence uniforme. Dans le chapitre 3 de [VAP 98], Vapnik fournit des conditions nécessaires et suffisantes de convergence uniforme "des deux côtés" en fonction des propriétés de la famille de fonctions $L_{\mathcal{H}}$ (voir aussi [GIN 84]). Ces conditions s'expriment toutes comme la convergence vers zéro du quotient de l'*entropie* (ou l'*ϵ-entropie*) de la famille des fonctions $L_{\mathcal{H}}$ par la taille de l'échantillon d'apprentissage, lorsque cette taille tend vers l'infini. Afin de conserver à cette exposé sa simplicté, nous ne donnons que la définition de l'entropie utilisée lorsque ℓ est une fonction indicatrice, c'est-à-dire que le problème traité est une tâche de discrimination.

DÉFINITION 31 (Entropie d'une famille de fonctions indicatrices) *Soit $\mathcal{H}$ une famille de fonctions h de $\mathcal{X}$ dans $\mathcal{Y}$ et ℓ la fonction de perte indicatrice associée (pour une fonction h et un couple (x, y) de $\mathcal{X} \times \mathcal{Y}$ donnés). Soit $\mathcal{Z}_m$ un élément de $(\mathcal{X} \times \mathcal{Y})^m$. En nous inspirant des notations utilisées pour définir la fonction de croissance, notons $\Pi_{L_{\mathcal{H}}}(\mathcal{Z}_m)$ le cardinal de $L_{\mathcal{H}}|_{\mathcal{Z}_m}$. Alors l'*entropie aléatoire *de la famille de fonctions $L_{\mathcal{H}}$ sur $\mathcal{Z}_m$ est donnée par :*

$$H_{L_{\mathcal{H}}}(\mathcal{Z}_m) = \ln\left(\Pi_{L_{\mathcal{H}}}(\mathcal{Z}_m)\right). \qquad [10.51]$$

*On définit également l'*entropie *de la famille des fonctions $L_{\mathcal{H}}$ sur un échantillon de taille m comme étant :*

$$H_{L_{\mathcal{H}}}(m) = \int_{(\mathcal{X}\times\mathcal{Y})^m} H_{L_{\mathcal{H}}}(\mathcal{Z}_m) dP^m\left(\mathcal{Z}_m\right) = \mathbb{E}\left(\ln\left(\Pi_{L_{\mathcal{H}}}(\mathcal{Z}_m)\right)\right). \qquad [10.52]$$

Dans le cas où la fonction de perte est une indicatrice, une condition nécessaire et suffisante de convergence uniforme "des deux côtés" est donc donnée par le théorème suivant.

THÉORÈME 39 *Soit $\mathcal{H}$ une famille de fonctions h de $\mathcal{X}$ dans $\mathcal{Y}$. Si ℓ est une fonction indicatrice, alors une condition nécessaire et suffisante de convergence "des deux côtés" du processus d'apprentissage est donnée par :*

$$\lim_{m \longrightarrow \infty} \frac{H_{L_{\mathcal{H}}}(m)}{m} = 0. \qquad [10.53]$$

Notons $\mathcal{Z}_m = ((x_i, y_i))_{1 \leq i \leq m}$ et $\mathcal{X}_m = (x_i)_{1 \leq i \leq m}$. Dans le cas où les fonctions de la classe $\mathcal{H}$ sont à valeurs binaires, on établit simplement l'égalité

$$\Pi_{L_{\mathcal{H}}}(\mathcal{Z}_m) = \Pi_{\mathcal{H}}(\mathcal{X}_m)$$

dont on déduit immédiatement

$$H_{L_{\mathcal{H}}}(m) \leq \ln\left(\Pi_{\mathcal{H}}(m)\right).$$

Nous avons donc ici une nouvelle illustration du lien fort existant entre les mesures de capacité de la famille de fonctions représentables par un modèle et les mesures de capacité des classes de fonctions de perte correspondantes. On remarquera que les conditions nécessaires et suffisantes de convergence uniforme "des deux côtés" permettent en fait d'obtenir un résultat plus fort que celui recherché, dans la mesure où elles établissent une convergence presque sûre du processus empirique étudié, au lieu d'une simple convergence en probabilité. La condition nécessaire et suffisante de convergence uniforme "d'un seul côté" est donnée par le théorème suivant :

THÉORÈME 40 (Théorème 3.8 dans [VAP 98]) *Soit $\mathcal{H}$ une famille de fonctions h de $\mathcal{X}$ dans $\mathcal{Y}$. Pour que la convergence uniforme "d'un seul côté" ait lieu sur une famille de fonctions $L_{\mathcal{H}}$ uniformément bornées, il est nécessaire et suffisant que pour tout triplet de réels positifs $(\delta, \epsilon, \zeta)$, il existe une famille de fonctions $L^*_{\bar{\mathcal{H}}} = \left\{\ell^*\left(\bar{h}, .\right)\right\}$, avec $\bar{h} \in \bar{\mathcal{H}}$, telle que :*

1) pour toute fonction $\ell\left(h, .\right)$ il existe une fonction $\ell^*\left(\bar{h}, .\right)$ satisfaisant les conditions

$$\ell\left(h, .\right) \geq \ell^*\left(\bar{h}, .\right)$$

et

$$\int_{\mathcal{X} \times \mathcal{Y}} \left(\ell\left(h, (x, y)\right) - \ell^*\left(\bar{h}, (x, y)\right)\right) dP(x, y) < \zeta \ ;$$

2) l'ϵ-entropie de la famille de fonctions $L^*_{\bar{\mathcal{H}}}$ satisfait l'inégalité :

$$\lim_{m \longrightarrow \infty} \frac{H_{L^*_{\bar{\mathcal{H}}}}(\epsilon, m)}{m} < \delta. \qquad [10.54]$$

Dans cette section, nous n'avons jusqu'à présent étudié le principe inductif de minimisation empirique du risque que d'un point de vue qualitatif. Il s'agissait de savoir quand la mise en œuvre de ce principe était pertinente, sans se soucier de considérations pratiques, comme la vitesse de convergence. Cependant, la seconde question est naturellement aussi importante que la première. Pour l'aborder, il convient au préalable d'introduire plusieurs notions supplémentaires. On considèrera que l'on dispose d'un taux de convergence asymptotique rapide pour un problème donné (caractérisé par la loi de probabilité P), s'il existe deux constantes positives b et c telles que, pour une taille m de l'échantillon d'apprentissage suffisamment grande, on ait

$$P^m \left\{ \sup_{h \in \mathcal{H}} |R(h) - R_m(h)| \geq \epsilon \right\} < b \exp\left(-c\epsilon^2 m\right). \quad [10.55]$$

m, taille d'échantillon assurant une précision donnée ϵ, est à choisir en fonction de P et $\mathcal{H}$, qui spécifient le problème d'apprentissage, ainsi que ϵ. De la même manière, la convergence rapide, non plus pour une unique mesure de probabilité P, mais pour une classe de mesures de probabilité $\mathcal{P}$ est caractérisée par

$$\sup_{P \in \mathcal{P}} P^m \left\{ \sup_{h \in \mathcal{H}} |R(h) - R_m(h)| \geq \epsilon \right\} < b \exp\left(-c\epsilon^2 m\right). \quad [10.56]$$

m, taille d'échantillon assurant une précision donnée ϵ, n'étant plus cette fois à déterminer qu'en fonction de $\mathcal{H}$ et ϵ.

Notons que la queue de distribution obtenue pour $\sup_{h \in \mathcal{H}} |R(h) - R_m(h)|$ est d'ordre $\exp(-c\epsilon^2 m)$, et qu'elle est donc presque aussi faible que la queue d'une gaussienne.

Notons aussi que si la condition 10.55 est vérifiée, alors naturellement par la remarque 2,

$$P^m \left\{ R(h_m) \geq \inf_{h \in \mathcal{H}} R(h) + 2\epsilon \right\} < b \exp(-c\epsilon^2 m).$$

Cela montre donc que la condition 10.55 entraîne une convergence rapide de $R(h_m)$ vers l'erreur minimale. On voit donc le caractère central de la condition 10.55. Le résultat est en outre indépendant de la distribution, dès lors que la propriété 10.56 est vérifiée. Déterminer des conditions nécessaires et suffisantes pour la propriété 10.56 est donc central.

DÉFINITION 32 (Entropie recuite) *Soit $\mathcal{H}$ une famille de fonctions h de $\mathcal{X}$ dans $\mathcal{Y}$. Alors, en conservant les notations précédentes, l'*entropie recuite *de la famille de fonctions $L_{\mathcal{H}}$ est définie comme :*

$$H_{ann, L_{\mathcal{H}}}(m) = \ln\left(\mathbb{E}\left(\Pi_{L_{\mathcal{H}}}(\mathcal{Z}_m)\right)\right). \quad [10.57]$$

Comme dans le cas de l'entropie, cette notion s'étend au cas des fonctions de perte qui ne sont pas des indicatrices ; on parle alors d'ϵ-*entropie* recuite. Ces définitions étant posées, on dispose des résultats suivants.

THÉORÈME 41 *Soit $\mathcal{H}$ une famille de fonctions h de $\mathcal{X}$ dans $\mathcal{Y}$. Si ℓ est une fonction indicatrice, alors une condition suffisante de consistance et de convergence rapide du processus d'apprentissage au sens de 10.55 est donnée par :*

$$\lim_{m \longrightarrow \infty} \frac{H_{ann, L_{\mathcal{H}}}(m)}{m} = 0. \qquad [10.58]$$

THÉORÈME 42 (Condition nécessaire et suffisante) *Soit $\mathcal{H}$ une famille de fonctions h de $\mathcal{X}$ dans $\mathcal{Y}$. Si ℓ est une fonction indicatrice, alors une condition nécessaire et suffisante de consistance et de convergence rapide du processus d'apprentissage au sens de 10.56, i.e. pour toute mesure de probabilité est donnée par :*

$$\lim_{m \longrightarrow \infty} \frac{\ln\left(\Pi_{L_{\mathcal{H}}}(m)\right)}{m} = 0. \qquad [10.59]$$

Vapnik nomme les théorèmes 39, 41 et 42 les trois étapes importantes (*milestones*) de la théorie statistique de l'apprentissage.

Pour conclure quant au célèbre lien entre apprenabilité et VC-dimension, on utilie le lemme de Sauer qui assure que $\Pi_{L_{\mathcal{H}}}(m)$ est polynomial en m pour m suffisamment grand si et seulement si la VC-dimension est finie, et qu'en cas contraire $\Pi_{L_{\mathcal{H}}}(m) = 2^m$ pour tout m. On en déduit alors le résultat suivant de Vapnik et Chervonenkis ([VAP 71, VAP 81, VAP 98]) :

THÉORÈME 43 *Dans le cadre d'une discrimination bi-classe, une condition nécessaire et suffisante de consistance et de convergence rapide du processus d'apprentissage au sens de 10.56, i.e. pour toute mesure de probabilité, est donnée par :*

$$\textit{VC-dim}(L_{\mathcal{H}}) < \infty, \ \textit{ou de manière équivalente VC-dim}(\mathcal{H}) < \infty.$$

REMARQUE 3 *Il est à noter que la VC-dimension ne caractérise pas seulement le fait que l'on ait convergence pour toute distribution au sens de 10.56. En outre, cette condition est équivalente au fait qu'il existe une vitesse de convergence indépendante de la distribution ; on n'obtiendrait pas, en VC-dimension infinie, une vitesse de convergence indépendante de la distribution, quelle que soit l'algorithme retenu. Ceci est établi par le théorème 47.*

Les résultats correspondant dans le cas de la régression sont beaucoup plus récents. [ALO 97] montre que, pour la fonction de coût quadratique $c(x, y, y') = (y - y')^2$ avec $\mathcal{Y}$ borné, l'existence d'une convergence uniforme de R_m vers R, indépendante de la distribution, est équivalente à la finitude de la fat-shattering dimension P_γ-dim$(\mathcal{H})$ pour tout $\gamma > 0$.

On peut, tout comme en VC-dimension pour la classification, dériver des bornes sur l'écart $\sup_{h\in\mathcal{H}} |R(h) - R_m(h)|$ en régression à partir de la fat-shattering dimension. Pour cela, on peut borner par le théorème 28 les nombres de couverture $\sup_{E;\#E\leq m} N(\epsilon, \mathcal{H}_{|E}, L_\infty)$. On sait que

$$\sup_{E;\#E\leq m} N(\epsilon, \mathcal{H}_{|E}, L_\infty) \geq N_m(\epsilon, \mathcal{H}) \qquad [10.60]$$

où $N_m(\epsilon, \mathcal{H}) = \sup_{Q=\frac{1}{m}\sum_{i=1}^m \delta_{t_i}} N(\epsilon, \mathcal{H}, L^1(Q))$. Or ces nombres de couverture sont fort utiles :

THÉORÈME 44 (Théorème 17.1 dans [ANT 99]) *Si* $c(x, y, y') = (y-y')^2$ *and* $\mathcal{Y} = [0, 1]$,

$$\sup_{h\in\mathcal{H}} |R_m(h) - R(h)| \leq \epsilon \qquad [10.61]$$

avec probabilité au moins $1 - 4N_{2m}(\epsilon/16, \mathcal{H}) \exp(-m\epsilon^2/32)$.

On en déduit donc immédiatement, en combinant les équations 10.60 et 10.61 une borne sur $\sup_h |R_m(h) - R(h)|$ en fonction de la dimension fat-shattering . Nous avons vu précédemment (théorèmes 26 et 27) que la dimension fat-shattering pouvait être finie même avec des réseaux arbitrairement grands ; ainsi, nous avons ici une justification des performances en généralisation de réseaux de très grande taille.

On peut remarquer certaines différences avec le cas de la classification.

En classification, la caractérisation de la convergence uniforme indépendante de la distribution reste la même si l'on se restreint aux distributions telle que l'erreur minimale est nulle ; c'est-à-dire que si on est assuré de l'existence de h vérifiant $R_m(h) = 0$, alors $\sup_{h;R_m(h)=0} |R(h) - R_m(h)| \to 0$ avec une vitesse indépendante de la distribution si et seulement si la VC-dimension est finie, de même que sans cette hypothèse sur l'existence de tels h, on a $\sup_h |R(h) - R_m(h)| \to 0$.

Dans le cas de la régression des cas pathologiques apparaissent lorsque l'erreur minimale est nulle (voir [ANT 99, exemple 19.6]) : lorsque les exemples sont sans bruit, il peut arriver que pour toute distribution d'exemples, un exemple suffise à déterminer exactement la fonction adéquate, bien que la fat-shattering dimension soit infinie.

10.4.2. *Minimisation empirique du risque corrigée*

La minimisation empirique du risque (MER) est connue efficace seulement dans le cas de la dimension VC finie. Toutefois, le résultat suivant de [DEV 96, page 290], basé sur un choix judicieux de la taille de la famille de fonctions, justifie la minimisation empirique du risque sur une union dénombrable de familles de dimension VC finie.

Nous présentons dans cette partie :

– ce principe d'induction sur une union dénombrable de familles de dimension VC finie 10.4.2.1 ;

– un résultat montrant que les mêmes performances ne peuvent être obtenues à partir d'une seule famille de fonctions de dimension VC finie 10.4.2.2 ;

– un résultat montrant que d'autres alternatives à la minimisation empirique du risque existent 10.4.2.3.

10.4.2.1. *Consistance universelle par minimisation incrémentale du risque empirique*

THÉORÈME 45 (Risque empirique incrémental ; théorème 18.1 dans [DEV 96]) *Soient $\mathcal{H}_1$, …, $\mathcal{H}_k$ des familles de fonctions dont les VC-dimensions V_1, …, V_k, … sont toutes finies. Soit $\mathcal{H} = \cup_n \mathcal{H}_n$. Supposons que chaque distribution conduise à $\inf L_{\mathcal{H}} = L^*$ (de telles classes de fonctions existent ; c.f. [DEV 96, ch.18] et les références qui y sont citées). Soit $k_m \to \infty$ et $V_{k_m}\ log(m)/m \to 0$ lorsque $m \to \infty$. Alors, pour toute distribution d'exemples, l'algorithme consistant à minimiser l'erreur empirique sur $\mathcal{H}_{k_m}$ a une erreur asymptotique égale à L^* avec probabilité* 1.

Interprétation : Dimensionnez votre nombre de paramètres en fonction de votre nombre d'exemples et tout devrait bien se passer.

REMARQUE 4 *– Ce paradigme sera appelé, par la suite, MER "incrémentale" - MER^I en abrégé. Aucune vitesse de convergence n'est donnée, et aucune vitesse de convergence ne peut être donnée uniformément en le classifieur Bayésien sous-jacent ("le" est un raccourci impropre de vocabulaire - le classifieur de Bayes n'est pas unique). Ceci peut être vérifié dans le théorème ci-dessous (dû à Benedek, Itai, [BEN 94]), qui établit que les classes VC peuvent être de mauvais approximateurs :*

– MER^I est très aisément utilisé dans des cas pratiques, contrairement à MER^I_C (présenté en section 10.4.2.3 et nécessitant l'utilisation d'ϵ-réseaux). Il conduit à la consistance universelle, mais à des problèmes algorithmiques NP-complets dans beaucoup de cas.

10.4.2.2. *Mauvaise approximation en dimension VC finie*

Nous allons maintenant signaler que l'on a besoin d'une famille incrémentale comme précédemment définie, car une seule famille de VC-dimension finie ne peut suffire.

THÉORÈME 46 (Mauvaise approximation des classes VC [BEN 94, DEV 96]) *Soit $\mathcal{H}_1, \ldots, \mathcal{H}_k, \ldots$ une suite de familles de fonctions à valeurs dans $\{0,1\}$ dont les dimensions VC $V_1, \ldots, V_k, \ldots$ sont finies. Soit $a_1, a_2, \ldots$ une suite de nombres réels décroissant vers zéro. Alors il existe une distribution de probabilité sur $\mathcal{X} \times \{0,1\}$ telle que pour k suffisamment grand,* $\inf L_{\mathcal{H}_k} - L^* > a_k$.

Ainsi, la convergence fournie par le théorème 45 (minimisation du risque empirique sur des espaces de fonctions de complexité "adaptée" au nombre d'exemples) peut être arbitrairement lente. On peut se demander s'il est possible de faire mieux, et si oui comment. Un résultat négatif fondamental est le suivant (théorème 14.4. dans [DEV 96]) :

THÉORÈME 47 (Borne inférieure de vitesse de convergence) *Considérons un algorithme qui à un échantillon associe un classifieur à valeurs dans $\{0,1\}$. Considérons une suite $\frac{1}{16} \geq a_0 \geq a_1 \geq a_2 \geq a_3 \geq \ldots$. Alors, il existe une distribution telle que $L^* = 0$ et $\forall n \in \mathbb{N}$, l'erreur en généralisation L_n a son espérance minorée comme suit :*

$$\mathbb{E}_{P^m}\ L_n \geq a_n$$

(où P^m est la mesure de probabilité associée à l'échantillon).

Interprétation : Sans hypothèse sur la distribution, on ne peut faire grand-chose sans avoir du temps devant soi.

Nous verrons plus bas que des hypothèses efficaces sont :

– le fait que l'erreur minimale soit nulle (forte accélération de la vitesse de convergence - voir section 10.4.1) ; on passe alors de vitesses du type $O(1/\sqrt{m})$ à des vitesses du type $O(1/m)$;

– le fait que L^* soit atteint par une famille union dénombrable de familles de dimension VC finie (*atteint*, et pas seulement *approché*), via la minimisation structurelle du risque ;

– des hypothèses sur la distribution, éventuellement marginalement (en X).

10.4.2.3. *Améliorer les résultats en ajoutant des hypothèses sur la distribution*

Considérons maintenant des résultats (non-uniformes sur l'ensemble de toutes les distributions) à propos de convergences rapides. Premièrement, certains résultats positifs sont possibles, au-delà de la VC-dimension finie, supposant que la loi de X est connue. Nous avons déjà vu en section 10.3.4 que des convergences étaient possibles en VC-dimension infinie, éventuellement pour toute distribution, mais avec des vitesses distribution-dépendantes. Voyons maintenant qu'avec des hypothèses sur la

distribution, on peut retrouver des vitesses (non-asymptotiques) en VC-dimension infinie. Considérons un espace totalement borné (i.e. de nombres de couverture finis pour tout ϵ) de fonctions à valeurs dans $[0,1]$. Alors avec Y à valeurs dans $\{0,1\}$ [7] :

THÉORÈME 48 (Minimisation incrémentale du risque empirique [KOL 61]) *Considérons $\mathcal{H}$ une famille de fonctions à valeurs dans $[0,1]$, avec des nombres de couverture pour L^1 $N(\epsilon)$ (pour la distribution marginale en X !). Considérons $\mathcal{H}_\epsilon$ un ϵ-réseau fini pour ϵ de taille $N(\epsilon)$. Considérons un algorithme minimisant le risque empirique pour L^2 parmi $\mathcal{H}_{\epsilon_m}$:*

$$\eta = argmin_{f \in \mathcal{H}_{\epsilon_m}} \hat{\mathbb{E}}(f(X) - Y)^2$$

$$g_\eta(x) = 1 \text{ si } \eta(x) > \frac{1}{2} \text{ et } g_\eta(x) = 0 \text{ sinon}$$

$$L_m = \mathbb{E}\chi_{g_\eta(X) \neq Y}$$

pendant un apprentissage sur un ensemble de taille m, avec ϵ_m choisi minimal tel que $2m \geq \log N(\epsilon_m)/\epsilon_m^2$. Alors, si $P(Y=1|X)$ appartient à $\mathcal{H}$,

$$P(L_m - L^* > \delta) \leq 2\exp(2m\epsilon_m^2 - m(\delta - 2\epsilon_m)^2/2)$$

$$\mathbb{E}_{P^m}\, L_m - L^* \leq (2+\sqrt{8})\epsilon_m + \sqrt{\pi/m}$$

Donc $L_m \to L^*$ avec probabilité 1.

Ceci sera noté, par la suite, MER_C^I (minimisation du risque empirique incrémental avec couvertures). Ce résultat provient de [DEV 96, chap. 28], où on peut trouver de nombreuses références sur des travaux liés, inspiré de [KOL 61]. Utiliser des nombres de couverture pour L^∞ au lieu de L^1 amène à une borne uniforme en la distribution sous-jacente.

Un point important est à noter. Nous parlons ici d'optimiser le risque empirique *parmi* un ϵ-squelette, chose délicate. En régression, des résultats similaires sont possibles en optimisant directement le risque empirique parmi la famille de fonctions complète. Le lecteur est renvoyé à [VID 97] pour le cas de la régression.

Examiner plus précisément la vitesse de convergence qui découle de ces lignes, comme dans [DEV 96], montre que les vitesses typiques sont de $\sqrt{\log(m)/m}$ pour des familles paramétriques (avec des nombres de couverture polynomiaux) et à

7. Notez que nous pouvons considérer les nombres de couverture et les réseaux dépendant seulement de la distribution marginale en X, car l'on considère les nombres de couverture des fonctions de prévision et non des fonctions de coût.

$1/m^{k/2}$ avec $k < 1$ pour des familles de fonctions régulières, selon le niveau de régularité (voir théorème 30).

Ceci permet de travailler sur tout ensemble totalement borné (i.e. tout ensemble dont les nombres de couverture sont finis pour tout $\epsilon > 0$) de fonctions, pourvu que l'on sache limiter la précision à laquelle on travaille sur $\mathcal{H}$ (i.e. on optimise le risque empirique sur un ϵ-squelette). Le fait que le nombre de couverture augmente plus vite que polynomialement (e.g. cadre VC) ou même plus vite que la "lente" exponentielle permettant la finitude de l'intégrale d'entropie uniforme (e.g. cadre Donsker), conduit à des bornes non-asymptotiques. En particulier, l'apprentissage pour des distributions marginales en X uniformes permet donc de travailler sur des espaces de Hölder même en grande dimension et sans forcer le coefficient α ; le cadre des modèles déformables est en particulier couvert. La force de ce résultat montre l'importance d'une connaissance a priori sur la distribution marginale de (X, Y) en X.

D'autres résultats, comme le suivant, extrait de [VID 97, page 188], peuvent être démontrés de même, utilisant ce savoir a priori, pour établir de combien d'exemples on a besoin pour garantir une précision donnée.

THÉORÈME 49 (Complexité en échantillon) *On se donne $\epsilon > 0$. Considérons $\mathcal{H}$ une famille de fonctions à valeurs dans $[0, 1]$, avec des nombres de couverture pour L^1 finis $N(\epsilon)$ (pour la distribution marginale sous-jacente en X, ou pour L^∞ si la distribution marginale est inconnue).*

Considérons $\mathcal{H}_\epsilon$ un ϵ-réseau de taille $N(\epsilon)$. Considérons un algorithme minimisant le risque empirique [8] *parmi $\mathcal{H}_\epsilon$, quand on apprend sur un ensemble d'apprentissage de taille m, avec m choisi minimal tel que $m \geq \frac{2}{\epsilon^2} \ln(\frac{N(\epsilon)}{\delta})$.*

Alors, avec probabilité au moins $1 - \delta$, $L_m - \inf L_{\mathcal{H}} \leq 2\epsilon$.

REMARQUE 5 *MER^I est raisonnablement implanté dans des cas pratiques (quoiqu'il conduise à des problèmes NP-complets dans beaucoup de cas), contrairement à MER^I_C (qui nécessite la construction plus ou moins explicite d'une ϵ-couverture). MER^I_C présente l'avantage de modéliser le paramétrage de précision finie par un ϵ-réseau, difficile (mais possible) à mettre en œuvre ; MER^I_C permet de modéliser des convergences plus lentes que $1/\sqrt{m}$ dans des cadres non-Donsker et non-VC, là où les résultats de type Glivenko-Cantelli ne fournissent que des convergences sans vitesse ; en particulier, MER^I_C est adapté au cadre Hölderien même si la dimension est grande devant le degré de la régularité, ainsi qu'aux espaces d'ensembles convexes.*

8. Pour la fonction de coût usuelle L^2.

10.4.3. *Minimisation structurelle du risque*

Beaucoup de résultats existent sous le nom de "minimisation structurelle du risque". Le théorème ci-dessous est extrait de [DEV 96, page 294] et a été initialement prouvé par Lugosi et Zeger [LUG 95, LUG 96].

THÉORÈME 50 (Minimisation structurelle du risque) *Soient $\mathcal{H}_1, \ldots, \mathcal{H}_k \ldots$ ayant des dimensions finies $V_1, \ldots, V_k, \ldots$ Soit $\mathcal{H} = \cup_n \mathcal{H}_n$. Supposons que toute distribution conduise à $L_{\mathcal{H}} = L^*$ (de telles classes de fonctions existent !). Considérons l'algorithme consistant à choisir $f \in \mathcal{H}$ minimisant l'erreur empirique plus $\sqrt{\frac{32}{m} V(f) \log(e \times m)}$, où $V(f)$ est V_k avec k minimal tel que $f \in \mathcal{H}_k$. Alors :*

– si une règle de Bayes (une fonction conduisant à une erreur L^*) appartient à $\mathcal{H}_k$, alors pour tous m et ϵ tels que

$$V_k \log(e \times m) \leq m\epsilon^2/512$$

l'erreur en généralisation est plus petite que ϵ avec risque plus petit que

$$\Delta \exp(-m\epsilon^2/128) + 8m^{V_k} \times \exp(-m\epsilon^2/512)$$

avec $\Delta = \sum_{j=1}^{\infty} \exp(-V_j)$ supposé fini.

– L'erreur en généralisation, pour toute distribution d'exemples, converge vers L^* avec probabilité 1 (consistance universelle).

Interprétation : Le second résultat était déjà vrai pour la minimisation du risque empirique "améliorée" comme dans le théorème 45. Le premier résultat garantit la convergence rapide, sur une famille de VC-dimension infinie. Notez toutefois que la rapidité est seulement garantie asymptotiquement, la constante ne dépendant que d'un classifieur de Bayes. La vitesse de convergence est toutefois *uniforme en la distribution pour un classifieur de Bayes donné*, et on assure *convergence asymptotique en* $O(1/\sqrt{m})$ (les facteurs logarithmiques peuvent être supprimés grâce à un résultat d'Alexander, voir [ALE 84]).

La première partie du résultat sonne comme un résultat de résistance au bruit au sens où quelle que soit la distribution, tant que la fonction optimale n'est pas modifiée, la vitesse de convergence est la même. La seconde partie est une consistance universelle comme déjà fournie par MER^I.

La minimisation structurelle du risque conduit, dans de nombreux cas intéressants de familles de fonctions, à la formulation de problèmes NP-complets. Les machines à vecteurs support sont souvent présentées comme des minimiseurs du risque structurel, point discutable puisque pour des raisons algorithmiques on utilise dans la procédure

d'optimisation une fonction objectif différente du risque garanti standard. Dans le cas de la rétro-propagation avec régularisation, la fonction objectif utilisée dans la procédure d'optimisation est bien celle correspondant à la mise en œuvre du principe inductif MSR (si les coefficients sont choisis pour cela), mais à l'inverse les algorithmes utilisables en pratique ne garantissent pas la convergence vers un optimum global (sauf à utiliser des méthodes dont la durée de convergence est totalement rédhibitoire). Notez qu'une approximation est faite dans les SVM, de manière à transformer le problème NP-complet de la minimisation structurelle du risque en un problème de programmation quadratique sur un domaine convexe, donc un problème de programmation convexe. Les équivalents linéaires des SVM sont parfois présentés comme plus efficaces que les SVM elles-mêmes. Il n'y a pas de raison théorique à cela dans le cadre général, mais il est joli de pouvoir réduire le problème de l'apprentissage à un problème d'optimisation linéaire, pour lequel beaucoup d'algorithmes classiques existent (voir par exemple [FLE 87]).

10.5. Extensions

Nous présentons enfin dans cette section quelques compléments liés aux développements récents de la théorie de l'apprentissage. Nous présentons en 10.5.1 des applications récentes des variables de Rademacher. Nous présentons en 10.5.2 les conditions de Massart et Tsybakov. D'autres directions de recherche concernent la définition de bornes non-asymptotiques dans le cas de l'apprentissage bayésien (bornes dont la justesse ne serait pas dépendante de la véracité de l'a priori de Bayes) ; ces bornes sont d'autant plus importantes que l'apprentissage bayésien de réseaux neuronaux a les avantages i) d'être optimal si l'a priori est vrai ii) d'être efficace en pratique.

Le lecteur intéressé est vivement encouragé à étudier [BOU 05] pour un "survey" de ces différents sujets.

10.5.1. *Les variables de Rademacher et leurs applications*

Il est connu de longue date que les inégalités de type VC-dimension sont trop conservatrices, au sens où elles sont certes bien vraies, mais d'un peu loin. Il est d'ailleurs remarquable que par nature, elles sont forcément conservatrices ; il s'agit d'inégalités dans le pire des cas sur la distribution. Elles peuvent même ne pas être applicables (i.e. ne pas conduire à des inégalités non dégénérées), comme dans le cas des ensembles de Glivenko-Cantelli de VC-dimension infinie, alors même que le supremum des déviations converge bel et bien. Ce paradoxe apparent mérite explication.

Tout d'abord, pour une famille $\mathcal{F}$ universellement Glivenko-Cantelli presque sûrement et de VC-dimension infinie (et on sait que de telles familles de fonction existent),

presque sûrement,

$$\forall P, Q_{1-\delta} \sup_{f \in \mathcal{F}} |\hat{\mathbb{E}} L_f - \mathbb{E} L_f| \to 0$$

(où $Q_{1-\delta}$ désigne l'opérateur "quantile à $1-\delta$") mais on ne dit pas (et on ne peut établir sans contradiction !) que

presque sûrement,

$$\sup_P Q_{1-\delta} \sup_{f \in \mathcal{F}} |\hat{\mathbb{E}} L_f - \mathbb{E} L_f| \to 0$$

si bien qu'une hypothèse sur la distribution est nécessaire pour obtenir une vitesse de convergence sur une telle famille de fonctions $\mathcal{F}$.

Malheureusement, il est bien peu commode de disposer d'hypothèses sur la distribution. Pour prendre en compte la distribution, *sans hypothèse sur celle-ci*, une façon est de prendre en compte *l'échantillon* (point d'ailleurs central dans le bootstrap).

Cela peut être fait comme suit (voir par exemle [VAA 96]) :

avec probabilité $1-\delta$ et sous des conditions très générales, $\sup_{f \in \mathcal{F}} |\mathbb{E} L_f - \hat{\mathbb{E}} L_f|$ est majoré par l'espérance sur les σ_i de

$$\frac{2}{n} \sup_{f \in \mathcal{F}} \sum_i \sigma_i f(X_i) + \sqrt{2 \log(2/\delta)/n}$$

où, dans le terme de gauche, les σ_i sont des variables aléatoires i.i.d. uniformes sur $\{-1, 1\}$.

10.5.2. *Les conditions de Massart et Tsybakov*

Considérons $\eta(x) = P(Y = 1|X)$ lors d'une tâche de discrimination bi-classe. Alors il est clair que le cas problématique est $\eta(x) = \frac{1}{2}$. L'erreur bayésienne est même directement liée à η et sa proximité à $\frac{1}{2}$:

$$L^* = \frac{1}{2} - \mathbb{E}|\eta(x) - \frac{1}{2}|.$$

Un fait moins clair mais bien réel est le fait que la vitesse de convergence vers ce L^* dépend de cette même quantité.

Précisément, la condition de Massart stipule que s'il existe $s > 0$ tel que $|\eta(x) - \frac{1}{2}| > s$ alors, avec $\mathcal{F}$ famille de fonctions à valeurs binaires de VC-dimension finie V, avec probabilité $1-\delta$ au moins,

$$\mathbb{E} L_{\arg\min_{f \in \mathcal{F}} \hat{\mathbb{E}} L_f} \leq C \frac{\log(N/\delta)}{n}.$$

Une extension, connue sous le nom de condition de Tsybakov [TSY 04], stipule que si

$$P(|2\eta(x) - 1| \leq s) = O(s^{\alpha/(1-\alpha)})$$

pour un certain $\alpha \in [0, 1]$, alors

$$\mathbb{E}L_{\arg\min_{f\in\mathcal{F}} \hat{\mathbb{E}}L_f} \leq C(\frac{\log(N/\delta)}{n})^{(1/(2-\alpha))}.$$

10.6. Conclusions et perspectives

Dans ce chapitre, nous avons étudié les performances en généralisation des réseaux de neurones, en nous plaçant dans le cadre fourni par la théorie statistique de l'apprentissage. Si nous avons plus particulièrement concentré notre attention sur l'architecture la plus utilisée, le perceptron multi-couche, et la tâche d'apprentissage pour laquelle la théorie est la plus avancée, la discrimination, les résultats que nous avons relatés demeurent d'une portée très générale. Nous avons finalement :

1) présenté les problématiques de discrimination, régression et estimation de densité ;

2) résumé les inégalités utiles montrant que les espérances ne sont pas trop loin des moyennes dans un certain nombre de cas ;

3) introduit les outils les plus classiques pour caractériser la capacité d'une famille de fonctions ;

4) présenté les résultats classiques combinant les points 2 et 3 pour établir que dans des familles de fonctions complexes, on pouvait avoir comme dans des cas paramétriques des convergences 10.3.4.1, 10.4.1, des vitesses de convergence 10.3.4.2, des bornes 10.4.1, des approximations asynmptotiques 10.3.4.3 sur les différences entre moyennes et espérance conduisant à des consistances 10.3.4.1, 10.3.4.2 ;

5) présenté des algorithmes plus complexes que la minimisation empirique du risque (minimisation incrémentale du risque empirique 10.4.2.1, minimisation structurelle du risque 10.4.3), justifiant le fait de dimensionner la régularisation en fonction du nombre d'exemples disponibles pour obtenir de meilleures vitesses de convergence ou des consistances *universelles* ;

6) présenté les directions de recherche courante (10.5).

Pour conclure, nous souhaitons mettre l'accent sur un dernier développement récent, particulièrement prometteur pour obtenir des bornes de meilleure qualité : l'utilisation de mesures de capacité locales et empiriques (calculées sur l'échantillon d'apprentissage). Cette approche a en particulier été développée par Bousquet et ses coauteur [BOU 02, BAR 05]. Ils la considèrent plus spécifiquement en conjonction avec l'utilisation de moyennes de Rademacher.

Remerciements

Les auteurs remercient Monsieur le Professeur Paul Deheuvels pour des références et discussions utiles. Les travaux de Yann Guermeur sont financés par l'ACI "Masse de Données". Les travaux d'Olivier Teytaud bénéficient du support du réseau d'excellence européen "PASCAL" et de l'ACI "MIST-R".

10.7. Bibliographie

[ADL 90] ADLER R., « An Introduction to Continuity, Extrema, and Related Topics for General Gaussian Processes », *vii + 160, IMS Lecture Notes-Monograph Series*, 1990.

[ALE 84] ALEXANDER K., « Probability inequalities for empirical processes and a law of the iterated logarithm », *Annals of Probability*, vol. 12, p. 1041–1067, 1984.

[ALO 97] ALON N., BEN-DAVID S., CESA-BIANCHI N., HAUSSLER D., « Scale-sensitive Dimensions, Uniform Convergence, and Learnability. », *J. ACM*, vol. 44, n°4, p. 615–631, 1997.

[AND 88] ANDERSEN N., GINÉ E., OSSIANDER M., ZINN J., « The central limit theorem and the law of iterated logarithm for empirical processes under local conditions », *Probability theory and related fields 77, 271-305*, 1988.

[ANT 97] ANTHONY M., « Probabilistic analysis of learning in artificial neural networks : the PAC model and its variants. », *Neural Computing Surveys*, vol. 1, p. 1–47, 1997.

[ANT 99] ANTHONY M., BARTLETT P., *Artificial Neural Network Learning : Theoretical Foundations*, Cambridge University Press, New York, 1999.

[ARO 50] ARONSZAJN N., « Theory of reproducing kernels », *Trans. Amer. Math. Soc.*, vol. 68, p. 337-404, 1950.

[BAR 93] BARRON A., « Universal approximation bounds for superpositions of a sigmoidal function », *IEEE Transactions on Information Theory*, vol. 39, p. 930–945, 1993.

[BAR 96] BARTLETT P., WILLIAMSON R., « The Vapnik-Chervonenkis dimension and pseudodimension of two-layer neural networks with discrete inputs », *Neural computation*, vol. 8, p. 653-656, 1996.

[BAR 98a] BARBE P., LEDOUX M., *Probabilité*, De la licence à l'agrégation, Belin, 1998.

[BAR 98b] BARTLETT P., « The sample complexity of pattern classification with neural networks : the size of the weights is more important than the size of the network », *IEEE Transactions on Information Theory*, vol. 44, n°2, p. 525–536, 1998.

[BAR 99a] BARRON A., BIRGÉ L., MASSART P., « Risk bounds for model selection via penalization », *Probab. Theory Relat. Fields*, vol. 113, p. 301-413, 1999.

[BAR 99b] BARTLETT P., SHAWE-TAYLOR J., « Generalization Performance of Support Vector Machines and Other Pattern Classifiers », SCHÖLKOPF B., BURGES C., SMOLA A., Eds., *Advances in Kernel Methods, Support Vector Learning*, p. 43–54, The MIT Press, Cambridge, 1999.

[BAR 05] BARTLETT P., BOUSQUET O., MENDELSON S., « Local Rademacher Complexities », *Annals of Statistics*, vol. 33, n°4, p. 1497-1537, 2005.

[BAU 89] BAUM E., HAUSSLER D., « What size net gives valid generalization ? », *Neural Computation*, vol. 1, p. 151-160, 1989.

[BEN 62] BENNETT G., « Probability inequalities for sum of independent random variables », *Journal of the American Statistical Association*, vol. 57, p. 33–45, 1962.

[BEN 94] BENEDEK G., ITAI A., « Nonuniform learnability », *Journal of Computer and Systems Sciences*, vol. 48, p. 311–323, 1994.

[BEN 95] BEN-DAVID S., CESA-BIANCHI N., HAUSSLER D., LONG P., « Characterizations of Learnability for Classes of $\{0, \ldots, n\}$-Valued Functions. », *Journal of Computer and System Sciences*, vol. 50, p. 74–86, 1995.

[BER 46] BERNSTEIN S., *The Theory of Probabilities*, Gastehizdat Publishing House, Moscow, 1946.

[BIR 67] BIRMAN M.-S., SOLOMJAK M.-Z., « Piecewise-polynomial approximation of functions of the classes W_p », *Mathematics if tge USSR Sbornik 73, 295-317*, 1967.

[BIS 96] BISHOP C. M., *Neural networks for pattern recognition*, Oxford University Press, Oxford, UK, 1996.

[BLU 55] BLUM J., « On the convergence of empiric distribution functions », *Annals of Mathematical Statistics*, vol. 26, p. 527-529, 1955.

[BOS 92] BOSER B., GUYON I., VAPNIK V., « A training algorithm for optimal margin classifiers », *COLT'92*, p. 144-152, 1992.

[BOT 97] BOTTOU L., « La mise en œuvre des idées de Vladimir N. Vapnik », THIRIA S., LECHEVALLIER Y., GASCUEL O., CANU S., Eds., *Statistique et méthodes neuronales*, p. 262–274, DUNOD, 1997.

[BOU 02] BOUSQUET O., Concentration Inequalities and Empirical Processes Theory Applied to the Analysis of Learning Algorithms, PhD thesis, Ecole Polytechnique, 2002.

[BOU 05] BOUSQUET O., BOUCHERON S., LUGOSI G., « Theory of Classification : A Survey of Some Recent Advances », *ESAIM Probability and Statistics*, vol. 9, p. 323-375, 2005.

[BRE 96] BREIMAN L., « Bagging Predictors », *Machine Learning*, vol. 24, p. 123–140, 1996.

[CAR 90] CARL B., STEPHANI I., *Entropy, compactness, and the approximation of operators*, Cambridge University Press, Cambridge, UK, 1990.

[CER 04] CERVELLERA C., MUSELLI M., « Deterministic design for neural network learning : an approach based on discrepancy », *IEEE transactions on Neural Networks*, vol. 15, n°3, p. 533–544, 2004.

[CHU 49] CHUNG K.-L., « An estimate concerning the Kolmogoroff limit distribution », *Transactions of the American Mathematical Society, 67 :36–50*, 1949.

[COR 95] CORTES C., VAPNIK V., « Support-Vector Networks », *Machine Learning*, vol. 20, p. 273–297, 1995.

[CYB 89] CYBENKO G., « Approximation by Superpositions of a Sigmoidal Function », *Mathematics of Control, Signals, and Systems*, vol. 2, p. 303–314, 1989.

[DEH 71] DEHARDT J., « Generalizations of the Glivenko-Cantelli theorem », *Annals of Mathematical Statistics 42, 2050-2055*, 1971.

[DEV 87] DEVROYE L., *A course in density estimation.*, Birkhauser Boston Inc., Cambridge, MA., 1987.

[DEV 96] DEVROYE L., GYÖRFI L., LUGOSI G., *A Probabilistic Theory of Pattern Recognition*, Springer-Verlag, New-York, 1996.

[DUD 66] DUDLEY R., « Weak convergence of measures on non separable metric spaces and empirical measures on Euclidean spaces », *Illinois J. Math.*, vol. 10, p. 109–126, 1966.

[DUD 73] DUDA R., HART P., *Pattern Classification and Scene Analysis*, Wiley, 1973.

[DUD 78] DUDLEY R., « Central limit theorems for empirical measures », *Annals of Probability*, vol. 6, p. 899–929, 1978.

[DUD 84] DUDLEY R.-M., « A course on empirical processes (Ecole d'été de Probabilité de Saint-Flour XII-1982) », *Lecture notes in Mathematics 1097, 2-141 (ed P.L. Hennequin)*, Springer-Verlag, New-York, 1984.

[DUD 87] DUDLEY R., « Universal Donsker classes and metric entropy », *Ann. Probab.*, vol. 15, n°4, p. 1306–1326, 1987.

[DUD 91] DUDLEY R., GINÉ E., ZINN J., « Uniform and universal Glivenko-Cantelli classes », *Journal of Theoretical Probability*, vol. 4, p. 485–510, 1991.

[EFR 79] EFRON B., « Bootstrap methods : another look at the jackknife », *Annals of Statistics*, vol. 7, p. 1–26, 1979.

[FLE 87] FLETCHER R., *Practical Methods of Optimization*, Wiley, 1987.

[GAS 97] GASCUEL O., « La dimension de Vapnik-Chervonenkis, application aux réseaux de neurones », THIRIA S., LECHEVALLIER Y., GASCUEL O., CANU S., Eds., *Statistique et méthodes neuronales*, p. 244–261, DUNOD, 1997.

[GEE 00] VAN DER GEER S., *Empirical Processes in M-estimation*, Cambridge University Press, 2000.

[GER 02] GERSTNER W., KISTLER W., *Spiking Neuron Models, Single Neurons, Populations, Plasticity*, Cambridge University Press, 2002.

[GIN 84] GINÉ E., ZINN J., « Some limit theorems for empirical processes », *Annals of Probability*, vol. 12, n°4, p. 929–989, 1984.

[GIN 90] GINÉ E., ZINN J., « Bootstrapping general empirical measures », *Annals of Probability*, vol. 18, p. 851–869, 1990.

[GIR 01] GIRARDIN V., LIMNIOS N., *Probabilités, cours et exercices en vue des applications*, Vuibert, 2001.

[GOL 95] GOLBERG P., JERRUM M., « Bounding the Vapnik-Chervonenkis dimension of concept classes parametrized by real numbers », *Machine Learning*, vol. 18, p. 131–148, 1995.

[GRü 67] GRÜNBAUM B., *Convex Polytopes*, vol. XVI de *Pure and Applied Mathematics*, Interscience Publishers, 1967.

[GUE 04] GUERMEUR Y., Large margin multi-category discriminant models and scale-sensitive Ψ-dimensions, Rapport n°RR-5314, INRIA, 2004.

[GUE 05a] GUERMEUR Y., MAUMY M., SUR F., « Model selection for multi-class SVMs », *ASMDA'05*, p. 507–516, 2005.

[GUE 05b] GUERMEUR Y., MAUMY M., SUR F., Notes sur le "théorème de Maurey-Carl", Rapport n°RR-5???, INRIA, 2005.

[GUO 02] GUO Y., BARTLETT P., SHAWE-TAYLOR J., WILLIAMSON R., « Covering Numbers for Support Vector Machines », *IEEE Trans. on Information Theory*, vol. 48, n°1, p. 239–250, 2002.

[GUR 95] GURVITS L., KOIRAN P., « Approximation and learning of convex superpositions », *Proceedings of EUROCOLT'95*, 1995.

[GUR 97] GURVITS L., KOIRAN P., « Approximation and Learning of Convex Superpositions », *Journal of Computer and System Sciences*, vol. 55, n°1, p. 161–170, 1997.

[GUR 01] GURVITS L., « A note on a scale-sensitive dimension of linear bounded functionals in Banach spaces », *Theoretical Computer Science*, vol. 261, n°1, p. 81–90, 2001.

[HAS 02] HASTIE T., TIBSHIRANI R., FRIEDMAN J., *The Elements of Statistical Learning - Data Mining, Inference, and Prediction*, Springer, 2002.

[HAU 92] HAUSSLER D., « Decision theoretic generalizations of the PAC model for neural net and other learning applications. », *Information and Computation*, vol. 100, p. 78–150, 1992.

[HAU 95a] HAUSSLER D., « Sphere packing numbers for subsets of the boolean n-cube with bounded Vapnik-Chervonenkis dimension. », *Journal of Combinatorial Theory A*, vol. 69, p. 217–232, 1995.

[HAU 95b] HAUSSLER D., LONG P., « A Generalization of Sauer's Lemma. », *Journal of Combinatorial Theory, Series A*, vol. 71, p. 219–240, 1995.

[HOE 63] HOEFFDING W., « Probability inequalities for sums of bounded random variables », *Journal of the American Statistical Association*, vol. 58, p. 13–30, 1963.

[HOR 89] HORNIK K., STINCHCOMBE M., WHITE H., « Multilayer feedforward networks are universal approximators », *Neural Networks*, vol. 2, p. 359–366, 1989.

[KAR 95] KARPINSKI M., MACINTYRE A., « Polynomial bounds for VC dimension of sigmoidal neural networks », *Proceedings of the ACM Symposium on Theory of Computing*, p. 200–208, 1995.

[KAR 97] KARPINSKI M., MACINTYRE A., « Polynomial bounds for VC dimension of sigmoidal and general Pfaffian neural networks », *Journal of Computer and System Sciences*, vol. 54, p. 169-176, 1997.

[KEA 90] KEARNS M., SCHAPIRE R., « Efficient distribution-free learning of probabilistic concepts », *Proceedings of the 31st Annual Symposium on Foundations of Computer Science*, vol. 1, IEEE Computer Society Press, p. 382–391, 1990.

[KEA 94] KEARNS M., SCHAPIRE R., « Efficient Distribution-free Learning of Probabilistic Concepts », *Journal of Computer and System Sciences*, vol. 48, n°3, p. 464–497, 1994.

[KHI 28] KHINTCHINE A. Y., « Sur la loi forte des grands nombres », *Comptes Rendus de l'Académie des Sciences*, vol. 186, 1928.

[KIE 61] KIEFER J., « On large deviations of the empiric d.f. of vector chance variables and a law of the iterated logarithm », *Pacific journal of mathematics, 11, pp. 649-660*, 1961.

[KOH 89] KOHONEN T., *Self-Organization and Associative Memory*, Springer-Verlag, Berlin, 1989.

[KOI 97] KOIRAN P., SONTAG E., « Neural Networks with Quadratic VC Dimension », *Journal of Computer and System Sciences*, vol. 54, n°1, p. 190–198, 1997.

[KOL 61] KOLMOGOROV A., TIHOMIROV V., « ϵ-entropy and ϵ-capacity of sets in functional spaces », *Amer. Math. Soc. Translations (2)*, vol. 17, p. 277–364, 1961.

[KOL 81] KOLCINSKI V., « On the central limit theorem for empirical measures », *Theory of Probability and Mathematical Statistics 24, 71-82*, 1981.

[LED 91] LEDOUX M., TALAGRAND M., *Probability in Banach Spaces*, Springer-Verlag, Berlin, 1991.

[LOR 66] LORENTZ G.-G., *Approximation of Functions*, Holt, Rhinehart, Winston, New York, 1966.

[LUG 95] LUGOSI G., ZEGER K., « Nonparametric estimation via empirical risk minimization », *IEEE Transactions on Information Theory*, vol. 41, p. 677–687, 1995.

[LUG 96] LUGOSI G., ZEGER K., « Concept learning using complexity regularization », *IEEE Transactions on Information Theor*, vol. 42, p. 48–54, 1996.

[LUG 04] LUGOSI G., « Concentration-of-measure inequalities », Lecture notes, Summer School on Machine Learning at the Australian National University, Canberra, 2004.

[MAC 93] MACINTYRE A., SONTAG E., « Finiteness results for sigmoidal "neural" networks », *Proceedings of 25th Annual ACM Symposium on the Theory of Computing*, p. 325–334, 1993.

[MCD 89] MCDIARMID C., « On the method of bounded differences », *Surveys in Combinatorics*, vol. 141, p. 148-188, 1989, Cambridge University Press.

[NAI 04] NAIM P., WUILLEMIN P.-H., LERAY P., POURRET O., BECKER A., *Réseaux bayésiens*, Eyrolles, Paris, 2004.

[NAT 89] NATARAJAN B., « On learning sets and functions », *Machine Learning*, vol. 4, p. 67–97, 1989.

[NIE 92] NIEDERREITER H., *Random Number Generation and Quasi-Monte-Carlo Methodes*, Society of Industrial and Applied Mathematics, 1992.

[OSS 87] OSSIANDER M., « A central limit theorem under metric entropy with L_2 bracketing », *Annals of probability 15, 897-919*, 1987.

[POI 35] POISSON S., « Recherche sur la probabilité des jugements, principalement en matière criminelle », *Comptes-Rendus hebdomadaires des Séances de l'Académie des Sciences*, vol. 1, p. 473-494, 1835.

[POL 82] POLLARD D., « A central limit theorem for empirical processes », *Journal of the Australian Mathematical Society A33, 235-248*, 1982.

[POL 84] POLLARD D., *Convergence of stochastic processes*, Springer-Verlag, N.Y., 1984.

[POL 90] POLLARD D., « Empirical Processes : Theory and Applications », *NFS-CBMS Regional Conference Series in Probability and Statistics*, vol. 2, Institute of Math. Stat. and Am. Stat. Assoc., 1990.

[RIC 91] RICHARD M., LIPPMANN R., « Neural Network Classifiers Estimate Bayesian a posteriori Probabilities », *Neural Computation*, vol. 3, p. 461–483, 1991.

[SAK 93] SAKURAI A., « Tighter bounds of the VC-dimension of three-layer networks. », *WCNN'93*, p. 540-543, 1993.

[SAP 90] SAPORTA G., *Probabilités Analyse des Données et Statistique*, Technip, France édition, 1990.

[SAU 72] SAUER N., « On the density of families of sets », *Journal of Combinatorial Theory (A)*, vol. 13, p. 145–147, 1972.

[SCH 98] SCHAPIRE R., FREUND Y., BARTLETT P., LEE W., « Boosting the Margin : A New Explanation for the Effectiveness of Voting Methods », *The Annals of Statistics*, vol. 26(5), p. 1651–1686, 1998.

[SCH 02] SCHÖLKOPF B., SMOLA A., *Learning with Kernels, Support Vector Machines, Regularization, Optimization and Beyond*, The MIT Press, 2002.

[SHA 91] SHAWE-TAYLOR J., ANTHONY M., « Sample sizes for multiple-output threshold networks. », *Network : Computation in Neural Systems*, vol. 2, p. 107–117, 1991.

[SHE 72] SHELAH S., « A combinatorial problem : Stability and order for models and theories in infinitary languages », *Pacific Journal of Mathematics*, vol. 41, p. 247–261, 1972.

[SON 92] SONTAG E., « Feedforward nets for interpolation and classification », *J. Comp. Syst. Sci*, vol. 45, p. 20–48, 1992.

[SON 98] SONTAG E., « VC Dimension of Neural Networks », BISHOP C., Ed., *Neural Networks and Machine Learning*, p. 69–95, Springer-Verlag, Berlin, 1998.

[STE 01] STEINWART I., « On the influence of the kernel on the consistency of support vector machines », *Journal of Machine Learning Research*, vol. 2, p. 67–93, 2001.

[TOU 99] TOULOUSE P., *Thèmes de Probabilités et Statistique*, Dunod, 1999.

[TSY 04] TSYBAKOV A., « Optimal aggregation of classifiers in statistical learning », *Annals of Statistics*, vol. 32, n°1, 2004.

[VAA 96] VAN DER VAART A., WELLNER J., *Weak Convergence and Empirical Processes, With Applications to Statistics*, Springer Series in Statistics, Springer-Verlag New York, Inc., 1996.

[VAA 98] VAN DER VAART A., *Asymptotic Statistics*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, 1998.

[VAP 71] VAPNIK V., CHERVONENKIS A., « On the uniform convergence of relative frequencies of events to their probabilities. », *Theory of Probability and its Applications*, vol. XVI,

n°2, p. 264–280, 1971.

[VAP 81] VAPNIK V., CHERVONENKIS A., « Necessary and sufficient conditions for the uniform convergence of the means to their expectations. », *Theory of Probability and its Applications*, vol. 26, p. 532–553, 1981.

[VAP 82] VAPNIK V., *Estimation of Dependences Based on Empirical Data.*, Springer-Verlag, N.Y, 1982.

[VAP 89] VAPNIK V., « Inductive principles of the search for empirical dependencies », *Proceedings of the 2nd Annual Workshop on Computational Learning Theory*, p. 3–21, 1989.

[VAP 98] VAPNIK V., *Statistical learning theory.*, John Wiley & Sons, Inc., N.Y., 1998.

[VID 97] VIDYASAGAR M., *A Theory of Learning and Generalization, with Applications to Neural Networks and Control Systems*, Springer-Verlag, New-York, 1997.

[WAH 90] WAHBA G., « Spline Models for Observational Data », *SIAM*, vol. 59 de *CBMS-NSF Regional Conference Series in Applied Mathematics*, 1990.

[WAH 99] WAHBA G., « Support Vector Machines, Reproducing Kernel Hilbert Spaces, and Randomized GACV », SCHÖLKOPF B., BURGES C., SMOLA A., Eds., *Advances in Kernel Methods, Support Vector Learning*, p. 69–88, The MIT Press, 1999.

[WAS 97] WASILKOWSKI G., WOZNIAKOWSKI H., « The exponent of discrepancy is at most 1.4778 », *Math. Comp., 66, pp. 1125-1132*, 1997.

[WIL 00] WILLIAMSON R., SMOLA A., SCHÖLKOPF B., « Entropy Numbers of Linear Function Classes », *COLT'00*, p. 309–319, 2000.

[WIL 01] WILLIAMSON R., SMOLA A., SCHÖLKOPF B., « Generalization Performance of Regularization Networks and Support Vector Machines *via* Entropy Numbers of Compact Operators », *IEEE Trans. on Information Theory*, vol. 47, n°6, p. 2516–2532, 2001.