Deuxième partie

L'apprentissage dans les cas difficiles

Cette partie se concentre sur les cas où l'apprentissage bénéficie plus difficilement de garanties statistiques. La théorie la plus classique, la VC-théorie, fournit des bornes quand:

- *On fait confiance au "professeur"*, qui fournit les sorties désirées. On verra qu'on peut réduire cette hypothèse.
- *Les points sont indépendants.* Ceci est généralement une hypothèse passée sous silence, et l'on peut ainsi voir des articles testant la validité de la VC-théorie, et concluant à sa validité expérimentale, sur des bases qui de toute évidence ne sont pas absolument pas distribuées avec indépendance (points successifs d'une série temporelle par exemple).
- *Les points sont identiquement distribués.* Pour de nombreux cas concrets cette hypothèse n'est pas valable. Les inégalités de Hoeffding et Bernstein s'adaptent très facilement à ce cas.
- *La VC-dim est finie.* Ceci peut se produire lorsque l'on a des a priori physiques, où lorsque l'on sait expérimentalement que telle famille de fonctions suffit: mais on sait que dans le cas général, une famille de VC-dim finie ne peut pas être un bon approximateur.

On fournit donc dans les parties qui suivent des études sur:

- *Les pire-cas de convergence en VC-dim finie*, qui restent donc dans le cadre de la VC-dim finie sur points iid, mais cherche les distributions "catastrophes" d'un point de vue de convergence des moyennes empiriques vers les espérances.
- *Les cas de points ne sont pas indépendants*: les séries temporelles. Ce chapître apparait crucial en vertue des nombreuses applications concrètes en contrôle.
- *Les cas où les points ne sont pas identiquement distribués*: on verra notamment des applications en imagerie (importantes pour l'imagerie médicale notamment).
- *Le cas où le professeur se trompe*, à cause du bruit.
- *Les cas où la VC-dim est infinie.*

Chapitre 2

Bornes inférieures pour les estimateurs empiriques, leave-one-out et cross-validation de l'erreur en généralisation

Résumé

Nous étudions dans le cadre PAC-agnostique la différence entre l'erreur empirique et l'erreur leave-one-out. En outre, nous améliorons les bornes inférieures pour la convergence uniforme des moyennes empiriques vers leurs espérances en VC-dimension finie. Enfin, nous obtenons des résultats similaires pour la cross-validation lorsque l'on utilise un trop grand nombre de sous-ensemble, ce qui confirme un résultat connu des praticiens selon lequel il faut utiliser un nombre réduit de sous-ensembles, quand bien même le nombre d'exemples est très grand.

Table des matières

2.1 Introduction

Ce travail a été réalisé en collaboration avec G. Gavin ([Gavin et al, 2001]).

La reconnaissance de forme consiste à modéliser un phénomène à partir d'un échantillon aussi appelé échantillon d'apprentissage. Le propos est d'établir une relation fonctionnelle h entre un espace d'entrée X et un espace de sortie Y tel que h minimise un critère donné. Un algorithme d'apprentissage est défini comme étant une fonction récursive de l'ensemble des échantillons d'apprentissage dans un certain espace h de telles fonctions. Une solution naturelle consiste en minimiser le critère pour la distribution empirique en espérant que les performances peuvent être généralisées aux observations non effectuées. Pour quantifier cet espoir, des résultats statistiques sont nécessaires.

Les résultats de la théorie statistique de l'apprentissage fournissent des bornes supérieures sur le nombre d'exemples d'apprentissage nécessaires pour avoir des performances en généralisation satisfaisantes, en terme de dimension de Vapnik-Chervonenkis (VC-dimension). La VC-dimension est une valeur combinatoire calculée sur H caractérisant sa "complexité". Le théorème de Vapnik et Chervonenkis traite de problèmes de classification deux-classes, ie $Y = \{0,1\}$. Il montre que la finitude de la VC-dimension est une condition suffisante pour avoir la convergence de l'erreur empirique vers l'erreur en généralisation (condition en outre nécessaire au pire cas sur la distribution). La seule hypothèse est que le phénomène peut être modélisé par une loi jointe sur $X \times Y$ dont les éléments sont tirés au sort de manière indépendante ou identiquement distribuée, sans hypothèse sur la distribution. Ce cadre autorisant le bruit et les recouvrements de classes est très général.

Toutefois, ce théorème amène à considérer des tailles d'échantillon beaucoup trop grandes pour les praticiens. De nombreuses améliorations ont été faites ces dernières années, insuffisantes encore pour les applications pratiques. Les idées les plus intuitives pour améliorer ces résultats consistent en passer outre la convergence uniforme et considérer des algorithmes particuliers. Le plus naturel est l'algorithme minimisant l'erreur empirique. Même si ce paradigme est souvent NP-complet, il peut être vu comme cas limite d'un grand nombre d'algorithmes existants.

Malheureusement, la taille d'échantillon requise pour cet algorithme est équivalente à celle requise pour la convergence uniforme. En fait des bornes inférieures pour cet algorithme sont égales, à des constantes près, à celles obtenus pour des convergences uniformes. Toutefois, ces constantes sont très larges (environ 10 000). Ici, nous réduirons l'écart d'un facteur environ 200.

Usuellement, les estimateurs par rééchantillonnage sont considérés comme plus efficaces pour estimer les performances en généralisation que l'erreur empirique. Plus bas, nous étudierons en détail l'estimateur leave-one-out et considérons l'extension à la cross-validation. On montre que dans le cadre précédent, il n'est pas meilleur que l'erreur empirique. En outre on montre que parfois l'erreur empirique est plus précise.

La partie 2.2 résume le cadre du "machine learning". La partie 2.3 définit la complexité d'échantillon et rappelle quelques résultats usuels. La partie 2.4 rappelle quelques notions sur le leave-one-out. La partie 2.5

donne les théorèmes principaux. La partie 2.4 rappelle quelques notions concernant l'estimateur leave-one-out. La partie 2.5 fournit les résultats principaux. La partie 2.6 est un résumé de corollaires plus explicites. La suite est faite de lemmes techniques et d'extensions à la cross-validation. On montrera notamment dans le cadre PAC le fait connu des praticiens qu'en cross-validation, il faut partitionner en un nombre borné de sous-ensembles.

2.2 Le cadre du Machine Learning

Soit X l'espace d'entrée, $Y = \{0;1\}$ l'espace de sortie, D une distribution sur $X \times Y$ et H l'espace des fonctions prédictives (fonctions de X vers Y). Pour $h \in H$, on note $er_D(h)$ l'erreur en généralisation définie comme la probabilité (pour la distribution D) pour que $h(x) \neq y$. Un exemple d'apprentissage z_n est un ensemble de n observations $(x_i,y_i) \in X \times Y$. On supposera que les observations sont tirées de manière iid (indépendante et identiquement distribuée) selon D. On note $er_{z_n}(h)$ l'erreur empirique définie comme la fréquence des mauvaises classifications sur l'échantillon d'apprentissage.

Un ensemble H a la propriété de convergence uniforme si l'erreur empirique converge uniformément en probabilité vers l'erreur en généralisation. Ceci amène la définition de la complexité d'échantillon.

Définition 2.1 *Soit $\varepsilon > 0$ et $\delta > 0$. La complexité d'échantillon $n_H(\varepsilon,\delta)$ pour un ensemble H pour la convergence uniforme est la plus petite valeur telle que $\forall n \geq n_H(\varepsilon,\delta)$*

$$\sup_D P_{D^n} \left(\sup_{h \in H} \left(|er_D(h) - er_{z_n}(h)| \geq \varepsilon \right) \right) \leq \delta$$

La propriété de la convergence uniforme assure une bonne efficacité en généralisation pour tout algorithme minimisant approximation l'erreur empirique, pourvu que la complexité d'échantillon est suffisamment larger. Dans la prochaine section, des bornes supérieures sur $n_H(\varepsilon,\delta)$ seront présentées. Nous verrons que les bornes sont beaucoup trop larges en pratiques. Pour avoir des améliorations, on peut restreindre la définition précédente à des algorithmes spécifiques. En fait, la convergence uniforme apparait une contrainte trop fort: pour un algorithme donné, on est seulement intéressé par la sortie de l'algorithme et non par l'ensemble des fonctions. Ces considérations amènent à la définition de complexité d'échantillon spécialisée à un algorithme d'apprentissage.

Définition 2.2 *Soit $\varepsilon > 0$ et $\delta > 0$. La complexité d'échantilon $n_A(\varepsilon,\delta)$ pour un algorithme A est la plus petite valeur telle que $\forall n \geq n_A(\varepsilon,\delta)$*

$$\sup_D P_{D^n} \left(|er_D(A(z_n)) - er_{z_n}(A(z_n))| \geq \varepsilon \right) \leq \delta$$

2.3 Bornes sur la complexité d'échantillon

Vapnik et Chervonenkis ont introduit une notion combinatoire calculée pour une famille de fonctions H, appelée la VC-dimension. Elle quantifie la notion de complexité de H. Sa finitude est une condition suffisante pour prouver que H a la propriété de convergence uniforme, sans faire d'hypothèse sur la distribution de probabilité D. En outre, Vapnik et Chervonenkis ont fourni dans [Vapnik, 1982] une complexité d'échantillon explicite dans leur fameux théorème:

Théorème 2.3 (Vapnik et Chervonenkis) *Pour tout $\varepsilon \geq 0$, $\delta \geq 0$*

$$n_H(\varepsilon,\delta) \leq \frac{64}{\varepsilon^2} \left(2d \log \frac{12}{\varepsilon} + \log \frac{4}{\delta} \right)$$

où d est la VC dimension de H

Clairement la finitude de d implique la propriété de convergence uniforme. Le problème de la généralisation est réduit à un problème combinatoire: le calcul de la VC-dimension.

Une des critiques principales de ce théorème est l'impraticabilité de son application. En fait, même pour des petites VC-dimensions, il exige des millions d'exemples pour apprendre aux précision et confiance

usuelles. Il serait intéressant de savoir s'il est possible de grandement réduire cette borne sur la complexité d'échantillon ou si le choix de ne faire aucune hypothèse sur la distribution de probabilité est trop ambitieux.

Durant ces dernières décennies, beaucoup d'améliorations ont été proposées dans la littérature. Alexander [Alexander, 1984] et Talagrand [Talagrand, 1994] ont proposé des améliorations significatives où le facteur logarithmique $\log\left(\frac{1}{\varepsilon}\right)$ est supprimé. Long [Long, 1998] propose la borne suivante.

Théorème 2.4 (Long 1998) *Pour tout* $\varepsilon \geq 0, \delta \geq 0$

$$n_H(\varepsilon,\delta) \leq \frac{576 \ln 41}{\varepsilon^2}\left(4d + \ln\frac{1}{\delta}\right)$$

En pratique, pour des précisions et confiances classiques, cela pourrait être moins intéressant que le résultat de Vapnik et Chervonenkis, parce que les constantes sont plus grandes. Mais d'un point de vue philosophique, il est très intéressant, car des bornes équivalentes sont maintenant montrées. En fait beaucoup de bornes inférieures sur la complexité d'échantillon en $\Omega(\varepsilon^2)$ ont été montrées. Pour autant que nous sachions, le meilleur résultat est dû à Bartlett et Anthony ([Anthony et al, 1999]).

Théorème 2.5 (Bartlett et Anthony) *Soit A un algorithme minimisant l'erreur empirique. Pour tout* $\varepsilon \geq 0, \frac{1}{64} > \delta \geq 0$

$$n_A(\varepsilon,\delta) \geq \frac{1}{\varepsilon^2}\max\left(\frac{d}{320}; \left(1-\varepsilon^2\right)\ln\frac{1}{8\delta\left(1-2\delta\right)}\right)$$

Ces deux théorèmes montrent que la VC-dimension est un critère pertinent pour définir la complexité d'un ensemble de classifieurs pour la convergence uniforme dans le cas agnostique. En d'autres termes, deux ensembles de classifieurs ayant la même VC-dimension ont le même comportement asymptotique dans le pire cas pour la convergence uniforme. Une question intéressante est la possibilité de réduire l'écart.

Pour cela, nous améliorons la borne inférieure. Etant donnée un ensemble de classifieurs H de VC-dimension d, on considère une distribution U_d et un algorithme A minimisant l'erreur empirique. Considérons x un ensemble de d points pulvérisé par H. Notons U_d la distribution de probabilité discrète uniforme sur $x \times \{0,1\}$, c'est à dire $P_{U_d}(\{(x_i,0)\} \cup \{(x_i,1)\}) = 1/2d$. L'erreur Bayésienne de U_d est égale à $\frac{1}{2}$.

2.4 Leave one Out, cross-validation et complexité d'échantillon

L'estimateur Leave-One-Out est calculé en éxécutant l'algorithme d'apprentissage n fois, chaque fois en ôtant un des n exemples d'apprentissage, et en testant le classifieur résultanta sur l'exemple supprimé; la proportion de tests ratés est l'estimateur leave-one-out. Considérer non plus n tests en utilisant un élément comme ensemble test mais D tests en utilisant n/D éléments comme ensemble test amène à la cross-validation en D-sous-ensembles. On pourrait imaginer, au lieu de partitionner en D sous-ensembles, de tester tous les sous-ensembles de taille n/D donnés. Nous n'étudierons pas ce cas ici. Le leave-one-out est en général considéré comme plus efficace que l'erreur empirique. L'intuition empirique des méthodes de rééchantillonage consiste en tester les classifieurs sur des données non observées.

Kearns & Ron [Kearns et al, 1997] proposent des bornes sur la complexité d'échantillon en considérons l'erreur leave-one-out au lieu de l'erreur empirique.

Théorème 2.6 *La complexité d'échantillon vérifie, en ignorant les dépendances en* δ*,*

$$n_A(\varepsilon,\delta) = \begin{cases} O\left(\frac{d}{\varepsilon^2}\right) \\ \Omega\left(\frac{d}{\varepsilon}\right) \end{cases}$$

Ces résultats laissent penser que des améliorations sont possibles et que le leave-one-out est un meilleur estimateur dans le cadre de la convergence uniforme dans le cadre agnostique. Toutefois, nous montrons une borne inférieure environ égale à la borne supérieure proposée par Kearns et Ron. En fait, sur la distribution de probabilité U_d, la probabilité pour que l'erreur leave-one-out soit exactement égale à l'erreur empirique tend vers 1 comme la taille n de l'ensemble d'apprentissage augmente.

2.5 Résultats principaux

Considérons les X_i variables aléatoires iid distribuées suivant U_d. Considérons $0 < \epsilon < \frac{1}{2}$. Définissons $N_i = card\{i/X_i = x_i\}$, $\Delta_i = |card\{i/X_i = x_i \wedge Y_i = 1\} - card\{i/X_i = x_i \wedge Y_i = 0\}|$. $\Delta = \sum \Delta_i$. Considérons ϵ_1,$t \geq 0$. Définissons $N = n/d - \epsilon_1 n$ et $N' = n/d + \epsilon_1 n$.

Considérons les évènements suivants:

A_i : $N_i \geq n/d - \epsilon_1 n = N$. $A = \cap A_i$.

A'_i : Pour $y \in \{0,1\}$, $card\{i/X_i = x_i \wedge Y_i = y\} \in [n/(2d) - \epsilon_1 n/2, n/(2d) + \epsilon_1 n/2]$. $A' = \cap A'_i$. Notez que $A' \Rightarrow A$.

B_i : $\Delta_i > 1$. $B = \cap_{i=1}^{d} B_i$.

C : $\Delta/n \geq \epsilon$

D : $\exists i/\Delta_i = 0$

E : $card\{i/N_i \text{ est pair}\} \geq \lceil d/2 \rceil$

η: $N_i = \eta_i$ (on identifie une famille de d-uples d'entiers et une famille d'évènements).

Les preuves des lemmes suivants sont donnés en partie A.

Lemme 2.7 (Erreur empirique) *La différence entre l'erreur empirique et l'erreur en généralisation après apprentissage par l'algorithme de minimisation du risque empirique est* $\frac{\Delta}{2n}$.

Lemme 2.8 (Erreur Leave-one-out, cas général) *Si B a lieu, alors l'erreur leave-one-out est égale à l'erreur empirique.*

Lemme 2.9 (Erreur Leave-one-out, cas particulier) *Si $D \cap A'$ a lieu, alors la différence entre l'erreur en généralisation et l'erreur leave-one-out est minorée par* $\frac{1}{2d} - \frac{d+1}{2}\epsilon_1$.

Lemme 2.10 (Tout plein de probabilités) *On a les inégalités suivantes:*

$$P(A) \geq 1 - d\exp(-2n\epsilon_1^2) \tag{2.1}$$

$$P(C|A) \geq 1 - \frac{1}{1 + \frac{t^2}{n + d\epsilon_1 n}} \text{ avec } 2\epsilon = \sqrt{\frac{d - d^2\epsilon_1}{2n}} - t/n \tag{2.2}$$

$$P(B|A) \geq \left(1 - 2\sqrt{\frac{1}{2(n/d - \epsilon_1 n)\pi}}\right)^d \tag{2.3}$$

$$P(D \cap A') \geq (\frac{1}{2} - 4d\exp(-n\epsilon_1^2)) \times \left(1 - \left(1 - \frac{\exp(-1/12)}{\sqrt{2N'\pi}}\right)^{\lceil d/2 \rceil}\right) \tag{2.4}$$

Les théorèmes suivants ont lieu pour la minimisation du risque empirique, pour tout ϵ_1 et t.

Théorème 2.11 (Erreur empirique) *La différence entre l'erreur empirique et l'erreur en généralisation est minorée par* $\frac{1}{2}(\sqrt{\frac{d - d^2\epsilon_1}{2n}} - t/n)$ *avec probabilité au moins* $P(A)P(C|A)$.

Preuve: Si C a lieu, alors la différence entre l'erreur empirique et l'erreur en généralisation est minorée par ϵ (lemmes 7,10). □

Remarquez que la preuve développée incluant le lemme 7 et la part nécessaire du lemme 10 est très courte et intuitive.

Théorème 2.12 (Erreur Leave-One-Out, cas général) *La différence entre l'erreur en généralisation et l'erreur leave-one-out est minorée par* $\frac{1}{2}(\sqrt{\frac{d - d^2\epsilon_1}{2n}} - t/n)$ *avec probabilité au moins* $P(A)(P(B|A) + P(C|A) - 1)$.

Preuve: Si B a lieu, alors l'erreur leave-one-out est égale à l'erreur empirique (lemme 8). Le théorème 11 conclut.□

Une fois encore, la preuve est courte et intuitive.

Théorème 2.13 (Erreur Leave-One-Out, cas particulier) *La différence entre l'erreur en généralisation et l'erreur empirique pour un nombre pair d'exemples est minorée par* $\frac{1}{2d} - \frac{d+1}{2}\epsilon_1$ *avec probabilité au moins* $P(D \cap A')$.

(la condition selon laquelle le nombre d'exemples est pair peut être supprimée avec un coût faible sur la confiance pourvu que la VC-dim est supérieure à 1)

Preuve: Ceci est une conséquence des lemmes 9 et 10. L'idée est que lorsqu'un point parmi les x_i a le même nombre d'instances positives et négatives, alors il y a un problème dans le leave-one-out.□

2.6 Corollaires

Corollaire 2.14 (Erreur empirique) *Si* $n \geq \frac{25d^2 \ln(6d)}{2}$ *et* $\delta < \frac{25}{36}$, *alors la précision garantie avec confiance au moins* δ *est minorée par*

$$\frac{1}{\sqrt{10}}\sqrt{\frac{d}{n}} - \frac{\frac{36}{25}\delta}{(1-\frac{36}{25}\delta)2\sqrt{n}}$$

La complexité d'échantillon garantissant avec confiance $1-\delta > \frac{11}{36}$ *une précision* $\epsilon \leq \frac{d\sqrt{2}(\sqrt{10}-\delta')}{5\sqrt{\ln(6d)}}$ *est au moins* $\frac{(\sqrt{10}-\delta')^2}{\epsilon^2}$, *avec* $\delta' = \frac{\frac{36}{25}\delta}{2(1-\frac{36}{25}\delta)}$.

Preuve: En posant $\epsilon_1 = \frac{1}{5d}$ et $t = \frac{\delta}{1-\delta}/\sqrt{n}$ dans le théorème 11.□

Corollaire 2.15 (Comportement asymptotique) *Comme* $n \to \infty$ *pour un* δ *donné, la précision de l'estimateur empirique de l'erreur en généralisation pour l'algorithme de minimisation du risque empirique est aymptotiquement plus grande qu'un équivalent de* $\frac{1}{2}(\sqrt{d/2} - \frac{\delta}{1-\delta})/\sqrt{n}$ *pour le seuil de confiance* $1-\delta$.

Ceci implique qu'asymptotiquement pour une précision $\to 0$, *la complexité d'échantillon est minorée par* $\frac{(\sqrt{d/2}-\frac{\delta}{1-\delta})^2}{4\epsilon^2}$.

Preuve:

Soit $\epsilon_1 = o(1/\sqrt{n})$ et $t\sqrt{n} = \delta/(1-\delta)$ dans le théorème 11.□

Corollaire 2.16 (Leave-One-Out: cas général) *Pour tout* $\delta \leq \frac{4}{5}(5/66 - \frac{1}{d\sqrt{70\Pi}})$, *la complexité d'échantillon garantissant une précision* ϵ *avec confiance au moins* $1-\delta$ *est minorée par*

$$\frac{(\frac{1}{2}\sqrt{\frac{d-\frac{1}{5}}{2}} - \frac{1}{\sqrt{11}})^2}{\epsilon^2}$$

pourvu que cette quantité soit $\geq 175d^3$.

Preuve: En posant $\epsilon_1 = \frac{1}{5d}$ et $t = \sqrt{n/11}$ dans le théorème 12. □

Corollaire 2.17 (Leave-One-Out: cas particulier) *La complexité d'échantillon garantissant une précision au moins* $\frac{2d-1}{6d^2}$ *avec confiance* $1-\delta$ *est au moins*

$$\frac{d^3 \times \left(\frac{\exp(-1/12)\sqrt{2\pi(1+1/(3d))}-0.15}{16\pi(1+1/(3d))}\right)^2}{\delta^2} - 1$$

pourvu que cette quantité soit plus grande que $18d^4\ln(16d)$.

Une autre borne inférieure possible, pour toute précision $< \frac{1}{2}$, *est* $\exp(-1/6)/(2\pi\delta^2)$.

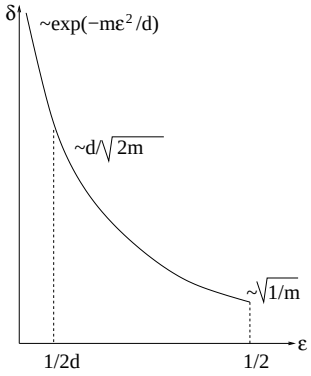

FIG. 2.1 – *Borne inférieure sur le risque δ pour l'estimation leave-one-out de la minimisation du risque empirique.*

Donc, le leave-one-out n'est pas PAC, puisqu'il n'est pas polynomial en $\ln(1/\delta)$. En outre, la complexité d'échantillon n'est pas linéaire en la VC-dimension.

Preuve: En posant $\epsilon_1 = 1/(3d^2)$ dans le théorème 13. Le "moins un" est la à cause de la condition sur n dans le théorème 13 (n pair).

Le second résultat est une optimisation pour $d = 1$ et peut être utilisé bien sûr pour toute VC-dimension au moins 1.□

La figure 2.1 illustre la borne inférieure sur le risque δ d'une différence au moinsϵ entre l'erreur en généralisation et l'erreur leave-one-out.

2.7 Optimalité de cette distribution

On considère l'ensemble des distributions sur d points pulvérisés, et prouvons que notre distribution U_d utilisée dans les preuves ci-dessus est optimale en un certain sens.

Soit les α_i les probabilités de x_i, et β_i la probabilité de $Y = 1$ conditionnellement à $X = x_i$. Supposons (sans perte de généralité) que $\beta_i \leq \frac{1}{2}$.

Considérons la variable aléatoire ϵ égale à la différence entre l'erreur en généralisation et l'erreur empirique, multipliée par $\sqrt{n}$.

Notons $\tilde{B}(n,p)$ la variable aléatoire $min(X,n-X)$ avec X la loi binomiale $B(n,p)$ de paramètre p. Considérons des tirages successifs des X_i et des Y_i, avec les Y_i, pour $X_i = x_d$, étant tirés en dernier. ϵ est un C (constant, après le tirage des X_i, et le tirage des Y_i pour $Y_i = x_j$ avec $j < d$) plus $\frac{1}{\sqrt{n}}(\tilde{B}(n_i,\beta_i) - \beta_i n_i)$, avec n_i le nombre de X_j tels que $X_j = x_d$. On néglige, dans la suite, les cas pour lesquel $n_i < An$ pour un A donné compris entre 0 et α_i: la probabilité de tels cas décroit exponentiellement, et donc n'interfère pas avec les équivalents asymptotiques d'erreur.

Ainsi, $\frac{1}{\sqrt{n}}\tilde{B}(n_i,\beta_i) - \beta_i n_i$ converge en loi vers

$$\sqrt{\alpha_i}\tilde{\mathcal{N}}_{n_i(1-\beta_i)}(0,\beta_i(1-\beta_i))$$

avec $\tilde{\mathcal{N}}_M(0,\sigma^2)$ égal à la loi normale de moyenne 0 et variance σ^2, mais à valeurs x plus grands que M remplacées par $2M - x$. Pour $\beta_d < \frac{1}{2}$ ceci converge vers

$$\sqrt{\alpha_i}\mathcal{N}(0,\beta_i(1-\beta_i))$$

Pour toute valeur de C on accroit l'espérance de la valeur absolute en augmentant β_d. Ainsi, $\beta_d \to \frac{1}{2}$ est meilleur. Toutefois, il y a une discontinuité en $\frac{1}{2}$ (remplacement de $\mathcal{N}$ par $\tilde{\mathcal{N}}_0$) qui est en faveur de $\beta_d = \frac{1}{2}$. Ceci implique que $\beta_d = \frac{1}{2}$ est optimal. L'équivalence entre les β_i implique que tous les β_i devraient être choisis égaux à $\frac{1}{2}$.

La concavité de $x \mapsto \sqrt{x}$ implique l'optimalité de $\alpha_i = \frac{1}{d}$.
Ainsi, nous avons prouvé le résultat suivant:

Proposition 2.18 *La distribution résultant de $\alpha_i = \frac{1}{d}$ et $\beta_i = \frac{1}{2}$ (c'est-à-dire U_d) est la distribution optimale parmi d points, pour maximiser l'équivalent asymptotique de l'espérance de la différence entre l'erreur en généralisation et l'erreur empirique.*

2.8 Extension à la cross-validation en D-sous-ensembles

Un point intéressant à propos de la cross-validation est le fait que les praticiens sont d'accord sur le fait que des valeurs de D tendant vers l'infini pour une partition en D sous-ensembles amène des résultats médiocres, en suggérant des valeurs bornées $D \simeq 10$. Le sujet de cette section est la preuve formelle de ce fait dans le formalisme PAC.

On considère la même distribution U_d que précédemment sur d points pulvérisés.
Pour tout $i \in [1,d]$, nous allons:

1. Evaluer la proportion de points tirés en x_i.
2. Evaluer la même quantité dans le j^e sous-ensemble.
3. Evaluer la difference entre le nombre d'exemples positifs en x_i et le nombre d'exemples négatifs.
4. Evaluer la même quantité dans le j-ième sous-ensemble.
5. Conclure qu'avec forte probabilité, pourvu que D soit assez grand, supprimer le j^e sous-ensemble ne va pas modifier un vote à la majorité en x_i, et donc que l'erreur évaluée par cross-validation en D sous-ensembles est égal à l'erreur empirique.

Ceci est fait plus formellement ci-dessous, dans le cas de $n = DN$ avec D et N entiers (pour préserver la clarté):

1. Avec probabilité au moins $1 - \delta_1$, la proportion de points tirés en x_i est au moins:

$$\frac{1}{d} - \epsilon_1 \qquad \text{où } \epsilon_1 = \sqrt{-\frac{\ln(\delta_1)}{2n}} \tag{2.5}$$

   (ceci est une conséquence de l'inégalité de Hoeffding)
2. Avec probabilité au moins $1 - \delta_2$, la proportion de points tirés en x_i dans le j^e sous-ensemble est au plus:

$$\frac{1}{d} + \epsilon_2 \qquad \text{où } \epsilon_2 = \sqrt{-\frac{\ln(\delta_2)D}{2n}} \tag{2.6}$$

3. Conditionnellement à l'évènement du point 1 (équation 2.5), la différence entre le nombre d'exemples positifs et le nombre d'exemples négatifs (en valeur absolue) en x_i a une espérance minorée par $\sqrt{\frac{n}{2}(1/d - \epsilon_1)}$, et variance majorée par $\frac{1}{4}n(1/d + \epsilon_1)$. Ainsi par l'inégalité de Chebyshev-Cantelli (voir par exemple [Devroye et al, 1996, annexe A.5]), cette différence est, avec probabilité $\geq 1 - \delta_3$, au moins égale à:

$$\sqrt{\frac{n}{2d}(1 - d\epsilon_1)} - \sqrt{\frac{1 - \delta_3}{\delta_3}\frac{n}{4}(1/d + \epsilon_1)} \tag{2.7}$$

4. Conditionnellement à l'évènement du point 2, la différence entre le nombre d'exemples positifs et le nombre d'exemples négatifs (en valeur absolue) en x_i est majorée par

$$2\sqrt{\frac{n}{Dd}}\sqrt{\frac{\ln(2D/\delta_4)}{(1+d\epsilon_2)}} \tag{2.8}$$

dans *chacun* des D sous-ensembles, avec probabilité au moins δ_4.

5. Quand la quantité bornée par l'équation 2.8 est plus petite que la quantité minorée à l'équation 2.7 (ce qui arrive pour tout i, uniformément, si D est suffisamment grand, avec probabilité au moins $(1-\delta_1-\delta_2-\delta_3-\delta_4)^d$), alors l'erreur évaluée par cross-validation sur D sous-ensembles est égale à l'erreur empirique.

Ceci implique que le corollaire 14 a lieu (avec des constantes différentes) dans le cas de la cross-validation en D sous-ensembles, pourvu que $D \to \infty$ comme $n \to \infty$ (ou même simplement, pourvu que $\liminf D$ soit suffisamment grand). Plus précisément:

Proposition 2.19 *Si δ_1, δ_2, δ_3, δ_4 et D sont tels que l'expression 2.8 est plus petite que l'équation 2.7, alors avec probabilité au moins $(1-\delta_1-\delta_2-\delta_3-\delta_4)^d$ la cross-validation en D sous-ensembles est égale à l'erreur empirique.*

Donc, pour D suffisamment grand, *la complexité d'échantillon nécessaire pour garantir une précision ϵ avec confiance $1-\delta$ pour la cross-validation en D sous-ensembles, pour ϵ et δ suffisamment petits, est $\Omega(\frac{d}{\epsilon^2})$.*

2.9 Remarques et travail à suivre

L'étude du leave-one-out dans quelques cas particuliers est faite dans [Devroye et al, 1996].

Une question intéressante est: est-ce-qu'une version du corollaire 17 a lieu pour la cross-validation en D sous-ensembles? Le coeur de la preuve du corollaire 17 est le fait que l'égalité peut avoir lieu entre le nombre de points positifs et le nombre de points négatifs sur un x_i donné, et qu'alors, l'estimateur leave-one-out a un très mauvais comportement. Un comportement analogue peut être souligné dans la cross-validation en D sous-ensembles. La preuve détaillée pourrait être l'objet d'un travail futur.

Une autre question intéressante pourrait être la quantification de D. Ceci pourrait être aisément réalisé par comparaison des équations 2.7 et 2.8, mais des optimisations étant possibles, nous préférons étudier ce point dans un travail futur.

Des résultats positifs à suivre pourraient inclure des *a priori* sur la distribution de Y conditionnellement à X, ou sur la distribution marginale sur X. Il est probable que, pourvu que $P(Y|X)$ soit loin de $\frac{1}{2}$, des améliorations importantes soient possibles.

Nos résultats suggèrent que, dans le pire cas, le leave-one-out est un estimateur pire que l'erreur empirique. Ceci est faux dans la pratique. Cela implique que des hypothèses devraient être incluses pour rendre la théorie plus adaptée à la pratique.

Bibliographie

[Anthony et al, 1999] M. Anthony, P.L. Bartlett, *Neural Network Learning: Theoretical Foundations, Cambridge University Press, 1999.*
[Devroye et al, 1996] L. Devroye, L. Gyorfi, G. Lugosi, *A probabilistic theory of pattern recognition, 1996.*
[Gavin et al, 2001] G. Gavin, O. Teytaud, Lower bounds for empirical and leave-one-out estimates proceedings of Ijcnn 2001.

[Long, 1998] P.M. Long, *The Complexity of Learning According to Two Models of a Drifting Environment, Proceedings of the 11th Annual Conference on Computational Learning Theory, 116-125, ACM press, 1998.*
[Kearns et al, 1997] Leave-one-out stability, *M. Kearns, D. Ron, Algorithmic Stability and Sanity-Check Bounds for Leave-One-Out Cross Validation, AT& T Labs research Murray Hill, New jerzey and MIT Cambridge, MA, 1997*
[Alexander, 1984] K. Alexander, *Probabilities Inequalities for Empirical Processes and a Law of the Iterated Logarithm, Annals of Probability, volume = 4, pages = 1041-1067, 1984*
[Talagrand, 1994] M. Talagrand, *Sharper Bounds for Gaussian and Empirical Processes, The Annals of Probability, 22(1), 28-76, 1994*
[Vapnik, 1982] V.N. Vapnik, *Estimation of Dependences Based on Empirical Data, 1982.*

Annexe A

Proofs

Lemmas 7,8,9 are direct consequences of the definition. We now consider the proof of lemma 10.

- first step: $P(A)$
  Equation (2.1) is simply a consequence of Hoeffding's inequality.
- second step: $P(B|\eta)$

$$P(\neg B_i|A) \leq 2\sqrt{\frac{1}{2N\pi}} \text{ (see [Devroye et al, 1996, appendix A.8])}$$

Notice that factor 2 can be removed for even values. This suggests that this could be optimized.

$$P(B_i|A) \geq 1 - 2\sqrt{\frac{1}{2N\pi}}$$

$$P(B|\eta) = \pi_i P(B_i|\eta)$$

(as the B_i are independent, when η is assumed)

$$P(B|\eta) \geq \pi_i(1 - 2\sqrt{\frac{1}{2N_i\pi}})$$

In particular

$$P(B|A) \geq (1 - 2\sqrt{\frac{1}{2N\pi}})^d$$

$$P(B|A) \geq (1 - 2\sqrt{\frac{1}{2(n/d - \epsilon_1 n)\pi}})^d$$

This is equation 2.3.

- third step: $P(C|A)$ (equation 2.2)

$$E(\Delta|A) \geq d\sqrt{N/2}$$

$$Var(\Delta|A) \leq d(2n/d - N)$$

The first result is extracted from [Devroye et al, 1996, appendix A.8], the second simply consists in bounding the variance of the absolute value of a random variable by its variance. So, by Chebyshev-Cantelli inequality (see [Devroye et al, 1996, appendix A.5]) applied to $-\Delta'$

$$P(-\Delta - E(-\Delta) > t|A) \leq d(2n/d - N)/(d(2n/d - N) + t^2)$$

$$P(\Delta < d\sqrt{N/2} - t|A) \leq \frac{1}{1 + \frac{t^2}{n + d\epsilon_1 n}}$$

$$P(\Delta/n < \frac{d\sqrt{N/2} - t}{n}|A) \leq \frac{1}{1 + \frac{t^2}{n + d\epsilon_1 n}}$$

$$P(\Delta/n < \sqrt{(d - d^2\epsilon_1)/(2n)} - t/n)|A) \leq \frac{1}{1 + \frac{t^2}{n + d\epsilon_1 n}}$$

This is exactly $P(\neg C|A)$ for $\epsilon = (\sqrt{(d - d^2\epsilon_1)/(2n)} - t/n)$.

- fourth step: $P(E) \geq \frac{1}{2}$ if n is even.
  The number of even N_i for n even can be seen as a Markov chain, with parameter $n/2$. The initial value is d, and the result is proved by induction on $k = n/2$.
- fifth step: $P(E \cap A') \geq P(A') + \frac{1}{2} - 1$.
- sixth step: $P(D \cap A')$.
  Assume that $\eta \Rightarrow A'$ and that η_1 is even.

$$P(\Delta_1 = 0|\eta) \geq \frac{exp(-1/12)}{\sqrt{2N'\pi}}$$

(see [Devroye et al, 1996, appendix A.8])

$$P(\Delta_1 \neq 0|\eta) \leq 1 - \frac{exp(-1/12)}{\sqrt{2N'\pi}}$$

Assume that $\eta_1, \ldots, \eta_{\lceil d/2 \rceil}$ are even. Then

$$P(\forall i \in [1, \lceil d/2 \rceil]/\Delta_i \neq 0|\eta) \leq (1 - \frac{exp(-1/12)}{\sqrt{2N'\pi}})^{\lceil d/2 \rceil}$$

The order of the x_i can be chosen arbitrarily without loss of generality. So, this lower bounds the probability of $D|\eta$ for the η's such that $\eta \Rightarrow E$. This, in conjunction with the previous point, implies that

$$P(D \cap (E \cap A')) \geq (P(A') - \frac{1}{2}) \times (1 - (1 - \frac{exp(-1/12)}{\sqrt{2N'\pi}})^{\lceil d/2 \rceil})$$

$$P(D \cap A') \geq (P(A') - \frac{1}{2}) \times (1 - (1 - \frac{exp(-1/12)}{\sqrt{2N'\pi}})^{\lceil d/2 \rceil})$$

- seventh step: $P(A')$.
  $P(\neg A') \leq 4d \times \exp(-\frac{n\epsilon_1^2}{2})$ by Hoeffding's inequality. This, in conjunction with the point above, leads to equation 2.4.

Chapitre 3

Apprentissage Robuste

Résumé

Les préoccupations de ce chapitre se concentrent sur deux thèmes:

- Robustesse au bruit: comment modifier un algorithme d'apprentissage pour qu'il prédise bien des sorties alors qu'on ne lui fournit que des données correspondant à des capteurs peu fiables? Les résultats concernent essentiellement le bruit en sortie.
- Robustesse au passage sur sous-zone: ici l'algorithme d'apprentissage travaille sur une distribution donnée, mais l'utilisateur peut avoir besoin de travailler seulement sur une zone restreinte.

Ce chapitre est rédigé à partir d'un travail réalisé en collaboration avec la société Elf
Les parties théoriques ont fait l'objet d'une publication ([Teytaud, 2001]).

Table des matières

3.1 Théorie de l'apprentissage avec bruit. Extension à la régression et étude pratique

3.1.1 Introduction

La théorie de l'apprentissage se préoccupe le plus souvent de déterminer une taille d'échantillon suffisante ou nécessaire pour garantir qu'à ϵ-près, avec confiance $1-\delta$, le modèle sélectionné pour modéliser un phénomène est "bien" choisi à travers une minimisation (éventuellement approximative) de l'erreur empirique. L'introduction de la notion de pire cas et d'uniformité des bornes en la distribution des exemples conduit alors à la notion de VC-dimension, permettant des bornes non-asymptotiques.

La théorie de l'apprentissage se scinde alors en deux grandes catégories:

- le cas de données pour lesquelles la famille de modèles peut atteindre une erreur aussi petite que désiré pourvu que les paramètres soient suffisamment bien choisis. Dans ce cas, la complexité d'échantillon est grosso-modo linéaire en la VC-dimension, linéaire en $1/\epsilon$, et logarithmique en δ.
- le cas de données pour lesquelles la famille de modèles ne peut descendre indéfiniment; pour une raison ou une autre (famille de modèles mal choisie, entrées insuffisantes, erreur bayésienne intrinsèque non nulle), il existe un taux d'erreur résiduelle inévitable. La complexité est alors grosso-modo linéaire en la VC-dimension, quadratique en $1/\epsilon$, et logarithmique en δ.

Des résultats intermédiaires dans le cas d'un faible taux d'erreur minimal existent: Vapnik dans [Vapnik, 1982] obtient ainsi une dépendance en ϵ proportionnelle non plus à ϵ^2 mais à $\epsilon^2/(L+\epsilon)$, avec L une borne sur le taux d'erreur minimal en classification.

Depuis quelques années, un cas particulier a été dégagé dans le deuxième cadre: le cas où un bruit explique la barrière infranchissable du taux d'erreur, et où si l'on pouvait supprimer le bruit, le taux d'erreur tendrait vers 0 (ou simplement, serait réduit). C'est-à-dire que l'on suppose bien que la famille de modèles est bien choisie, mais pas que le bruit est nul. Formellement, l'algorithme ne dispose plus d'entrées (X,Y) parfaites, mais d'entrées $(X,\hat{Y})$, avec $\hat{Y}$, dans le cas de la classification deux classes $\{-1,1\}$, égal à Y avec probabilité $1-\eta$ et à $-Y$ avec probabilité η, l'apparition de bruit étant supposée indépendante identiquement distribuée. Un modèle un peu plus général est le modèle CPCN, "Constant Partitionning Classification Noise"; ici η est simplement supposé constant sur chaque zone d'une partition finie du produit entrée/sortie. Un modèle encore plus général, cas des erreurs malicieuses, suppose que l'*apparition* de bruit est indépendante identiquement distribuée, mais que le bruit se manifeste par une interversion de classe selon le bon vouloir d'un adversaire sans limite de capacités de calcul, ayant accès à toute l'information désirée, même l'état interne de l'algorithme d'apprentissage.

Le modèle des statistical queries (SQ) proposé par Kearns et très utilisé dans le domaine de la théorie de l'apprentissage avec bruit doit être détaillé: au lieu de disposer, comme dans le cas PAC, d'un oracle fournissant des exemples, un algorithme SQ dispose d'un oracle, fournissant, pour un prédicat Q sur l'espace des exemples et une précision τ, une évaluation à τ près de la probabilité pour qu'un exemple vérifie Q. La précision se doit d'être polynomiale en les mêmes paramètres que les algorithmes PAC.

3.1.2 Etat de l'art

Taux de bruit constant

On peut résumer rapidement les résultats connus sur ce modèle, provenant de Simon, Talagrand, Aslam, Decatur, ([Aslam et al, 1996, Decatur, 1997, Decatur, 1995, Kearns et al, 1993, Gentile et al, 1998]) par les bornes suivantes sur la complexité d'échantillon[1]:

1. $m(\epsilon,\delta,\eta) = O(\frac{VC-\log(\epsilon\delta(1-2\eta))}{\epsilon^2(1-2\eta)^2})$
2. $m(\epsilon,\delta,\eta) = \Omega(\frac{VC-\log(\delta)}{\epsilon(1-2\eta)^2})$
3. Aux facteurs logarithmiques près, pour toute classe apprenable en temps polynomial $poly(n)$ pour la loi des exemples, $m(\epsilon,\delta,\eta) = O(\frac{poly(n)-\log(\delta)}{\epsilon^2(1-2\eta)^2})$ avec $poly$ un polynome et n la longueur des instances. La complexité en temps est en outre polynomiale.
4. Si la classe est SQ, alors m a une dépendance en ϵ en $O(\frac{(\log 1/e)^2(\log\log 1/\epsilon)^2}{\epsilon})$.
5. Si un algorithme est SQ, alors il est PAC dans le cas CPCN.

Par réalisme, les algorithmes se doivent d'utiliser non η, mais une borne sur η; la dépendance étant alors en cette borne.

Les bornes précédentes sont valables seulement dans le cadre d'un apprentissage pour lequel l'erreur minimale dans le cas sans bruit est nul. On peut s'interroger sur le cas où cette erreur minimale est non nulle, mais petite. L'écart entre les deux bornes $1/m$ et $1/\sqrt{m}$ constaté selon que l'on travaille à erreur minimale nulle et erreur minimale non nulle est une préoccupation ancienne et Vapnik dans [Vapnik, 1982] montre que la probabilité d'avoir un écart plus grand que $1/\epsilon$ entre l'erreur de la fonction minimisant l'erreur empirique et l'erreur minimale est majorée par $8m^V e^{-\frac{m\epsilon^2}{8(L+\epsilon)}}$ si l'erreur minimale est majorée par L (si la VC-dim V est ≥ 3, sinon $8(m+1)^V e^{-\frac{m\epsilon^2}{8(L+\epsilon)}}$). Cela laisse supposer que les résultats de complexité précédents, valables dans le cas où l'erreur minimale est nulle, pourraient avoir des extensions au cas d'erreur minimale du cas sans bruit est légère.

Erreurs malicieuses

Les résultats proviennent de [Kearns et al, 1993, Decatur, 1995]:

- S'il existe deux classes pulvérisées par les points[2] (hypothèse très légère!), alors $m(\epsilon,\delta,\eta)$ n'est jamais fini si on n'a pas $\eta < \frac{\epsilon}{1+\epsilon}$.
- Cas de classes déséquilibrées, apprentissage mono-classe: s'il existe des points $x_1,x_2,\ldots,x_t,x'$ et des classes $c_1,\ldots,c_t$ vérifiant

$$x_i \in c_j \iff i \neq j$$
$$\forall j \; x' \in c_j$$

(hypothèse raisonnable malgré les apparences, beaucoup plus légère qu'une contrainte de VC-dimension!) pour $t > 1$ alors η (taux d'erreurs *malicieuses*) doit être plus petit que $\frac{\epsilon}{t-1}$ si l'on n'a que des exemples positifs.

1. Il est important de remarquer qu'ϵ est la précision obtenue sur le taux d'erreur *dans le cas sans bruit.*
2. Il ne s'agit pas d'une erreur de frappe... Cette notion est une notion duale de celle de pulvérisation d'un ensemble de points par une famille d'applications.

- Cas de classes déséquilibrées (i.e. on suppose ici qu'on a des exemples d'une seule classe), même sans que les erreurs soient malicieuses: le taux d'erreur maximal acceptable pour la finitude de $m(\epsilon,\delta,\eta)$ est inférieur à $\frac{\epsilon}{1+\epsilon}$, si la classe est positivement incomparable, c'est-à-dire s'il existe 3 points u v et w et deux classes c_1 et c_2 tels que $u \in c_1 \setminus c_2$, $v \in c_1 \cap c_2$, $w \in c_2 \setminus c_1$ (hypothèse très légère!). Si $\delta \leq 1/d$ et la VC-dim au moins d, alors le taux d'erreurs *malicieuses* maximal acceptable est inférieur à $\epsilon/(d-1)$.
- Si un algorithme est SQ, et si l'erreur optimale dans le cas sans bruit est nulle, alors il existe un algorithme PAC résistant à un taux d'erreur malicieuses de $\Omega(\epsilon/\log(1/\epsilon))$, avec une complexité d'échantillon $O(\frac{(\log 1/e)^2(\log\log 1/\epsilon)^2}{\epsilon})$ (on ne garde ici que les dépendances en ϵ).
- Si une classe est apprenable dans le cas sans erreur avec une complexité d'échantillon en $m(\epsilon)$, alors on peut PAC-apprendre avec bruit en complexité d'échantillon $O(m(\epsilon)^3)$, en résistant à un bruit malicieux de taux $min(\epsilon/8, \log(m)/m)$. L'algorithme est détaillé en partie 3.1.4. Au prix d'une constante multiplicative sur le taux d'erreurs malicieuses toléré, on peut se ramener à un complexité d'échantillon $O(m(\epsilon))$ (voir [Decatur, 1995, p101]).

3.1.3 Quel modèle de bruit?

Le bruit additif, ou d'autres formes particulières de bruit, sont hors-sujet ici. L'idée sous-jacente est une proportion petite de points endommagés, arbitrairement dangereusement pour l'apprentissage. Ces points sont dits "aberrants".

En partie 3.1.3, nous définissons un autre modèle de bruit. En partie 3.1.3, on considère quelques conséquences de nos résultats au regard des autres modèles de bruit.

Un nouveau modèle de bruit

On considère deux modèles et les montrons ensuite presqu'équivalents:

- Bruit de sortie: un exemple (X,Y) est mal étiqueté avec probabilité $p(X,Y)$. Le taux de bruit η est la proportion d'exemples bruités, ie Ep.
- Bruit général: une proportion η d'exemples ne sont pas distribués selon la loi initiale P, mais selon une autre loi Q.

Le bruit en sortie peut être modélisé par un bruit général: Q est la loi des $(X,1-Y)$ avec X distribué selon P conditionnellement à l'occurence de bruit. Réciproquement, le bruit général avec η et Q peut être modélisé par des bruits en sortie dès que $\neg Q$ est absolument continu par rapport à P, avec p la densité de $\neg Q$ par rapport à P.

Alors, les résultats suivants peuvent facilement être établis, pour une famille de VC-dimension finie (des extensions sont possibles pour des résultats non-indépendants de la distribution):

- Si un algorithme d'apprentissage ne minimise pas asymptotiquement l'erreur empirique (ie si $E_m(f_m) = \frac{1}{m}\sum_{i=1}^{m} f_m(z_i)$ ne converge pas vers $\inf_f E(f)$, avec f_m la fonction de coût associée au classifieur sélectionné par l'algorithme d'apprentissage, f une fonction de coût associée à un classifieur, E l'espérance dans le cas sans bruit, et $z_i = (X_i,Y_i)$ le i^e exemple), alors considérons un taux d'erreur nul. Alors $|E_m(f_m) - E(f_m)| \to 0$. Ainsi, $E(f_m) \not\to \inf_f E(f)$. Aussi, l'algorithme ne réussit pa à fournir une complexité d'échantillon finie pour $\epsilon \leq \frac{\eta}{1-\eta}$.
- Considérons maintenant un algorithme d'apprentissage qui minimise l'erreur empirique. Alors, distinguons E^p l'espérance pour P, E^Q l'espérance pour Q. Soit f_m choisi par l'algorithme et soit g minimisant l'erreur dans le cas sans bruit.

$$E_m^{(1-\eta)P+\eta Q} f_m \leq E_m^{(1-\eta)P+\eta Q} g$$
$$E_m^P f_m + \eta' E_m^Q f_m \leq E_m^P g + \eta' E_m^Q g$$

with $\eta' = \frac{\eta}{1-\eta}$.

$$Ef_m \leq L + \eta' E_m^Q g - E_m^q f_m + O(\sqrt{\frac{V}{m}})$$

avec $L = E^p g$ et V la VC-dimension. Ceci est une convergence en $1/\sqrt{m}$ vers au plus $\eta/(1-\eta)$.

– Considérons le même cas, avec $L = 0$. Alors

$$E_m^P f_m + \eta' E_m^Q f_m \leq E_m^P g + \eta' E_m^Q g$$

amène à

$$E_m^P f_m \leq O(V/m) + \eta' E_m^Q g$$

$$E f_m \leq \eta' + O(\sqrt{\frac{V}{m}(\frac{V}{m} + \eta')})$$

Pour tout $A \in]0,1[$, utiliser $m \geq H/\epsilon$ pour H suffisamment grand ($H = \Omega(V/(1-A))$) amène une complexité d'échantillon borné linéairement en $O(1/\epsilon)$ résistant à un taux d'erreur η tel que $\frac{\eta}{1-\eta} < A\epsilon$.

– On peut aisément trouver P et Q tels que pour la minimisation du risque empirique, $\frac{\eta}{1-\eta}$ est le taux d'erreur optimal. Ceci suppose seulement que deux fonctions dans la classe coïncident sur un point et ne coïncident pas sur un autre. Notez que ceci peut être fait aussi bien avec bruit généraliste qu'avec bruit en sortie.

Discussion

Notre modèle simple a les avantages suivants:

– Il est plus naturel (du moins pour nous!) que le bruit constant ou le bruit malicieux. Il inclut des bruits sur l'entrée comme sur la sortie.

– Le résultat négatif 1 dans l'état de l'art plus haut à propos des erreurs malicieuses est retrouvé sous une forme similaire.

– Les résultats positifs sont moins puissants que ceux avec erreurs malicieuses au sens où on considère un modèle de bruit plus faible. Mais les constantes sont meilleures, les facteurs logarithmiques disparaissent, l'optimalité en termes de tolérance au taux de bruit est atteinte, pour le point 4. Le point 5 a le même avantage, plus une complexité d'échantillon significativement meilleure pour une optimale robustesse au taux de bruit.

Ces résultats simples dans un cadre réaliste suggèrent que la minimisation du risque empirique pourrait être plus efficace dans beaucoup de cas pratiques que les algorithmes compliqués ci-dessous. Il est intéressant de noter quels résultats restent intéressant si nous considérons que la généralité des erreurs malicieuses ne vaut pas la perte dans les résultats. Le résultat 3, avec ses considérations pratiques (complexité en temps), dépasse la minimisation du risque empirique requise dans cette section. En outre, les algorithmes ci-dessous améliorent la complexité en temps, lorsqu'on les applique à des algorithmes PAC. On pourrait considérer que les résultats de robustesse jusqu'à une précision ϵ (alors qu'avec erreur malicieuse, ϵ est minoré par une fonction de η) sont plus forts que le bruit constant ou le bruit CPCN. Ceci est une réelle force de ce modèle dans le cadre SQ, quand le bruit est restreint à la sortie et quand les aires sont connues.

Notez que notre modèle nécessite seulement, dans cette partie, une minimisation asymptotique du risque empirique. L'algorithme décrit ci-dessous, pour les erreurs malicieuses, est une solution pour une telle minimisation.

3.1.4 Algorithme pour la résistance aux erreurs en classification - adaptation à la régression, discussion sur les méthodes à marges

L'algorithme suivant est proposé par Kearns et Li pour résister aux erreurs malicieuses avec bonne complexité d'échantillon.

On suppose donné un algorithme A, effectuant un apprentissage à complexité d'échantillon $m_A(\epsilon,\delta)$, dans le cas sans bruit. Typiquement, $m_A(\epsilon,\delta)$ évolue en $\frac{1}{\epsilon} - \log(\delta)/\epsilon$.

On éxécute r fois l'algorithme A, avec $r > 2s^2 \ln(3/\delta)$, avec un nombre d'exemples s à chaque fois égal à $m_A(\epsilon/8,\frac{1}{2})$. Tous ces échantillons sont supposés indépendants. On choisit ensuite m tel que $r \times exp(-\frac{m\epsilon}{24}) < \delta/3$, et on se donne un échantillon D de taille m.

On obtient donc r fonctions possibles que l'on peut comparer sur un échantillon de taille m. On choisit alors celle qui minimise le nombre empirique de mauvaises classifications sur D. Cette fonction, avec probabilité $1-\eta$, est juste à ϵ près. Le nombre d'exemples mis en jeu est donc, grossièrement, en $(\frac{1}{\epsilon})^3$; l'algorithme résiste à tout taux d'erreurs $< min(\epsilon/8, \log(s)/s)$.

Résumons l'algorithme ainsi défini. Etant donné un échantillon de taille M, on va le partitionner en un échantillon de taille m, et r échantillons de tailles s; le nombre s est déterminé par $m_A(\epsilon/8, \frac{1}{2})$; en gros s évolue en $\sqrt[3]{M}$. Le nombre m doit, lui, être en $\sqrt[3]{M}$, et r en $M^{(\frac{2}{3})}$. On a donc utilisé A pour sélectionner des hypothèses raisonnablement bonnes, puis l'on a sélectionné parmi celles-ci la meilleure par minimisation de l'erreur empirique sur un échantillon séparé. Il s'agit ici de l'algorithme permettant de lutter efficacement contre des erreurs *malicieuses*. Il faut noter qu'il est polynomial si l'algorithme initial l'est. Au prix d'une perte sur le taux d'erreurs malicieuses tolérable (facteur multiplicateur constant), [Decatur, 1995, p101] propose de fixer le nombre de sous-échantillons utilisés; la complexité d'échantillon devient alors linéaire en la complexité d'échantillons du cas sans bruit.

Sans cette modification, dans le cadre d'un algorithme à complexité d'échantillon en $1/\epsilon$ dans le cas sans bruit, l'algorithme initial est à éxecuter $M^{\frac{2}{3}}$ fois sur des échantillons de taille $\sqrt[3]{M}$ (à des constantes près). La complexité en temps devrait être raisonnable; l'algorithme est à éxecuter un grand nombre de fois, mais sur des échantillons inversement plus petits, donc si l'algorithme est en temps sur-linéaire (ce qui est le cas d'à peu près tous les algorithmes possibles...) on va en fait *plus vite* que l'algorithme initial. La comparaison de l'erreur en test est, elle, (souvent) linéaire. Donc cet algorithme rapproche la complexité en temps de la complexité linéaire (la complexité en temps est ici considérée par rapport au nombre d'exemple fournis en apprentissage, non par rapport au ϵ et δ souhaités - l'algorithme reste polynomial et raisonnable dans ce cadre aussi). Par contre la complexité d'échantillons n'est pas en $O((\frac{1}{\epsilon})^2)$ comme pour la plupart des algorithmes sur des espaces non trop vastes de fonctions, mais en $O((\frac{1}{\epsilon})^3)$. On voit ici l'intérêt de la modification proposée par Decatur: la complexité d'échantillon est largement meilleure.

Considérons maintenant l'algorithme proposé par Decatur pour combattre les erreurs CPCN. Il y a ici quelques déformations par rapport à la version de Decatur, car Decatur travaille dans le cadre de requêtes SQ et non du PAC-apprentissage, et nous souhaitons conserver le formalisme PAC traditionnel:

- L'algorithme suppose connues les différentes zones (et non les bruits - si l'on dispose d'évaluations du bruit, on peut sauter la première partie). Supposons que ces zones soient en nombre k.
- On suppose donné un algorithme A, permettant de PAC-apprendre la famille de fonctions, dans le cas sans bruit.
- L'algorithme A peut (presque - Kearns montre que cela est vrai pour une large classe d'algorithmes) toujours se décomposer en requêtes SQ, c'est-à-dire en un nombre fini (dépendant des données) d'évaluations de la probabilité pour qu'un exemple vérifie un certain prédicat. Cette décomposition est simplement nécessaire pour les démonstrations, mais l'explicitation de cette possibilité n'est sans doute pas utile et une adaptation intuitive est préférable pour des raisons de complexité. Formellement, on remplace un oracle fournissant des couples entrées-sorties, par un oracle renvoyant une évaluation de probabilité de réalisation, garantie avec une certaine qualité, lorsqu'on lui propose un prédicat admettant un couple entrée-sortie comme variables. La qualité est choisie par le programme, mais son inverse doit rester polynomiale.
- Un nombre polynomial de fois, avec à chaque fois un nouvel échantillon, et en supposant une famille de bruits $(\eta_i)_{i\in[1,k]}$ différente à chaque fois[3]:
- On éxécute l'algorithme A, mais à chaque fois qu'il a besoin de la probabilité empirique pour que $f(X,Y)$ soit vrai, pour un prédicat f quelconque à variables X et Y (entrée/sortie), il le fait en utilisant les données débruitées en supposant que les bruits soient environ η_i.
- On dispose donc d'une famille d'hypothèses différentes construites par l'algorithme A. Par le même lemme que dans le cas de l'algorithme de Kearns & Li, ces hypothèses ayant de bonnes raisons statistiques de n'être pas trop mauvaises, une opération de sélection par test sur un échantillon séparé fournit statistiquement un modèle performant.

Il est intéressant de constater que malgré la différence de cadre avec le cas de Kearns et Li, l'algorithme a toutefois de fortes similitudes; l'idée générale de déterminer un nombre important d'hypothèses comme si le

3. Le choix de la famille de bruits se fait simplement par énumération, en discrétisant.

bruit n'existait pas, puis choisir la bonne hypothèse par test sur un échantillon séparé. On peut se demander si dans des cas pratiques, ne pas garantir le caractère disjoint des échantillons pourrait être préférable malgré le manque de rigueur: ainsi la complexité en échantillon serait largement moindre.

La complexité de l'algorithme de Decatur en échantillon a pour remarquable avantage la précision en m en $O(1/\sqrt{m})$. Il est malheureusement à craindre que l'algorithme utilisé, passant par un débruitage algébrique éxecuté sur diverses paramètres de bruit testés les uns après les autres (exhaustivement à un pas près!) mène à des constantes décourageantes. En outre sa forme explicite passe par une forme SQ, et donc nécessite, avant d'être pratiquable, une réécriture de l'algorithme sans bruit, ou une adaptation instinctive. Il est important de noter qu'en définitive, cet algorithme effectue simplement un grand nombre de débruitages, choisit à chaque fois un modèle à partir des données débruitées, puis sélectionne sur un échantillon test le meilleur de ces modèles. On pourrait alors imaginer d'effectuer un algorithme classique de débruitage, voire des algorithmes différents, avec des paramètres différents, et tester sur un échantillon séparé l'efficacité des modèles ainsi construits pour choisir le meilleur. La théorie demande que chaque tentative de débruitage soit faite à partir d'un échantillon distinct indépendant.

Quelles sont alors les différences entre les deux algorithmes?

- L'algorithme de Decatur fournit un meilleur équivalent en m; toutefois, un algorithme réalisant un bon compromis entre r et s, comme le suggère Decatur à propos de l'algorithme de Kearns et Li (voir plus haut), permet des résultats similaires.
- L'algorithme de Kearns & Li permet de résister à un bruit malicieux avec $\eta \leq min(\epsilon/8, \ln(s)/s)$, s étant la complexité d'échantillon de l'algorithme dans le cas sans bruit associée à une précision $\epsilon/8$ et une confiance $\frac{1}{2}$.
- En pratique, Decatur effectue des débruitage (que nous préfèrerons choisir différents même si nous perdons les justifications théoriques), l'algorithme de Kearns & Li non.

L'hypothèse d'indépendance entre les échantillons utilisés pour construire les hypothèses départagées par un test final sur un échantillon test est la seule différence avec les techniques d'early-stopping ou weight-decay ou un choix des hyperparamètres d'une SVM par test sur un échantillon séparé. Les résultats pratiques seront là pour donner une idée de l'importance de cette hypothèse, et surtout de sa valeur lorsque l'échantillon en apprentissage est limité.

Les SVMs et d'autres algorithmes (rétropropagation, Lp-machines...) devraient logiquement (pour suivre du moins les recommandations du théorème de minimisation du risque structurel) minimiser la somme d'un terme de marge et d'une erreur γ-empirique. En fait, l'erreur γ-empirique est généralement remplacée par une somme de distances à l'hyperplan séparateur dans le cas de la classification (avec ϵ-insensibilité) et à la fonction de régression (avec ϵ-insensibilité aussi) dans le cas de la régression. Cela peut clairement entrainer une forte erreur dans le cas d'un bruit poivre et sel; il serait alors clairement préférable de revenir au terme d'erreur initial. Le fait que dans diverses classes précises, les algorithmes tol'erant des bruits importants sont basés sur des minimisations parfaites d'erreur empirique encourage la fidélité au terme initial d'erreur empirique. Pour cela on peut imaginer d'adapter l'algorithme des Lp-machines, en relachant les contraintes correspondant aux inégalités non satisfaites. Il est clair que l'on se ramène alors à un problème NP-complet, mais rappelons que le simplexe, au pire cas, est de toute facon déjà exponentiel.

[Mukherjee et al, 1997] souligne l'importance du choix de ϵ dans l'epsilon-insensibilité des SVMs. L'ϵ-insensibilité peut se généraliser à la rétropropagation par exemple; notons aussi que la rétropropagation permet sans problème de moduler le choix de ϵ en fonction de la sortie. Plus le bruit est fort, plus ϵ doit être choisi grand. Ceci peut se formaliser de la façon suivante: supposons que l'on ait un bruit sur la sortie majoré par une constante ϵ. Choisissons un apprentissage par ϵ-insensibilité. Alors si l'apprentissage est réussi, avec ϵ-insensibilité sur toutes les données, on peut conclure à une convergence en $O(1/m)$ du nombre de points en dehors de ce ϵ-tube. La convergence d'une régression étant en général en $O(1/\sqrt{m})$, le gain est clair; en connaissant le niveau de bruit et en l'intégrant au modèle, on garantit une réussite dans une proportion des cas croissant plus vite que dans le cas de la régression simple (qui par Markov assure une convergence en $O(1/\sqrt{m})$.

3.1.5 Extension théorique à la régression

Extension à la régression de l'algorithme de Kearns et Li

Supposons donné un échantillon $D = \{(X_1,Y_1),...,(X_m,Y_m)\}$ d'apprentissage, en régression, avec $Y \in [A,B]$. Considérons l'échantillon $D' = \{(X_1,Z_1,S_1),...,(X_m,Z_m,S_m)\}$, avec les Z_i indépendantes et uniformément distribuées sur $[A,B]$ et $S_i = 1$ si $Z > Y$ et 0 si $Z \leq Y$. Cet échantillon correspond à un problème de classification. L'erreur en classification dans le cas sans bruit fournit une borne sur l'erreur L^p dans le cadre de la régression sans bruit. Clairement, le cas d'une erreur malicieuse en régression sur D, c'est-à-dire d'une apparition indépendante identiquement distribuée des bruits et du choix par un adversaire sans limites de puissance de calcul, de connaissances (il peut tout à fait faire ses choix en fonction de l'état interne de l'algorithme d'apprentissage), on se ramène à un bruit malicieux en classification sur D'.

L'algorithme proposé pour apprendre une classe $\mathcal{F}$ de fonctions avec bruit est donc finalement:

• Construire un échantillon D' en remplacant chaque exemple (X,Y) de D par $((X,Z),C)$, Z étant choisi aléatoirement, uniformément sur $[A,B]$, C étant égal à 1 pour $Z > Y$ et 0 sinon.

• Choisir $f \in \mathcal{F}$ tel que l'ensemble $\{(x,z)/f(x) \leq y\}$ par l'algorithme de Kearns et Li.

On comprend que le caractère aléatoire de Z puisse gêner le praticien. L'algorithme proposé ci-dessous vise à remédier à ce problème.

Dérivation directe d'un algorithme d'apprentissage

Supposons que l'on cherche à effectuer une régression au moyen de modèles $f \in \mathcal{F}$. Supposer qu'il existe f permettant une prédiction à ϵ près sur toute entrée X donne une justification nouvelle des algorithmes à ϵ-insensibilité; et effectuer un apprentissage de type classification sur, avec probabilité $\frac{1}{2}$, $(X,Y+\epsilon,1)$ ou $(X,Y-\epsilon,0)$ avec l'ensemble des sous-graphes $\{\{(x,y)/y \leq f(x)\}/f \in \mathcal{F}\}$, ramène le problème de régression à un problème de classification. L'erreur L^1 en régression est inférieure ou égale à $\epsilon + |B-A|\epsilon'$, avec ϵ' le taux d'erreur obtenu en classification et A et B bornes inf et sup sur les fonctions.

L'algorithme est donc le suivant:

• Choisir ϵ tel que l'on ne cherche pas à obtenir une précision meilleure que ϵ.

• Choisir f minimisant $\epsilon' = P(|Y - f(X)| > \epsilon)$

Un tel algorithme semblera sans doute plus satisfaisant d'un point de vue praticien. Le choix de ϵ peut toutefois apparaitre contraignant. On pourra par exemple choisir $\epsilon = \theta(\sqrt{\frac{VC-dim}{m}})$ de l'ordre de la garantie sur la précision de ϵ'. Les garanties sont alors :

- Résistance au même taux d'erreurs malicieuses que l'algorithme utilisé en classification. Typiquement, linéaire en ϵ pour ϵ taux d'erreur accessible par le modèle dans le cas sans bruit.
- Complexité d'échantillon, dans le cas sans bruit, ou dans le cas avec bruit inférieur à ϵ, inversement linéaire en la précision et non quadratique.

Le lecteur attentif aura remarqué que l'algorithme que l'on propose effectue une classification sur les $(X,Y+\epsilon)$ et $(X,Y-\epsilon)$, et pourra objecter que cette transformation ne préserve pas le caractère indépendant identiquement distribué. En effet! Il faudrait en fait ne garder qu'un des deux exemples. Cela dit, intuitivement, il semble souhaitable d'ajouter ces exemples supplémentaires. Le lecteur encore plus attentif aura aussi pu remarquer qu'on peut mieux faire en supposant que l'erreur minimale dans le cas sans bruit est en fait nulle.

Minimisation de quantiles

Une solution intuitive pour effectuer un apprentissage avec bruit est de minimiser non l'erreur moyenne mais l'erreur médiane, ou l'erreur correspondant à un quantile. L'objectif de cette partie est de montrer que dans un certain cadre, cette intuition est justifiée.

Considérons donc $\mathcal{F}$, famille de fonctions à valeurs dans $[-1,1]$, et cherchons $f \in \mathcal{F}$ tel que pour un certain E, aussi petit que possible, $P(|f(X) - Y| > E) \leq q$, avec q choisi entre 0 et 1.

Soit V la VC-dim de l'ensemble $\{\chi_{\{(x,y,E)/|f(x)-y|<E\}}/f \in \mathcal{F}\}$. Il s'agit, géométriquement, de la VC-dimension de l'ensemble des tubes de largeur E autour de f. Elle est notamment finie lorsque la VC-dimension de $\mathcal{F}$ est finie. Notons $\mu(f,E) = P(|f(X) - Y| > E)$ et $\mu_m(f,E)$ son évaluation empirique à l'étape m: $\mu_m(f,E) = \frac{1}{m}\sum_{i=1}^{m} \delta_{|f(X)-Y|>E}$.

$\mu - \mu_m$ est majoré, uniformément sur $f \in \mathcal{F}$ et $E \in [0,1]$, avec probabilité $1-\delta$, par

$$\epsilon = 2\frac{V(1+\log(2m/V)) - \log(\delta/4)}{m}\left(1+\sqrt{1+\frac{\mu_m}{\frac{V(1+\log(2m/V))-\log(\delta/4)}{m}}}\right)$$

par commodité, on va dire $O(1/m + \sqrt{\frac{\mu_m}{m}})$. Si f est choisi tel que $\mu_m = q$, cela donne $O(1/m + \sqrt{q/m})$. Si q est choisi évoluant en $1/m$, cela donne $O(1/m)$. Il reste à voir quelle borne nous donne ce résultat dans le cas sans bruit.

$$P(|f(X) - Y| > E) \geq (1-\beta) \times P_{SB}(|f(X) - Y| > E) + \beta \times Proba_{\inf}(E)$$

avec β la probabilité d'apparition de bruit, $Proba_{\inf}(E)$ la probabilité minimale sur x et y pour qu'une sortie bruitée sur x soit à distance supérieure à E de y. On considèrera par la suite un bruit poivre et sel, c'est-à-dire que dans le cas d'apparition de bruit la sortie est iid suivant une loi quelconque. Cette loi fournit une borne inférieure sur $Proba_{\inf}$. L'équation précédente nous donne alors

$$P_{SB}(|f(X) - Y| > E) \leq \frac{q + O(1/m)}{1-\beta}$$

Tout ceci n'a de sens que si $E(q)$ est petit, et si l'on peut choisir f vérifiant $\mu(f,E) = P(|f(X) - Y| > E) \leq q$. Il s'agit d'effectuer un apprentissage E-insensitif, vu comme une tâche de classification, avec une borne en $1/m$. Cela est en particulier possible en VC-dimension finie.

3.1.6 Résultats pratiques

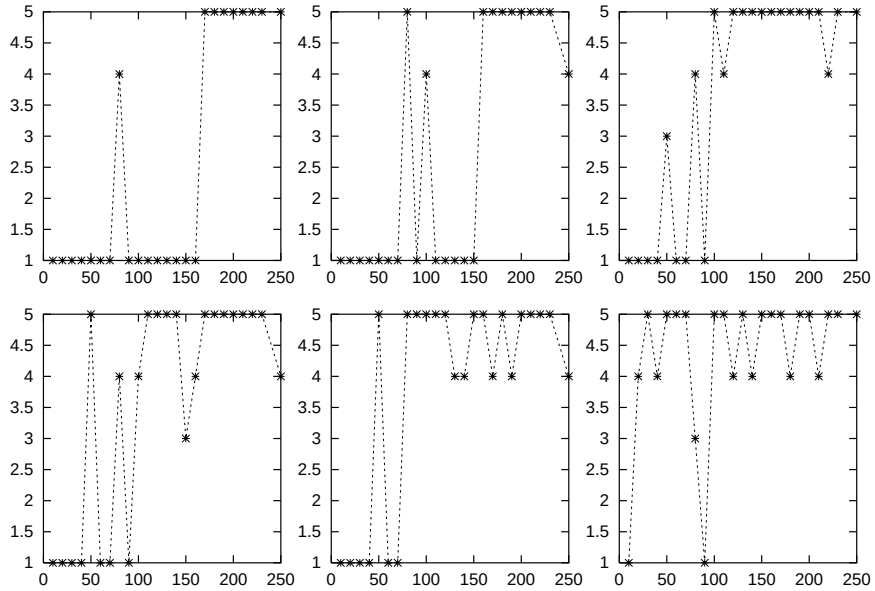

FIG. 3.1 – *Nombre N optimal de sous-ensembles pour différentes valeurs du nombre d'exemples et différentes proportions d'exemples arbitrairement bruités. Le jeu de données est improvisé pour l'occasion et les proportions d'exemples bruitées sont respectivement* 0,0.01,0.02,0.04,0.08,0.16. *Dans tous les cas lorsque le nombre d'exemples est assez grand N décolle de* 1. *Plus le taux de bruit est important, plus ce décollage se fait tôt.*

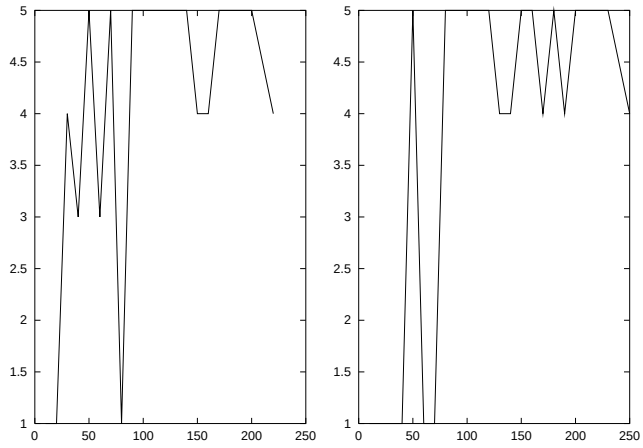

FIG. 3.2 – *à gauche, petit bruit additif et 8 % des exemples choisis au hasard, à droite, bruit additif fort: dans les deux cas l'algorithme est rapidement meilleur que l'algorithme initial. Comme on pouvait s'y attendre, l'algorithme est moins adapté lorsque le bruit est essentiellement additif. Cela dit la tendance s'inverse pour un bruit additif très fort, ce qui est logique puisqu'un tel bruit se rapproche d'un bruit du type considéré dans le modèle.*

Autres expérimentations: même sur des données peu bruitées, l'algorithme s'avère efficace:

- Sur un système chaotique à très grand nombre de points, l'algorithme s'est avéré beaucoup plus rapide que l'algorithme classique et un peu plus efficace.
- Sur les données M. l'algorithme ne fait pas mieux que l'algo classique si l'on n'ajoute pas de bruit. Il s'avère efficace si l'on bruite une proportion donnée d'exemples, et est presque aussi efficace qu'un apprentissage sur les données non bruitées. La ligne N de la figure 3.3 représente la différence entre l'erreur quadratique obtenue (divisée par la variance de la sortie sur l'échantillon d'apprentissage) par partition en $N+2$ ensembles et l'algorithme initial. L'erreur est environ 0.17 et varie peu lorsque le bruit augmente. Chaque passage en dessous de zéro signifie que l'algorithme modifié est meilleur que l'algorithme initial. En figure 3.4, on teste l'influence de l'epsilon-insensibilité sur les données M.:

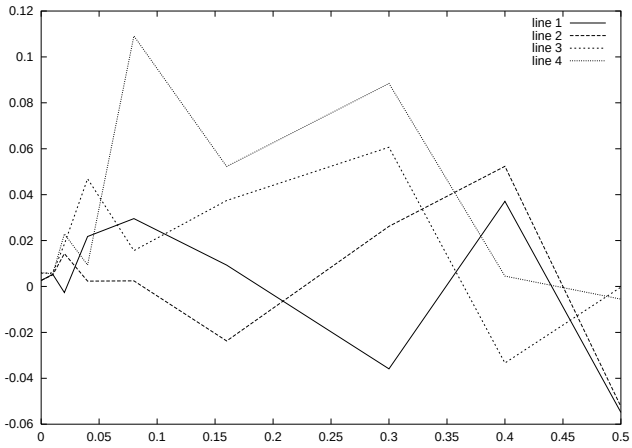

FIG. 3.3 – *Efficacité sur les données M. en fonction du bruit ajouté (ligne N = partition en $N+2$). Ca n'est pas formidable. Le nombre de points est un peu faible pour ce genre d'algorithmes basés sur des splittings.*

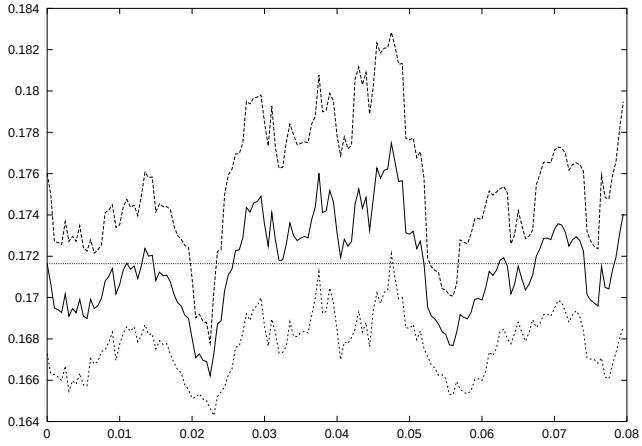

FIG. 3.4 – *Erreur en généralisation (L^2 divisée par la variance) en fonction de ϵ, pour un apprentissage avec ϵ-insensibilité avec intervalles de confiance. La ligne horizontale est le cas (presque) sans insensibilité (sans insensibilité du tout c'est moins bon pour des raisons de stabilité numérique sans doute).*

3.1.7 Conclusions et travaux futurs

La théorie de l'apprentissage avec bruit est un domaine fort amusant au niveau théorique, mais les résultats de la "learning theory with noise" comme on dit restent cantonnés à un type de bruit bien précis: une proportion donnée de points a un comportement aberrant arbitrairement mauvais.

En pratique:
Finalement les algorithmes **basés sur la décomposition en sous-ensembles (voir partie 3.1.4)** semblent adaptés pour:

- Très clairement expérimentalement les bruits du type "une proportion fixée d'exemples est absurde".
- Des bruits additifs très forts

Dans les deux cas, un nombre d'exemples important est requis; cela est cohérent avec la théorie de la VC-dimension: le nombre d'exemples par sous-base doit être grand devant la VC-dimension.

Dans les autres cas, **l'epsilon-insensibilité**, avec ϵ choisi proche de l'amplitude du bruit additif, semble plus utile. *Même si le choix de ϵ pose problème*, les résultats empiriques confirment qu'un **choix de ϵ petit** est efficace: non seulement l'**erreur est en moyenne plus petite**, mais en outre **la variabilité descend**: on a de meilleures garanties au pire cas (ce qui est peut-être plus dû à un problème de stabilité numérique plus qu'à un traitement du bruit théoriquement adapté). L'approche par quantiles semble donc sympathique car robuste aux deux types de bruit, et intuitivement très satisfaisante. En outre, à partir d'une SVM ou d'une backprop c'est vite programmé.

3.2 Généralisation sur une sous-zone

3.2.1 Introduction

On se préoccupe ici de développer des algorithmes d'apprentissage robustes à une spécialisation sur une sous-zone. Les hypothèses classiques en théorie de l'apprentissage supposent que le test sera fait sur un échantillon distribué suivant la même loi que l'échantillon d'apprentissage. Dans certains cas concrets malheureusement l'utilisateur a le mauvais goût d'avoir des priorités plus ciblées que le statisticien. Dans le cas de familles de fonctions paramétrées, le problème se ramène au choix des paramètres; par exemple, on conçoit bien que si l'on a la connaissance à priori du fait que le bon modèle est linéaire, alors des exemples abondants permettront de bien fixer les paramètres, et donc l'utilisateur n'arrivera pas à choisir une zone vraiment problématique. Les résultats formels permettant d'étayer cette intuition seront donnés, puis l'on

envisagera le cas non paramétrique (ou paramétrique ennuyeux). Si vraiment l'utilisateur est de très mauvais goût et se permet de choisir une zone arbitrairement petite, alors on choisira de minimiser l'erreur L^∞ sur une zone aussi vaste que possible. Le choix le plus raisonnable sera sans doute toutefois de supposer que l'utilisateur choisira une zone ayant certaines bonnes propriétés que l'on verra plus bas - on montrera alors qu'une bonne méthode consiste à minimiser l'erreur L^p pour p grand.

3.2.2 Apprentissage sur une distribution différente de la distribution-test

[Decatur, 1995] fournit un survey de résultats au pire cas, incluant des résultats évidents sur l'erreur sur une nouvelle distribution en fonction de l'écart entre les deux distributions, et des résultats plus difficiles prévoyant des pires cas un peu trop pessimistes sans doute pour un usage pratique. On privilégiera ici une approche qui se veut plus réaliste.

Spécialisation de l'apprentissage sur une sous-zone bornée pour L^2

On suppose A algorithme d'apprentissage sur D iid suivant P de taille m. $A(D)$ est ainsi une application de X dans Y. Afin de simplifier les notations, on notera $\tilde{f}(x') = |f(x) - y|$ pour $x' = (x,y)$. On constate ainsi que déterminer une fonction de petite erreur L^p en régression, est exactement analogue à déterminer une fonction de petite norme L^p suivant une loi jointe. L'intérêt de cette manœuvre est juste l'allègement de notations.

On qualifiera par la suite **d'admissible** une méthode donnant une borne sur le taux d'erreur en généralisation sur une nouvelle loi telle que la borne soit optimale pour certains choix de cette nouvelle loi. La notion d'admissibilité dépend donc de la classe de distributions autorisée pour les nouvelles zones.

Appelons $L_p(f)$ la racine p-ième de l'intégrale de $|f(x)|^p$. On note que l'erreur L^p, pour tout p, est majorée par $L_\infty(f)$, et ce même si on restreint f à une zone de X plus petite qu'en apprentissage. Une bonne solution semble donc, pour avoir une stabilité par restriction sur une zone plus petite, l'utilisation de normes L^∞.

On note $L_p^g(f)$ l'erreur L_p obtenue en restreignant la loi P par la loi Q de densité continue par rapport à f de densité g.

Alors $L_1^g(f)$ est majoré par $L_p(f) \times \sqrt[q]{\int_P g^q}$ avec $\frac{1}{p}+\frac{1}{q} = 1$ (théorème de Hölder). On a en particulier $L_1^g(f)$ majoré par $L_2(f)\sqrt{\int_P g^2}$. Donc minimiser l'erreur L_1 sur la loi restreinte passe ainsi par la minimisation de l'erreur L^2 sur la loi initiale. On peut de même montrer en remplacant f par f^p, que minimiser l'erreur L^p sur le nouvel ensemble est possible en considérant l'erreur L^{2p} (puissance $\frac{1}{p}$). Donc pour garantir l'erreur L^p sur une loi restreinte, une solution est de minimiser l'erreur L^{2p}.

Borne sur l'erreur
Si l'erreur L^{2p} est majorée par K, alors l'erreur L^p pour la mesure de densité g par rapport à P est majorée par $\sqrt[p]{L_2(g)} \times K$. C'est-à-dire que si g consiste en une restriction à une zone de probabilité $\mathcal{P}$, alors l'erreur L^p est majorée par $\frac{1}{\sqrt[2p]{\mathcal{P}}} \times K$.

La recherche d'une bonne erreur L^1 sur une sous-zone peut donc se faire par recherche de bonne erreur L^2. Il n'est pas question bien sûr d'affirmer que cette méthode soit la seule possible: mais dans le cadre où l'on s'est placé, cette méthode est optimale au sens où l'inégalité de Hölder est optimale: c'est-à-dire que pour une valeur donnée de $L^2(g)$, si on ne choisit pas f minimisant l'erreur L_{2p}, il y a des choix de g (g proportionnel à f), tels que l'on n'aura pas le meilleur résultat.

La méthode consistant à minimiser L^{2p} est admissible pour minimiser L^p sur une zone de densité g par rapport à P telle que $L^2(g)$ soit bornée.

On note que seul L^2 se minimise commodément d'un point de vue algorithmique, à moins d'utiliser une rétropropagation par exemple, qui se généralise facilement à tout p (donc un algorithme non borné en temps).

A $L^2(g)$ fixée, cette borne est optimale. Par contre pour obtenir une borne valable pour tout g (donc toute distribution absolument continue par rapport à P), il est nécéssaire de borner l'erreur L^∞.

Spécialisation de l'apprentissage sur une sous-zone arbitraire

On considère donc, en reprenant le problème à zéro, des x_i et des y_i tiré iid suivant une loi jointe (X,Y), et l'on cherche à ce que $P(|Y - f(X)| > M)$ soit petit.

On considère le classifieur C_f qui à x et y associe 1 si $|Y - f(X)| > M$ et -1 si $|Y - f(X)| < M$. Il s'agit bien d'une recherche de classifieur, parmi une famille de fonctions. Sa VC-dimension n'est pas la VC-dimension de l'ensemble des sous-graphes, mais la VC-dimension de l'ensemble des "tubes" autour de ces fonctions. Toutefois, les nombres de couvertures sont généralement facilement obtenus à partir des nombres de couverture usuels.

La probabilité d'erreur est alors majorée, si l'on réussit à placer toutes les données dans la classe 1, par $P_{pb} = O(h/n)$ ($h = VC - dim$ des tubes autour des fonctions...). La probabilité d'erreur après restriction à une sous-loi de densité g par rapport à P est alors majorée par $P_{pb} \times L_\infty(g)$. S'il s'agit d'une restriction à une zone de probabilité $\mathcal{P}$, alors, elle est majorée par $\frac{P_{pb}}{\mathcal{P}}$.

Borne
La probabilité d'erreur à M près sur la nouvelle distribution est majorée par la probabilité d'erreur à M près sur l'ancienne distribution, multipliée par $\frac{1}{\mathcal{P}}$

Cette borne peut probablement être fortement améliorée par l'usage de fat-shattering dimension et de marge. Une SVM semble adaptée grâce au critère d'ϵ-insensitivité.

Spécialisation de l'apprentissage sur une sous-zone: cas paramétrique

Supposons donc que l'on cherche une application m_θ de petite erreur L^p. θ est le paramètre de l'application m_θ. Il est bien clair qu'un nombre réel permet d'encoder autant de réels qu'on le souhaite, et que toute famille de fonctions peut ainsi se dire « paramétrique ». Les résultats qui suivent auront pour hypothèses des paramétrisations raisonnables (définition plus bas). On considère ci-dessous pour distance dans l'espace des paramètres la distance euclidienne.

Un théorème très général peut s'obtenir à partir de conditions d'entropie uniforme (bracketing entropie ou entropie de couverture, au choix, voir [Van der Vaart et al, 1996]), et d'une hypothèse d'existence d'un développement de Taylor à l'ordre 2, et quelques conditions raisonnables. Pour la plupart des cas concrets on pourra se contenter des hypothèses suivantes, souvent vérifiées :

- $\theta \mapsto m_\theta(x)$ est Lipschitzien pour tout x. Le coefficient de Lipschitz n'a pas besoin d'être constant pour x variable; il suffit qu'il soit L^2.
- L'espérance pour la loi des exemples de m_θ admet un développement de Taylor à l'ordre 2.

Alors si l'on choisit le paramètre comme minimisant l'erreur L_1 empirique, il converge en $O(1/\sqrt{m})$. On ne parle ici que d'erreur L^1, mais la généralisation à L^p est immédiate. Cela est détaillé après le résultat suivant destiné à alléger les hypothèses (tout en restant plus concret que le cas très général évoqué plus haut):

- $m_\theta(x) - m_{\theta'}(x)$ est majoré (en valeur absolue) par $k(x) \times d(\theta,\theta')^\alpha$ pour un certain $\alpha > 0$. $k(x)$ est supposé L^2.
- L'espérance pour la loi des exemples de m_θ admet un développement de Taylor à l'ordre 2.

Alors si l'on choisit le paramètre comme minimisant l'erreur L^1 empirique, il converge en $O(1/\sqrt[4-2\alpha]{m})$. Si l'on choisit le paramètre comme minimisant l'erreur L^p, $k(x) \sup_\theta |m_\theta(x)|$ étant supposé L^2, et la convergence est toujours en $O(1/\sqrt[4-2\alpha]{m})$.

La convergence dans l'espace des paramètres a l'avantage d'être indépendante de la zone choisie. Plus précisément, s'il existe un θ_0 optimal, minimisant l'erreur L^p en tout point, alors l'erreur L^p en généralisation, lorsque l'on minimise l'erreur L^p empirique, converge en $O(1/m^{\alpha/(4-2\alpha)})$. Pour $\alpha = 1$, le plus vraisemblable, on constate donc une convergence bien classique en $O(1/\sqrt{m})$.

3.2.3 Résultats pratiques

Sur un benchmark de dimension 2 en régression, avec une application sous-jacente sinusoidale, et pour une rétropropagation à 7 neurones cachés, on apprend puis l'on teste sur un grand nombre de sous-zones bornées pour L^2 de même probabilité L^2 pour la probabilité initiale. On minimise l'erreur L^p en apprentissage, et on teste l'erreur L^2 en généralisation sur toutes ces sous-zones; on considère la pire erreur obtenue.

p	2	4	6	8
Erreur en généralisation pour 650 exemples	0.93	0.87	0.89	
pour 50 exemples	1.07	0.90	0.90	0.91

De multiples essais confirment ce résultat: **En pratique:** La stabilité de l'apprentissage est grandement améliorée par une utilisation de plus grandes valeurs de p.

3.2.4 Conclusion

On peut dire en conclusion que dans le cas paramétrique, la généralisation à une distribution différente n'est pas un problème. Cela suppose une compréhension physique (ou empirique) suffisante des phénomènes en jeu pour proposer une modélisation simple. Par contre, le cas général pose problème: l'argument est finalement qu'il est nécéssaire de minimiser l'erreur L^p, pour p grand; les SVMs apparaissent donc déconseillées puisqu'elles traitent des apprentissages pour L^1. Les méthodes linéaires (généralisées, donc incluant des méthodes à noyaux) par contre semblent plus sympathiques puisqu'elles minimisent l'erreur L^2. L'erreur L^p pour p plus grand apparait difficile à manier; le seul algorithme qui ne semble pas gêné par ce choix est la rétropropagation, qui fonctionne presque de même pour tout p.

En conclusion pratique, pour avoir une bonne garantie lorsque l'on travaille sur une sous-zone, on a intérêt à minimiser l'erreur L^p pour p grand. Si l'on veut garantir un minimum vital même dans de très mauvais cas, alors on en vient à minimiser des quantiles, ce qui est aussi montré par ailleurs une bonne solution pour résister au bruit. Tout cela justifie doublement l'ϵ-insensibilité.

Dernière remarque enfin, à propos de choses qui marchent mal: utiliser une SVM pour sélectionner une sous-base représentative d'exemples, avec l'espoir qu'apprendre sur cette sous-base évite de se spécialiser trop sur les zones où la densité d'exemples est forte. De manière générale, les SVMs peuvent certes permettre d'extraire des sous-bases (les support vectors), mais dans nos expérimentations ces sous-bases ne sont pertinentes QUE pour un apprentissage par SVM. Mieux vaut, dans d'autres cas, utiliser un algorithme classique genre K-means ou algorithme de transfert.

Bibliographie

[Aslam et al, 1996] J.-A. Aslam, S.-E. Decatur, *On the Sample Complexity of Noise-Tolerant Learning*, information processing letters, 1996.

[Decatur, 1997] S.-E. Decatur, *PAC Learning with Constant-Partition Classification Noise and Applications to Decision Tree Induction*, Proceedings of the Fourteenth International Conference on Machine Learning, 1997.

[Decatur, 1995] S.-E. Decatur, *Efficient Learning from Faulty Data*, Thesis, 1995.

[Kearns et al, 1993] M. Kearns, M. Li, *Learning with malicious errors SIAM Journal on Computing, 22, 807–837, 1993.*

[Gentile et al, 1998] C. Gentile, D.-P. Hambold, *Improved Lower Bounds for Learning from Noisy Examples: an Information-Theoretic Approach, Proceedings of the 11th Annual Conference on Computational Learning Theory, ACM Press, 1998.*

[Mukherjee et al, 1997] S. Mukherjee, E. Osuna, F. Girosi, *Non-linear Prediction of Chaotic Time Series Using Support Vector Machines, Proc. of IEEE NNSP'97, Amelia Island, FL, 1997*

[Teytaud, 2001] O. Teytaud, *Learning with noise. Extension to regression. Proceedings of Ijcnn 2001.*

[Vapnik, 1982] V. Vapnik, *Estimation of Dependencies Based on empirical data. Springer-Verlag, New York, 1982*

[Van der Vaart et al, 1996] A.-W. van der Vaart, J.-A. Wellner, *Weak convergence and Empirical Processes, Springer, 1996*

Chapitre 4

Vitesse de Convergence des modèles déformables et du recalage d'images

Résumé

Nous étudions les modèles déformables (eg, les "snakes") et le recalage d'images (eg, pour la radiothérapie) dans le cadre de la théorie de l'apprentissage.

Tout d'abord, rappelons le problème du recalage d'image. Soit I une image et J une image à "recaler" sur I, ie sur laquelle on applique $g \in \Gamma$ tel que $g(J) \simeq I$, la dissimilarité étant encodée par $L(g(J),I)$. Pour des raisons physiques, on suppose que certains $g \in \Gamma$ sont moins vraisemblables que d'autres; ceci est codé par une régularisation; ainsi, on minimise $L(g(J),I) + R(g)$. Ceci restreint g à une sous-famille $\Gamma' \subset \Gamma$. Usuellement, g est atteint par une descente de gradient. Dans la suite et de manière à avoir des notations homogènes, on va noter F et L (à nouveau) tels que $L(f,I) = L(g(J),I)$, sans perte de généralité ($F = \{g(J)/g \in \Gamma\}$).

Considérons ensuite les modèles déformables. Considérons un ensemble F de modèles possibles (disons, les positions possibles d'un corps (eg une tumeur) dans une image) et une image I, et cherchons $f \in F$ tel que f est probablement proche de la position du corps représenté par I.

Les modèles déformables sont typiquement (en 2D) codés par des listes de points. Usuellement, un point initial est choisi (automatiquement, comme dans [Lai et al, 1993], ou manuellement par l'utilisateur). Alors une descente de gradient est utilisée, basée sur la similarité empirique entre l'image I et le modèle f, plus un terme de régularisation.

Ce principe est basé sur deux hypothèses:

1. La descente de gradient trouve approximativement un point empirique optimal.
2. Les valeurs emiriques sont liées aux "réelles", les "réelles" étant celles résultant de grilles de précision infinie sans bruit.

Le premier point dépend de résultats de convexité.

Le second point est très similaire aux problèmes traités dans [Devroye et al, 1996, Vidyasagar, 1997, Van der Vaart et al, 1996]. Malheureusement, les résultats basés sur la VC-théorie ne s'appliquent pas ici, parce que la VC-dimension est usuellement infini pour les cas courants de modèles déformables. En tout cas, [Vidyasagar, 1997, chap. 6] fournit des bornes non-asymptotiques, dans le cadre "probablement approximativement correct" (PAC) de Valiant, dans le cas d'un apprentissage avec connaissance a priori sur la distribution. De telles bornes peuvent être utilisées ici. En outre, des convergences asymptotiques (plus rapides que les convergences résultant de bornes asymptotiques) peuvent être prouvées. De tels résultats peuvent être trouvés dans [Van der Vaart et al, 1996] et sont rappelés plus bas.

Dans ce papier nous fournissons:

- Des conditions sous lesquelles une convergence faible en $O(\frac{1}{\sqrt{n}})$ a lieu, selon la dimension, la famille de modèles déformables, avec n le nombre de pixels.
- Des bornes non-asymptotiques, dans l'esprit du modèle d'apprentissage de Valiant, sur la différence entre:
  - le L *empirique* du f obtenu et le L *réel* de ce f

- le L *réel* du f obtenu et le L *réel* du meilleur f possible sous réserve que l'on ait minimisé le L empirique sur $f \in F(\epsilon)$ avec $F(\epsilon)$ une discrétisation de F (des extensions prenant en compte le compromis avec R sont possibles)

Ce travail a été (très partiellement) publié ([Teytaud et al, 2001]).

Table des matières

4.1 Introduction

Ce travail est à l'intersection des modèles déformables (ou du recalage d'image, cependant nous présentons essentiellement les modèles déformables, le recalage d'images étant mathématiquement (au moins du point de vue statistique) similaire) et du processus empirique (en incluant la théorie de l'apprentisage). Une introduction aux modèles déformables peut être trouvée dans [McInerney et al, 1996], alors que des introductions à l'apprentissage asymptotique et non-asymptotique peut être trouvé dans [Van der Vaart et al, 1996] et dans [Vidyasagar, 1997, Devroye et al, 1996] respectivement.

- X = **domaine**: [0,1]^d
- Y = **sortie**: [0,1]
- I = **image**:application de $X \to Y$.
- F = famille de **modèles**: applications $X \times Y \to$ [0,1].
- $\tilde{F} = \{x \mapsto \tilde{f}(x) = f(x,I(x))/f \in F\}$.
- Quand pour tout $f \in F$, il existe f' tel que $\forall (x,y) \in X \times Y$ $f(x,y) = |f'(x) - y|$ a lieu, alors on définit F' l'ensemble des tels f'. Notons que $\tilde{f}(x) = |f'(x) - I(x)|$.

Objectif: Trouver $f \in F$ tel que $Val(f) = \int_X \tilde{f} dx$ soit petit. Par la suite, on notera $Val_n(f) = \frac{1}{n}\sum_{i=1}^n \tilde{f}(X_i)$. Les X_i sont précisés plus tard.

4.2 Arrière plan mathématique

4.2.1 Résultats statistiques fondamentaux

Les nombres de couverture et la VC-dimension, dont les définitions sont rappelées ci-dessous, sont bien connus en théorie de l'apprentissage statistique, où ils ont fourni beaucoup de bornes non-asymptotiques. [Vidyasagar, 1997] fournit un survol clair de tels résultats.

Une suite de variables aléatoires X_i converge **faiblement** vers X si pour tout f continu et borné $E(f(X_n)) \to E(f(X))$. Une suite de variables aléatoires X_n est $O(u_n)$ **faiblement** si $\forall \epsilon \exists K /$ lim inf $P(|X_n| \leq Ku_n) \geq 1 - \epsilon$. Un **processus stochastique** indexé par une famille T est une famille de variables aléatoires à valeurs dans $\mathbb{R}$ indexées par T et dépendant d'un même univers.

$\mathcal{F}_{B,k,d}$, avec $k > 0$ réel, est la famille des ensembles $\{(x,t)/f(x) < t\}$ pour f de [0,1]$^{d-1}$ vers [0,1] tels que $|||f|||_k$ est bien défini et borné par B, avec

$$|||f|||_k = \max_{\sum k_i \leq \lfloor \alpha \rfloor} \sup_x |\frac{\partial^{\sum k_i} f(x)}{\partial x_1^{k_1} \partial x_2^{k_2} \dots \partial x_d^{k_d}}| \tag{4.1}$$

$$+ \max_{\sum k_i = \lfloor \alpha \rfloor} \sup_{x \neq y} \frac{|\frac{\partial^{\sum k_i} f(x)}{\partial x_1^{k_1} \partial x_2^{k_2} \dots \partial x_d^{k_d}} - \frac{\partial^{\sum k_i} f(y)}{\partial x_1^{k_1} \partial x_2^{k_2} \dots \partial x_d^{k_d}}|}{|x - y|^{\alpha - \lfloor \alpha \rfloor}} \tag{4.2}$$

avec $\lfloor\alpha\rfloor$ le plus grand entier strictement plus petit que α (NB: égal à $\alpha-1$ si α est entier!). Cet espace de fonctions est appelé un espace de **Hölder**.

La notation $E(X\{F\})$ avec F un évènement et X une variable aléatoire note l'espérance de $X\times\chi_F$, avec χ_F la fonction caractéristique de F égale à 1 si F a lieu et 0 sinon. Ceci sera utile pour noter la condition de Lindeberg.

On utilise par la suite la notion de **masse de Dirac régulée**: $s\delta_{x,h}$ est une variable aléatoire uniformément distribuée sur un hypercarré d'arête h centrée sur x. $x_{i,n}$ note le i^e point sur une grille de n points de $[0,1]^d$. $x_{i,n}$ est défini par la suite seulement pour n une puissance d^e d'un entier (des généralisations à des grilles non régulières sont possibles).

$Z'_{n,i}$ est une variable aléatoire de densité $p\times s\delta_{x_{i,n},\frac{h}{sqrt[d]{n}}}+(1-p)s\delta_{-x_{i,n},\frac{h}{sqrt[d]{n}}}$. Ainsi, $Z'_{n,i}$ est une variable alétoire localisée environ en $x_{i,n}$ avec probabilité p, et environ en $-x_{i,n}$ avec probabilité $1-p$. Le "environ" est là pour la précision finie de l'imagerie médicale par exemple. Le choix $h=1$ est le cas dans lequel la précision est telle que la valeur d'un pixel est choisie au hasard dans des voxels distincts, et soumise à du bruit. $h<1$ amène à des effets de grille: la convergence marginale n'est pas garantie, comme la précision (côté d'une grille) converge vers 0. Ceci amène une convergence en $O(1/\sqrt[d]{n})$ au lieu de $O(1/\sqrt[d]{n})$. $h>1$ amène à des effets de grille similaires, et $O(1/\sqrt[d]{n})$ au lieu de $O(1/\sqrt[d]{n})$. D'autres bruits pourraient être considérés de même; par exemple, on pourrait considérer un moyennage sur des pixels ou des pixels élargis, soumis au bruit. Cette extension est laborieuse et sera négligée pour préserver la clarté. On veut maintenant définir $Z_{n,i}$, tel que $\sqrt{n}Z_{n,i}(f)$ soit égal à l'évaluation empirique du snake f au point $x_{n,i}$. Ainsi on doit prendre I en compte. Ceci amène à $Z''_{n,i}=Z'_{n,i}\times I(|Z'_{n,i}|)$, et $Z_{n,i}$ avec densité proportionnelle à la densité de $Z''_{n,i}$, et mesure globale $1/\sqrt{n}$ (on utilise $|(x_1,...,x_d)|=(|x_1|,|x_2|,...,|x_d|)$).

Avec $\mathcal{F}$ une famille de fonctions, et X une variable aléatoire, $\|X\|_{\mathcal{F}}$ dénotant $\sup_{f\in F}|E(f(X))|$. Ceci peut être étendu à des mesures autres que de probabilité. $N_{[\,]}(\epsilon,\mathcal{F})$, appelé bracketing nombre de couverture $\mathcal{F}$, est le cardinal de la plus petite famille (si elle est finie) de $(f_i,g_i)_{i\in I}$ tels que $\forall f\in\mathcal{F}\exists i\in I/f_i\le f\le g_i$ et $\int g_i-f_i\le\epsilon^2$.

Les bracketing nombres de couvertures ont été étudiées dans [Van de Geer, 1991, Birman et al, 1967].

Etant donné $Z_{n,i}$ pour $1\le i\le n$, $N'_{[\,]}(\epsilon,\mathcal{F},n)$ est le cardinal (dépendant de n) de la plus petite famille (si elle est finie) de $(F_i)_{i\in I}$ telle que $\forall f\in\mathcal{F}\exists i\in I/f\in F_i$ et $\forall i/\sum_{i=1}^n E(\sup_{f,g\in F_i^2}(Z_{n,i}(f)-Z_{n,i}(g))^2)\le\epsilon^2$. p est la qualité de l'image dans le cas d'un bruit aléatoire (d'autres formes de bruit sont discutés plus bas; principalement, on est intéressé par la convergence des moyennes empiriques vers leurs espérances, dont pourvu que le bruit préserve localement l'ordre de l'espérance de la fonction de coût, il peut être traité par les résultats qui suivent); $\frac{1}{2}<p\le 1$. $p=1$ est le cas (inintéressant) idéal, $p\to\frac{1}{2}$ abaisse la qualité. Les $Z_{n,i}$ pour $i\in[1,n]$ seront considérés indépendants, ce qui nous apparait être l'hypothèse la plus problématique.

Pour f et g dans $\{0,1\}^{[0,1]^d}$, $d(f,g)=\frac{1}{2}\int 1-fg$.

Etant donnés un modèle déformable f, on définit sa **valeur**: $Val(f)=(2p-1)\int_{[0,1]^d}f(x)I(x)$. Son évaluation empirique sera appelée **valeur empirique**: $Val_n(f)=\frac{1}{\sqrt{n}}\sum f(Z_{n,i})$. Ceci est une variable aléatoire.

Définition 4.1 (Réseaux, nombres de couverture & bracketing nombres de couverture) *Etant donné un espace F de fonctions, un* **ϵ-réseau** *ou* **ϵ-couverture** *de F est un ensemble infini de fonctions $\{f_1,f_2,\ldots,f_N\}$ (non nécessairement inclus dans F) tel que pour tout $f\in F$, il existe i tel que $\int|f-f_i|\le\epsilon$.*

Quand N est minimal, il est appelé le **nombre de couverture** *$N(F,\epsilon,d)$ de F.*

De même, pour F un espace de fonctions, un **ϵ-bracket** *$[l,u]$ avec l et u tels que $\int u-l\le\epsilon$ est l'ensemble des $f\in F$ tels que $l\le f\le u$. L'*ϵ**-bracketing nomber de couverture de** *F est le nombre minimal $N_{[\,]}(F,\epsilon,d)$ d'ϵ-brackets $[l_i,u_i]$ tels que $F\subset\cup[l_i,u_i]$.*

*L'***entropie** *(resp.* **bracketing-entropie***) est le logarithme des nombres de couverture (resp. bracketing nombres de couverture).*

Les nombres d'entropie ont été utilisées en processus empirique dans [Dudley, 1967] (où Dudley attribue l'idée à Strassen) et dans [Sudakov, 1969, Strassen et al, 1969].

Définition 4.2 (VC-dimension) *Considérons C une classe de fonctions à valeurs dans $\{0,1\}$. Définissons $\Delta_n=\sup_{(x_1,\ldots,x_n)\in X^n}card\{c^{-1}(1)\cap\{x_1,\ldots,x_n\}|c\in C\}$ le n^e coefficient de pulvérisation de C, et la* VC-dimension *de C est $VC(C)=\sup_n\{m|D_m=2^m\}$.*

On aura besoin par la suite d'autres formes de (bracketing) nombres de couvertures. [Van der Vaart et al, 1996] fournit des résultats basés sur de tels nombres, donnant pour références [Koul, 1970, Shorack, 1973, Shorack, 1980, Van Zujilen, 1978, Shorack et al, 1986, Marcus et al, 1984, Shorack et al, 1986].

Définition 4.3 ("Entropie spéciale") *Etant donnés $Z_{n,i}$ variables aléatoires pour $1 \leq i \leq n$, $N'_{[\ ]}(\epsilon,\mathcal{F},n)$ est le cardinal (dépendant de n) de la famille la plus petite (si elle est finie) de $(F_i)_{i\in I}$ tels que $\forall f \in \mathcal{F} \exists i \in I / f \in F_i$ et $\forall i / \sum_{i=1}^n E(\sup_{f,g\in F_i^2}(Z_{n,i}(f) - Z_{n,i}(g))^2) \leq \epsilon^2$.*

Théorème 4.4 (Borne de Chernoff) *Considérons X une variable binomiale somme de n variables de Bernoulli indépendants de paramètre p (égales à 1 avec probabilité p et 0 sinon).*

Alors pour tout $\Gamma \in [0,1]$,

$$P(X/n \geq (1+\Gamma)p) \leq \exp(-\Gamma^2 np/3)$$
$$P(X/n \leq (1-\Gamma)p) \leq \exp(-\Gamma^2 np/2)$$

L'inégalité de Hoeffding, prouvée par Hoeffding en 1963 [Hoeffding, 1963], généralise les bornes de Chernoff sous forme multiplicative. Elle est bien connue en théorie de l'apprentissage. Notez qu'elle n'implique par la forme multiplicative des bornes de Chernoff.

Théorème 4.5 (Inégalité de Hoeffding) *Soit $X_1,\ldots,X_n$ des variables aléatoires indépendantes respectivement à valeurs dans $[a_1,b_1],\ldots,[a_n,b_n]$.*

$$P(\frac{1}{n}\sum(X_i - EX_i) \geq \epsilon) \leq \exp(-\frac{2n\epsilon^2}{\frac{1}{n}\sum|a_i - b_i|^2}) \tag{4.3}$$
$$P(\frac{1}{n}\sum(X_i - EX_i) \leq -\epsilon) \leq \exp(-\frac{2n\epsilon^2}{\frac{1}{n}\sum|a_i - b_i|^2})$$
$$P(\frac{1}{n}\sum|X_i - EX_i| \geq \epsilon) \leq 2\exp(-\frac{2n\epsilon^2}{\frac{1}{n}\sum|a_i - b_i|^2}) \tag{4.4}$$

Les théorèmes centraux limites généralisés (avec entropie uniforme) sont dus à [Dudley, 1978, Pollard, 1982, Kolcinski, 1981], et leurs équivalents bracketing sont dus à [Dudley, 1978, Dudley, 1984, Ossiander, 1987, Andersen et al, 1988]. Un survol général de tels résultats peut être trouvés dans [Van der Vaart et al, 1996]

Théorème 4.6 (Convergence rapide) *Pour tout $n \in \mathbb{N}$, on considère $Z_{n,i}$, pour $i \in [1,n]$, des variables aléatoires indépendantes. On suppose que les hypothèses suivantes sont vérifiées:*

$$\sum_{i=1}^n E \parallel Z_{n,i} \parallel_{\mathcal{F}}^2 \{\parallel Z_{n,i} \parallel_{\mathcal{F}} > \eta\} \to 0 \text{ for all } \eta > 0 \tag{4.5}$$

$$\sup_{(f,g)\in\mathcal{F},d(f,g)<\delta_n} \sum_{i=1}^n E(Z_{n,i}(f) - Z_{n,i}(g))^2 \to 0$$
$$\text{pour tout } \delta_n \text{ décroissant vers } 0 \tag{4.6}$$

$$\int_0^{\delta_n} \sqrt{\log N'_{[\ ]}(\epsilon,\mathcal{F},n)}d\epsilon \to 0$$
$$\text{for all } \delta_n \text{ décroissant vers } 0 \tag{4.7}$$

Alors $E_n = \sum_{i=1}^n Z_{n,i} - E(Z_{n,i})$ est asymptotiquement tight dans l'espace des fonctions totalement bornées de $\mathcal{F}$ dans $\mathbb{R}$. Ceci signifie que pour tout $\epsilon > 0$ il existe un ensemble compact K tel que $\liminf P(E_n \in K^\delta) \geq 1 - \epsilon$ pour tout $\delta > 0$, avec $K^\delta = \{y / d(y,K) < \delta\}$ le δ-élargissement de K.

Il converge en distribution pourvu que la suite des fonctions de covariance converge point à point sur $\mathcal{F}^2$.

La condition 4.6 peut être supprimée si la partition dans $N'_{[\ ]}$ peut être choisie indépendamment de n.

La finitude des bracketing nombres de couverture sera suffisante pour un théorème plus faible:

Théorème 4.7 (Convergence faible) *On suppose que les $N_{[\,]}(\epsilon,\mathcal{F})$ sont finis pour tout ϵ, et que tous les f_i et g_i peuvent être choisis dans $\mathcal{F}'$, tels que la convergence presque sûre de $Val_n(f)$ vers $Val(f)$ a lieu pour tout $f \in \mathcal{F}'$. Alors $Val_n(f)$ converge vers $Val(f)$, presque sûrement, uniformément en $f \in \mathcal{F}$.*

Preuve:

• Considérons $\epsilon > 0$.

• Soit I, f_i et g_i choisis de manière à réaliser $N_{[\,]}(\epsilon,\mathcal{F})$.

• Pour $i \in I$, avec probabilité 1, il existe n_i tel que $n > n_i$ implique que $|Val_n(f_i) - Val(f_i)| \leq \epsilon$ et $|Val_n(g_i) - Val(g_i)| < \epsilon$. Avec $n > max_i n_i$, ceci est uniforme en $i \in I$.

• Les affirmations suivants ont lieu pour tout n:

$$\begin{aligned} Val_n(f) - Val(f) &\leq (Val_n(f) - Val_n(g_i)) \\ &\quad +(Val_n(g_i) - Val(g_i)) \\ &\quad +(Val(g_i) - Val(f_i)) \\ &\quad +(Val(f_i) - Val(f)) \\ &\leq 0 + \epsilon + \epsilon + 0 = 2\epsilon \end{aligned}$$

et

$$\begin{aligned} Val(f) - Val_n(f) &\leq (Val(f) - Val(g_i)) \\ &\quad +(Val(g_i) - Val(f_i)) \\ &\quad +(Val(f_i) - Val_n(f_i)) \\ &\quad +(Val_n(f_i) - Val_n(f)) \\ &\leq 0 + \epsilon + \epsilon + 0 \end{aligned}$$

• Donc, $|Val_n(f) - Val(f)| \leq 2\epsilon$. □

Le résultat suivant est adapté de [Vidyasagar, 1997, p188].

Théorème 4.8 (Bornes non-asymptotiques) *Soit $F_{\epsilon_0/2}$ une $\frac{1}{2}\epsilon_0$-couverture de F'. Alors, l'algorithme de minimisation du risque empirique appliqué à $F_{\epsilon_0/2}$ est PAC (probablement approximativement correct) avec précision $\epsilon > \epsilon_0$ et confiance $1-\delta$ dans F, pourvu que n (nombre d'exemples) est choisi plus grand que $\frac{8}{\epsilon^2}\ln(\frac{Card\ F_\epsilon}{\delta})$. En outre, avec confiance au moins $1 - Card\ F_{\epsilon_0/2} exp(-n\epsilon^2)$, la différence entre l'erreur en généralisation et l'erreur empirique, pour tout algorithme choisissant son hypothèse dans $F_{\epsilon_0/2}$, est bornée par $e_0/2$.*

4.2.2 Nombres de couverture

Les lemmes qui suivent seront utiles pour les applications. Pour f et g dans $\{0,1\}^{[0,1]^d}$, définissons $d(f,g) = \frac{1}{2}\int 1 - fg$.

Lemme 4.9 *Pour tout δ_n décroissant vers 0,*

$$\sup_{d(f,g)<\delta_n} \sum_{i=1}^{n} E(f(Z_{n,i}) - g(Z_{n,i}))^2 \to 0$$

Preuve:

Soit A égal à $\sum_{i=1}^{n} E(f(Z_{n,i}) - g(Z_{n,i}))^2$, et h l'arête élémentaire de la grille ($h = \frac{1}{\sqrt[d]{n}}$).

$$\begin{aligned} A &\leq 4E_{\cup s\delta_{i,n}}(\chi_{\{f\neq g\}}) \\ &\leq 4(\frac{h}{h'})^d d(f,g) \\ &\leq 4(\frac{h}{h'})^d \delta_n \end{aligned}$$

Ceci fournit le résultat attendu.□

Ce lemme 9 sera utile pour vérifier la condition 4.6.

Lemme 4.10 *Avec les notations précédentes,*

$$N'_{[\ ]}(\epsilon,\mathcal{F},n) \leq N_{[\ ]}(\epsilon',\mathcal{F})$$

avec $\epsilon' = \epsilon \times (\frac{h'}{h})^{\frac{d}{2}}$.

Preuve:

Considérons I, f_i et g_i réalisant $N_{[\ ]}(\epsilon',\mathcal{F})$. Ainsi $d(f_i,g_i) \leq \epsilon'^2$.

Soit F_i l'ensemble des $f \in \mathcal{F}$ tels que $f_i \leq f \leq g_i$. Pour tout j,

$$\begin{aligned}
&\sum_{i=1}^{n} E \sup_{f,g\in F_j} |f(Z_{n,i}) - g(Z_{n,i})|^2 \\
&\qquad \leq \frac{1}{n}\sum_{i=1}^{n} E_{s\delta_{i,n}} \sup_{f,g\in F_j} \chi_{f\neq g} \\
&\qquad\qquad \leq \frac{1}{n}\sum_{i=1}^{n} E_{s\delta_{i,n}} \chi_{f_i\neq g_i} \\
&\qquad\qquad\qquad \leq (\frac{h}{h'})^d \epsilon'^2
\end{aligned}$$

Ainsi la preuve est complète.□

Le lemme 17 sera utile pour vérifier la condition 4.7.

Lemme 4.11 (Nombres de couverture de $\tilde{F}$) *Supposons que pour tout $f \in F$ $f(x,y) = |f'(x) - y|$. Alors*

$$N(\tilde{F},\epsilon,L_1) \leq N(F',\epsilon,L_1)$$
$$N_{[\ ]}(\tilde{F},\epsilon,L_1) \leq N_{[\ ]}(F',\epsilon,L_1) \text{ if } Y \text{ is restricted to } \{0,1\}$$

Proof:Le premier résultat provient de

$$\forall f,g \in F, d(f',g') < \epsilon \Longrightarrow d(\tilde{f},\tilde{g}) < \epsilon$$

où $d(a,b) = \int |a(.) - b(.)|$. Le second provient de

$$a' \leq f' \leq b' \Longrightarrow a'' \leq \tilde{f} \leq \tilde{b}''$$

où $a''(x) = \tilde{a}(x)$ si $I(x) = 0$ et $a''(x) = \tilde{b}(x)$ si $I(x) = 1$, et $b''(x) = \tilde{b}(x)$ si $I(x) = 0$ et $b''(x) = \tilde{a}(x)$ si $I(x) = 1$.□

Ceci montre que les nombres de couverture de F' peuvent être utilisés au lieu des nombres de couverture de $\tilde{F}$.

Un résultat très important est le suivant, fournissant des bornes sur les nombres de couverture dans le cas de modèles réguliers. Les espaces de fonctions régulière ont été étudiés dans [Kolmogorov et al, 1961, Lorentz, 1966, Birman et al, 1967, Dudley, 1984].

Théorème 4.12 (Adapté de Kolmogorov-Thikomirov ([Kolmogorov et al, 1961])) *Soit F une famille de fonctions de Hölder de coefficient α, β le plus grand entier strictement plus petit que α, M borne sur les dérivées jusqu'à α. Alors, les ϵ-nombres de couverture de F pour la norme infinie sont bornés de la manière suivante:*

$$N(\epsilon,F,\|\ .\ \|_\infty) \leq (L+1)^{m-1} \times (\lfloor 2M/\epsilon \rfloor)^{\frac{(d+1)^\beta}{\beta!}}$$

Proof:On suit les lignes de [Van der Vaart et al, 1996, chap 2.7] (et [Kolmogorov et al, 1961]), avec, en outre, les constantes requises pour avoir des bornes explicites.

On définit $k. = \sum k_i$ et $D_k = (\frac{\partial}{\partial x_1})^{k_1} \times \ldots (\frac{\partial}{\partial x_d})^{k_d}$.

Définissons $\delta = \epsilon^{1/\alpha}$. Considérons un δ-réseau de X, de points $x_1,\ldots,x_m \in X$. Alors,

$$m \leq \lceil \frac{\sqrt{d}}{2\delta} \rceil^d \tag{4.8}$$

Pour tout vecteur $k = (k_1,\ldots,k_d) \leq \beta$, et pour tout $f \in F$, on note

$$A_k f = (\lfloor \frac{D^k(x_1)}{\delta^{\alpha-k.}} \rfloor, \ldots, \lfloor \frac{D^k f(x_m)}{\delta^{\alpha-k.}} \rfloor)$$

Le vecteur $\delta^{\alpha-k} A_k f$ consiste en les valeurs $D^k f(x_i)$ discrétisées sur une grille d'arête $\delta^{\alpha-k}$.

Considérons deux fonctions f et g avec les mêmes valeurs discrétisées $\delta^{\alpha-k} A_k f$. Grâce à la formule de Taylor

$$f(x) - g(x) \leq \sum_{k.\leq\beta} D^k(f-g)(x_i) \frac{(x-x_i)^k}{k!} + R$$

où R est borné par $\frac{M}{\alpha!} \parallel x - x_i \parallel^\alpha$.

Comme il y a toujours, pour $x \in X$, un x_i tel que $\parallel x - x_i \parallel \leq \delta$, Ceci implique que (avec la notation $h^k/k! = \Pi_{i=1}^d (h_i^{k_i}/k_i!)$)

$$\begin{aligned} \parallel f - g \parallel_\infty &\leq \sum_{k.\leq\beta} \delta^{\alpha-k} \frac{\delta^{k.}}{k!} + \frac{M}{\alpha!}\delta^\alpha \\ \parallel f - g \parallel_\infty &\leq \delta^\alpha (e^d + \frac{M}{\alpha!}) \\ \parallel f - g \parallel_\infty &\leq C\epsilon \end{aligned}$$

où $C = e^d + \frac{M}{\alpha!}$.

Ceci implique que $N(F, C\times\epsilon, \parallel . \parallel_\infty)$ n'est pas plus grand que le nombre de discrétisations ci-dessus. On a maintenant à évaluer ce nombre de discrétisations. Précisément, on considère le nombre de Af, avec

$$Af = \begin{pmatrix} A_{k1} f \\ A_{k2} f \\ \ldots \\ A_{kl} f \end{pmatrix}$$

où $k1, k2, \ldots, kl$ est l'ensemble des k avec $k. \leq \beta$. Ceci implique que l est le nombre de répartitions possibles de β boules identiques dans $d+1$ ensembles:

$$l = \frac{(d+1)^\beta}{\beta!}$$

Ainsi, chaque colonne de la matrice peut alors avoir u^l différentes valeurs, avec u le nombre de valeurs possibles pour un $A_k(f)$. u est borné par $\lfloor 2M/\delta^{\alpha-k.} \rfloor$. Ainsi, $u^l \leq \lfloor 2M/\delta^{\alpha-k.} \rfloor^{\frac{(d+1)^\beta}{\beta!}} \leq \lfloor 2M/\epsilon \rfloor^{\frac{(d+1)^\beta}{\beta!}}$.

On utilisera ceci seulement pour la première colonne, ie pour un des x_i, disons x_0. En outre, on demande que les x_i soient choisis de manière à ce que pour tout $i > 0$, il y ait $j < i$ tel que $\parallel x_i - x_j \parallel \leq 2\delta$.

Ainsi, considérons f, et considérons le nombre de valeurs possibles pour $\lfloor \frac{D^k(x_i)}{\delta^{\alpha-k}} \rfloor$, $\lfloor \frac{D^k(x_j)}{\delta^{\alpha-k}} \rfloor$ étant donnés pour $j < i$.

Grâce à l'hypothèse ci-dessus, il y a j tel que $\parallel x_i - x_j \parallel \leq 2\delta$. Ainsi, grâce à la formule de Taylor:

$$|D^k f(x_i) - \sum_{k.+l.\leq\beta} D^{k+l} f(x_i) \frac{(x_i - x_j)^l}{l!}| \leq \frac{M}{(\alpha - k.)!}(2\delta)^{\alpha - k.}$$

Ainsi, avec $B_k f = \delta^{\alpha - k.} A_k f$,

$$\begin{aligned} &|D^k f(x_i) - \sum_{k.+l.\leq\beta} B_{k+l} f(x_i) \frac{(x_i - x_j)^l}{l!}| \\ \leq &\underbrace{|D^k f(x_i) - \sum_{k.+l.\leq\beta} D^{k+l} f(x_i) \frac{(x_i - x_j)^l}{l!}|}_{\leq \frac{M}{(\alpha - k.)!}(2\delta)^{\alpha - k.}} + \underbrace{|\sum_{k.+l.\leq\beta} (D^{k+l} f(x_i) - B_{k+l} f(x_i)| \frac{(x_i - x_j)^l}{l!}}_{\leq \sum_{k.+l.\leq\beta} \delta^{\alpha - k. - l.} \frac{(2\delta)^{l.}}{l!}} \end{aligned}$$

Ainsi notons $L = (\frac{M}{(\alpha)!}(2)^{\alpha} + \sum_{l.\leq\beta} \delta^{-l.} \frac{(2\delta)^{l.}}{l!})$. On a prouvé que les valeurs possibles pour $A_k f$, pour un x_i donné, sont dans un intervalle de longueur $\leq L$. Comme $A_k f$ est entier, il y a au plus $L + 1$ valeurs possibles.

Ceci implique que le nombre maximal de choix pour Af est borné par

$$\left(\lfloor \frac{2M}{\epsilon} \rfloor \right)^{\frac{(d+1)^{\beta}}{\beta!}} \times (L+1)^{m-1}$$

d'où le résultat attendu.□

Corollaire 4.13 (Bracketing nombres de couverture) *Les bracketing nombres de couverture $N_{[\,]}(\mathcal{F},2\epsilon,L_r(Q))$ du même ensemble de fonctions pour 2ϵ pour tout $r \geq 1$ est borné par le nombre de couverture $N(\mathcal{F},\epsilon,L_\infty)$ pour ϵ, pour toute mesure de probabilité Q.*

Proof:Considérer simplement les centres f_i du ϵ-réseau, puis les $[f_i - \epsilon, f_i + \epsilon]$.□

Corollaire 4.14 (Bracketing nombres de couverture des sous-graphes) *Considérons $\mathcal{F}'$ l'ensemble des sous-graphes de fonctions de $\mathcal{F}$. Alors*

$$N_{[\,]}(2\epsilon,\mathcal{F}',L_1) \leq N(\mathcal{F},\epsilon,L_\infty)$$

Remarque 4.15 *Dans [Teytaud et al, 2001] les auteurs utilisent un lemme technique (voir lemme 17) pour montrer que la borne ci-dessus sur les sous-graphes de fonctions de $[0,1]^{d-1} \to [0,1]$ avec une condition de Hölder pouvaient être étendue à une autre borne pour les sous-ensembles de $[0,1]^d$ avec une condition de Hölder similaire (voir figure 4.3.2 pour cerner le besoin d'une telle extension dans le cadre des modèles déformables), pourvu que la surface soit bornée. Malheureusement, ceci amène à des constantes multiplicatives déraisonnables. On va supposer par la suite que les sous-ensembles de $[0,1]^d$ sont choisis d'une telle manière qu'une application π existe avec une dérivée de norme bornée telle que via π les modèles déformables sont à prendre de la forme précédente. Ceci est en gros équivalent à considérer que la topologie de la forme recherchée est connue. Le coefficient multiplicateur est maintenant la borne sur la dérivée de π.*

Les nombres de couverture existent pour d'autres formes de fonctions. Par exemples, les nombres de couverture des ensembles convexes ont été étudiées dans [Bronstein, 1976, Dudley, 1974]. On suppose ici que $d \geq 2$. Une preuve du théorème suivant peut être trouvée dans [Van der Vaart et al, 1996, p163]. $\mathcal{F}$ est la famille des convexes dans $[0,1]^d$.

Théorème 4.16 *Pour toute mesure de probabilité Q absolument continue par rapport à la mesure de Lebesgue avec densité bornée*

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

Lemme 4.17 *Soit $\mathcal{F}$ égal à l'ensemble des modèles déformables dans $[0,1]^d$ d'intérieur connexe, de frontière de courbure bornée par C_{max} et de surface bornée par S_{mac}. Il existe un ensemble fini (de cardinal $N(d,C_{max},S_{max})$) de paramétrisations tel que tout $f \in \mathcal{F}$ peut être défini en utilisant de telles paramétrisations.*

Preuve:

La preuve est écrite pour la dimension 3 mais peut être adaptée à toute dimension.

Le caractère borné de la courbure conduit à ϵ tel que chaque point i de la frontière du modèle déformable f peut être le centre d'un disque D_i tangent au modèle déformable de centre O_i et de rayon ϵ, avec f localement paramétré sur le disk de centre O_i et de rayon 2ϵ par $z = f_i(x,y)$ avec $\| \nabla f_i \|$ borné par un K fixé. On choisit D_i tangent au modèle déformable.

On choisit alors $\epsilon' > 0$ tel qu'une variation de O_i d'amplitude bornée par ϵ' et une variation de n_i (normale unitaire à D_i) d'amplitude bornée par ϵ' conduit à une variation de $\| \nabla f_i \|$ plus petite que K et ue variation de $\{x_i,y_i,f(x_i,y_i)\}$ plus petite que ϵ. Ceci fait que de telles variations amènent à des paramétrisations valides dans D_i (on rappelle qu'on définit ϵ suffisamment petit pour rendre les paramétrisations f_i valides dans un disque de rayon 2ϵ).

On choisit maintenant un ensemble fini S_1 de points de $[0,1]^d$ et un ensemble fini S_2 de points sur la sphère unité de $\mathbb{R}^d$, tels que chaque boule de rayon ϵ' dans $[0,1]^d$ intersecte S_1, et chaque boule de rayon ϵ' centrée sur un point de la sphère unité de $\mathbb{R}^d$ intersecte S_2. Soit D l'ensemble des disque de centres dans S_1 et de normal dans S_2, de rayon ϵ'. S_1 et S_2 peuvent être choisis indépendamment de f.

Pour chaque D_i, on considère le $\tilde{D}_i$ le plus proche dans D. D étant fini, f est paramétré sur des disques choisis dans un ensemble fini. Malheureusement, nous n'avons pas prouvé que chaque disque était utilisé un nombre fini borné de fois. Ceci provient simplement du fait que revenir dans la même paramétrisation requiert une surface minorée, et ainsi le nombre de retours est borné.□

Corollaire 4.18 *Soit $\mathcal{F}$ égal à l'ensemble des sous-ensembles de $[0,1]^d$ d'intérieur connexe, avec frontière de courbure bornée par C_{max}, $|||f|||_\alpha$ borné par 1 avec $\alpha > 2$, et de surface bornée par S_{max}. Alors avec q borne sur la densité de Q par rapport à la mesure de Lebesgue,*

$$\log N_{[\,]}(\epsilon,\mathcal{F},L_r(Q)) \leq K(d,S_{max},C_{max}) q^{((d-1)/\alpha)} (\frac{1}{\epsilon})^{r(d-1)/\alpha}$$

Preuve:

Grâce au lemme 17, on peut borner le nombre de paramétrisations possibles par $N(d,C_{max},S_{max})$. Ainsi les modèles déformables de $\mathcal{F}$ peuvent être ϵ-bracketés par l'union des $N(d,C_{max},S_{max})$ brackets $(f_i,g_i)_{i\in I}$ sur des paramétrisations locales (pour les intersections de paramétrisations, on peut simplement considérer, avec coût linéaire sur ϵ, le max des g_i et le *min* des f_i). Les bracketing nombres de couverture sont ainsi bornés linéairement (pour un C_{max} fixé) par les bracketings nombres du corollaires 14.□

On note que l'existence de C_{max} est garantie par l'existence de $|||f|||_\alpha$ pour $\alpha \geq 2$. Ainsi, cette hypothèse peut être supprimée.

4.3 Concrètement

4.3.1 Le bruit

Les résultats donnés ci-dessus et basés sur des théorèmes antérieurs seraient de peu d'intérêt dans un cadre complètement déterministe. La géométrie discrète fournit des bornes sur les approximations déterministes lorsque les points ne sont pas bruités.

Basiquement, on considère X_i une variable aléatoire choisie indépendamment uniformément dans le i^e pixel (voir définitions plus haut). Ensuite, plusieurs sortes de bruit peuvent être considérées:

- Bruit aléatoire: $X_i' = (X_i,B)$ avec $B = 1$ avec probabilité p et 0 avec probabilité $1-p$. Alors, un modèle f est remplacé par nf tel que $nf(X_i') = nf(X_i,B) = (1-B) \times f(X_i) - B \times f(X_i) = (1-2B)f(X_i)$.
- Bruit additif: $X_i' = (X_i,B)$ avec B une variable aléatoire centrée. Alors, un modèle f est remplacé par nf tel que $nf(X_i') = nf(X_i,B) = f(X_i) + B$.

Les notations avec nf et $X'_i = (X_i,B)$ sont choisies de manière à garder le même formalisme simple du problème de convergence de l'évaluation empirique des moyennes des $f(X_i)$. Ceci amène à la remarque suivante:

Remarque 4.19 *Notez que les résultats de convergences ont lieu pour l'espérance après un bruit éventuel. Ainsi, nous n'avons pas, en section 4.3.2, à considérer différentes formes de bruit. D'autre part, les bornes ou vitesse de convergence obtenues ont lieu* pour Val_n et Val avec bruit, *et ils ont à être traduits en résultats dans le cas sans bruit. Par exemple, si on considère un bruit aléatoire (voir ci-dessus), les espérances sont multipliées par* $1-2p$ *quand du bruit est ajouté. Ceci fournit une évaluation calculable du facteur de dépendance de la vitesse de convergence en le taux de bruit:* $1-2p$. *Dans le cas d'un bruit additive borné, la seule différence de vitesse de convergence est basée sur le fait que nous supposons une borne sur les valeurs absolues des pixels, laquelle est modifiée par l'ajout de bruit. Pour des examples non bornés, des théorèmes comme 6 peuvent toujours être appliqué (le formalisme est exactement le même avec ou sans borne!) mais bien sûr, l'estimateur est évalué en termes de variance (théorème 6, la limite de la convergence uniforme est donnée par la convergence simple des variances) et les variances sont plus grandes avec bruit.*

4.3.2 Applications

Applications non-asymptotiques

Le théorème 8 fournit des bornes $B(\tilde{F},n,\delta)$ telles qu'avec confiance au moins $1-\delta$

$$\forall \tilde{f} Val(\tilde{f}) \leq Val_n(\tilde{f}) + B(\tilde{F},n,\delta)$$

Une solution classique en théorie de l'apprentissage quand de tels résultats ont lieu, est que l'on devrait minimiser

$$\text{argmin}_{\bar{F}\in\{\bar{F}_1,\ldots,\bar{F}_k\},\tilde{f}\in\bar{F}} Val_n(\tilde{f}) + B(\tilde{F},n,\delta)$$

Cette suggestions étant justifiée par le fait que, dans certains cas, on peut utiliser des bornes similaires de la forme

$$\forall \tilde{f} Val(\tilde{f}) \geq Val_n(\tilde{f}) + B'(\tilde{F},n,\delta)$$

où B' est lié à B (équivalent à des facteurs constants ou logarithmiques près) et où la distribution est la pire possible. Ceci peut être justifié, en outre, par le fait qu'un meilleur comportement asymptotique peut être garanti par cet algorithme. Ceci est connu dans la littérature comme la "Minimisation Structurelle du Risque" (ou "Minimisation du Risque Structurel" dans certaines religions), et [Devroye et al, 1996] fournit un résumé de tels résultats justifiant ce point de vue dans le cas d'une union infinible dénombrable de familles de VC-dimension finie ou dans le cas d'estimateurs par squelettes. Néanmoins, le premier argument, basé sur des bornes supérieures et inférieures, est seulement justifié dans une étude au pire cas. Le second, basé sur des comportements asymptotiques, est la consistence, ce qui signifie que pour toute distribution l'algorithme converge vers le coût optimal dans la famille. Pour des ensembles bien choisis de fonctions, la consistance pourrait être universelle, dans le sens où le coût optimal est le plus petit possible parmi *toutes* les fonctions.

Considérons l'algorithme suivant, directement inspiré de [Vidyasagar, 1997, p188]:

- Considérons δ un risque donné.
- Définissons $\epsilon > 0$ aussi grand que possible tel que $n \geq \frac{8}{\epsilon^2}\ln(\frac{k}{\delta})$, avec k le cardinal d'une $\epsilon/2$-couverture de F'.
- Considérons une $\epsilon/2$-couverture $f'_1,\ldots,f'_k$ de l'ensemble des modèles.
- Soit $f' = \text{argmin}_{g\in f'_1,\ldots,f'_k} Val_n(g)$.

Alors, avec probabilité au moins $1-\delta$, pour tout $h' \in H'$, $Val(f') \leq Val(h') + \epsilon$ et $Val(f') \leq Val(f) + \epsilon$.

Si les nombres de couverture sont tous finis (la famille est dite *totalement borné*) cet algorithme garantit une convergence vers le modèle déformable optimal), et fournit une borne supérieure avec confiance $1-\delta$

sur l'erreur: l'algorithme est dit PAC avec précision ϵ et confiance $1-\delta$. Une remarque importante soulignée dans [Vidyasagar, 1997, p188] est qu'on n'a pas besoin d'utiliser une ϵ-couverture optimale.

Ainsi les résultats théoriques ci-dessus ont les conséquences pratiques ci-dessous:

Conséquences pratiques
L'algorithme suivant a précision ϵ avec confiance $1-\delta$: – Considérons F une famille de modèles déformables. – Choisir une ϵ-couverture de F. – Minimisons le risque empirique, avec le nombre de pixels choisi ci-dessus, sur une ϵ-couverture. – Avec confiance $1-\delta$, le modèle résultant est le meilleur avec précision ϵ. Les nombres de couverture nécessaires pour cette conclusion peuvent être choisis grâce au corollaire 14 dans le cas des modèles réguliers. Notez que ces nombres de couverture sont vrais pour les sous-graphes de fonctions, concrètement, cela signifie qu'on considère des modèles déformables comme montré en figure4.3.2 (gauche). L'extension au cas incluant la figure 4.3.2 (droite) peut être fait grâce au lemme 17 (qui entraîne des nombres de couverture impratiquables) ou grâce à la remarque 15.

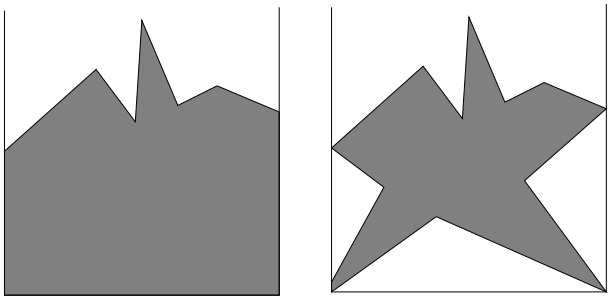

FIG. 4.1 – *Gauche: cas géré par le théorème 12. Droite: cas géré par le lemme 17 ou la remarque 15. A gauche, la frontière est définissable par une simple fonction de l'abscisse, pas à droite.*

Un tel résultat justifie un choix incrémental de la taille d'une famille de Hölder de fonctions dépendant du nombre d'exemples: accroître (lentement) α ou M entraîne la consistance universelle, ie la convergence asymptotique vers le meilleur modèle.

Applications asymptotiques

Dans le cadre asymptotique, on a les conséquences qui suivent:

Conséquences pratiques
Le modèle déformable choisi empiriquement converge faiblement en $O(1/\sqrt{n})$ (parce que l'ensemble des $\int \tilde{f}$ est uniformément évalué avec précision faiblement $O(1/\sqrt{n})$) avec n le nombre de pixels dans chacun des cas suivants, avec d la dimension ($X = [0,1]^d$): – Avec des fonctions de coût $\tilde{f}$ binaires: – Avec des corps convexes, quand $d \leq 2$. – Avec des frontières α-Hölder, avec $\alpha > d-1$. – Avec des fonctions de coût $\tilde{f}$ à valeurs réelles, avec les $\tilde{f}$ des fonctions α-Hölderiennes, quand $\alpha > d$.

4.4 Remarques, généralisations et conclusions

Notez que le lemme technique 17 ne préserve pas la nature explicite du théorème 12, et que l'hypothèse requise dans la remarque 15 n'est pas très naturelle. Une extension du théorème 12 en direction du lemme 17 serait utile. Voir la figure 4.3.2 pour bien comprendre la nécessité d'une adaptation du théorème 12.

Nous travaillons sur des fonctions de coût intégrées sur les intérieurs des modèles déformables. Nos résultats peuvent ils se généraliser au cas plus usuel: des intégrales sur les frontières? La réponse est positive à la fois pour des modèles réguliers ou des modèles convexes. Considérez la classe $\mathcal{F}''$ des η-élargissements intérieurs[1] de frontières de modèles déformables. L'argument pour préserver les résultats de convergence rapide est le suivant:

Fait 1: Les logarithmes des bracketing nombres de couverture de $\mathcal{F}' = \{A \setminus B/(A,B) \in \mathcal{F}\}$ sont bornés (à un facteur multiplicatif près) par les bracketing nombres de couverture de $\mathcal{F}$.

Fact 2: Les bracketing nombres de couverture des frontières élargies d'éléments de $\mathcal{F}''$ sont bornés par ceux de $\mathcal{F}'$.

Fact 3: La condition 4.5 est clairement vérifiée, les conditions 4.6 et 4.7 sont prouvées grâce aux lemmes 9 et 10.

Un inconvénient de cette approche est que nous considérons des η élargissements de frontières, avec η indépendants de n. Intuitivement, on préfèrerait peut-être η équivalent à $1/\sqrt[d]{n}$. Nous n'avons pas complété la preuve du fait que cela préserve les résultats de vitesse, mais nous le supposons. La raison est que de tels résultats sont souvent vrais pourvu que f_n et g_n, respectivement les $\eta = 1/\sqrt[d]{n}$-élargissements intérieurs de modèles déformables f et g, vérifient le fait que $E(f_n \times g_n) - E(f_n) \times E(g_n)$ converge simplement en f et g, ce qui est probablement vrai sous de légères hypothèses sur I.

Les modèles déformables usuels sont placés par descente de gradient. Quelle part de nos résultats est perturbée par ce cadre pratique? La validation (ie, les bornes fournies sur la différence entre coût empirique et coût en généralisation) peut être faite de la même manière; la convergence des valeurs empiriques vers les "vraies" valeurs est un fait *uniforme*. La convergence vers la valeur optimale est un autre problème. Le fait que converger vers un minimum global au lieu d'un minimum local n'est même pas prouvé être une meilleure solution, comme la finitude de la descente de graident peut être vue comme une régularisation supplémentaire (surtout dans le cas d'un choix par l'utilisateur d'un point initial - [Lai et al, 1993] considère un choix automatique du point initial par transformée de Hough).

Le terme de régularisation pour des fonctions régulières est basé sur les normes L^∞ des dérivées. Nous ne savons pas si des résultats peuvent être établis sur les normes L^2 des dérivées.

Finalement, on peut critiquer l'intérêt pratique de ces résultats. Au-delà des conclusions théoriques, quels sont les résultats concrets? On peut donner trois conclusions pratiques:

- On affirme qu'en dimension 3 $\alpha > 2$ est théoriquement nécessaire pour garantir une convergence rapide. Peut-être que $\alpha = 2$ étant le cas tangent (2.001 est suffisant!), il est suffisant en pratique pour garantir une convergence en $O(1/\sqrt{n})$.
- On affirme que la convexité est suffisante en dimension ≤ 2 seulement.
- En dimension 4 (ce qui a un sens dans certaines applications, comme les films cardiaques) $\alpha > 3$ est requis. Pour la même raison que ci-dessus, $\alpha = 3$ serait probablement suffisant. Il serait intéressant de tester expérimentallement la conclusion selon laquelle $\alpha = 2$ n'est pas suffisant.

Des travaux récents comme [Radulovic et al, 2000] permettent d'étudier de nouvelles propriétés des processus empiriques en tirant parti de différentes hypothèses de régularité sur la distribution des exemples. Il serait certainement intéressant de recommencer l'étude ci-dessus en adoptant ce point de vue. Le principe est d'utiliser, comme approximation de P, non pas la somme des masses de Dirac en les X_i, mais $\frac{1}{nh}\sum_{i=1}^{n} K(\frac{x-X_i}{h})$ avec K tel que $\int xK(x)dx = 1$, $\int x^2K(x)dx$ et $\int K^2(x)$ soient finis. h est choisi tel que $nh^2 \to \infty$ et $nh^4 \to 0$. Ceci peut se reformuler comme une convolution. Dans le cas général, sans régularité, Yukich [Yukich, 1992] et Van der Vaart [Van der Vaart, 1994] ont montré que cet estimateur était à peu près aussi bon que l'estimateur initial (ie, par masses ponctuelles), en passant, en fait, par des bornes sur la différence entre ces deux estimateurs. Mais Radulovic et Wegkamp [Radulovic et al, 2000] vont plus loin, en montrant que cet estimateur est meilleur dans certains cas, car il permet des résultats positifs dans le cas

1. Le η-élargissement intérieur de la frontière de f est l'ensemble des x tels que $f(x) = 1$ et $\exists y / f(y) = -1$ et $d(x,y) < \eta$.

de familles prégaussiennes, non nécessairement Donsker (Donsker pour une loi = vérifiant la conclusion du théorème 6 pour des exemples iid, ie $Z_{n,i}(f) = \frac{1}{\sqrt{n}} f(X_{n,i})$ avec les $X_{n,i}$ iid suivant cette loi). En revanche, des conditions plus fortes sont imposées sur la régularité ; dans ce cadre, ils démontrent l'aspect nécessaire et suffisant du caractère prégaussien pour des fonctions indicatrices. En outre, ils soulignent que l'estimateur classique peut échouer sur des familles prégaussiennes non-Donsker. Cet estimateur est donc *supérieur*, et non simplement équivalent, à l'estimateur usuel.

Les avancées suivantes ne sont pas effectuées dans ce papier:

1. En utilisant les bornes de Chernoff sous forme multiplicative on peut aisément obtenir de bien meilleures inégalités pour des échantillons iid que par l'inégalité de Hoeffding, quand l'espérance de $\tilde{f}$ (avec f le meilleur modèle) est faible et quand les pixels sont des variables de Bernoulli. Dans le cas de variables non-binaires une autre solution consiste en utiliser l'inégalité de Bernstein, qui améliore celle de Hoeffding dans le cas de petite variance.
2. Notre notion de pixel peut choquer (valeur choisie aléatoirement uniformément sur l'ensemble du pixel). On aimerait peut-être considérer une grille rigide, par exemple en considérant un angle aléatoire (sans aléatoire du tout, des contre-exemples clairs bloquent toute borne intéressante).
3. On peut généraliser l'approche en considérant F dépendant de n ou bien la somme de L et R.

Bibliographie

[Andersen et al, 1988] N.-T. Andersen, E. Giné, M. Ossiander, J. Zinn, *The central limit theorem and the law of iterated logarithm for empirical processes under local conditions, Probability theory and related fields 77, 1988.*

[Birman et al, 1967] M.-S. Birman, M.-Z. Solomjak, *Piecewise-polynomial approximation of functions of the classes W_p, Mathematics if tge USSR Sbornik 73, 295-317, 1967.*

[Bronstein, 1976] E.-M. Bronstein, *Epsilon-entropy of convex sets and functions, Siberian Mathematics Journal 17,393-398, 1976.*

[Devroye et al, 1996] L. Devroye, L. Gyorfi, G. Lugosi, *A probabilistic theory of pattern recognition, 1996.*

[Dudley, 1967] R.-M. Dudley, *The sizes of compact subsets of Hilbert spaces and continuity of Gaussian processes. Journal of Functional Analysis 1, 290-330, 1967.*

[Dudley, 1974] R.-M. Dudley, *Metric entropy of some classes of sets with differentiable boundaries. Journal of Approximation Theory 10, 227-236, 1974. Correction; Journal of Approximation Theory 26, 192-193, 1979.*

[Dudley, 1978] R.-M. Dudley, *Central limit theorems for empirical measures, Annals of probability 6, 1978. Correction: Annals of probability 7, 1978.*

[Dudley, 1984] R.-M. Dudley, *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.*

[Hoeffding, 1963] W. Hoeffding, *Probability inequalities for sums of bounded random variables. J. Amer. Statist. Assoc. 58, pp13-30, 1963.*

[McInerney et al, 1996] T. McInerney, D. Terzopoulos, *Deformable Modeles in Medical Image Analysis: A Survey, Medical Image Analysis, 1(2), 1996.*

[Kolcinski, 1981] V.-I. Kolcinski *On the central limit theorem for empirical measures. Theory of probability and mathematical statistics 24, 1981.*

[Kolmogorov et al, 1961] A.-N. Kolmogorov, V.-M. Tikhomirov, *ϵ-entropy and ϵ-capacity of sets in functional spaces, Amer. Math. Soc. Transl. 17, pp 277-364, 1961.*

[Koul, 1970] H. Koul, *Some convergence theorems for ranks and weighted empirical cumulatives. Annals of Mathematical Statistics 41, 1768-1773, 1970.*

[Lai et al, 1993] K.-F. Lai, R.-T. Chin, *On regularization, Formulation and Initialization of the Active Contour Models (Snakes),Proc. 1st Asian Conf. on Computer Vision, 542-545, 1993.*

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

[Marcus et al, 1984] M. Marcus, J. Zinn, *The bounded law of the iterated logarithm for the weighted empirical distribution in the non-iid case. Annals of Probability 12, 334-360, 1984.*

[Ossiander, 1987] M. Ossiander, *A central limit theorem under metric entropy with L_2 bracketing, Annals of probability 15, 1987.*

[Pollard, 1982] D. Pollard, *A central limit theorem for empirical processes. Journal of the australian mathematical society, A33, 1982.*

[Radulovic et al, 2000] D. Radulovic, M. Weckamp, *Weak convergence of smoothed empirical processes: beyond Donsker classes. High Dimensional Probability II, E. Gine, D. Mason and J. Wellner, Editors. Birkhauser, 2000.*

[Shorack, 1973] G.R. Shorack, *Convergence of reduced empirical and quantile processes with application to functions of order statistics in the non-iid case. Annals of Statistics 1, 146-152, 1973.*

[Shorack, 1980] G.R. Shorack, *The weighted empirical process of row independent random variables with arbitrary distribution functions. Statistica Neerlandica 33,169-189 (1980).*

[Shorack et al, 1986] G.R. Shorack, J. Beirlant, *The appropriate reduction of the weighted empirical process. Statistica Neerlandica 40, 123-12, 1986.*

[Shorack et al, 1986] G.R. Shorack, J.A. Wellner, *Empirical Processes with applications to statistics, Wiley, New-York, 1986.*

[Strassen et al, 1969] V. Strassen, R.M. Dudley, *The central limit theorem and ϵ-entropy. Lecture Notes in Mathematics 89, 224-231. Springer-Verlag, New York, 1969.*

[Sudakov, 1969] V.N. Sudakov, *Gauss and Cauchy measures and ϵ-entropy, Doklady Akademii Nauk SSSR 185, 51-53, 1969.*

[Teytaud et al, 2001] O. Teytaud, D. Sarrut, *Convergence speed of deformable models, proceedings of Ijcnn 2001.*

[Vidyasagar, 1997] M. Vidyasagar, *A theory of learning and generalization, Springer 1997.*

[Van der Vaart, 1994] A.W. Van der Vaart, *Weak convergence of smoothed empirical processes. Scandinavian Journal of Statistics, 21, 501-504, 1994.*

[Van de Geer, 1991] S. Van de Geer, *The entropy bound for monotone functions. Report 91-10. University of Leiden, 1991.*

[Van der Vaart et al, 1996] A.-W. van der Vaart, J.-A. Wellner, *Weak convergence and Empirical Processes, Springer, 1996*

[Van Zujilen, 1978] M. Van Zujilen, *Properties of the empirical distribution function for independent not identically distributed random variables. Annals of Probability 6, 250-266, 1978.*

[Yukich, 1992] J.E. Yukich, *Weak convergence of smoothed empirical processes. Scandinavian Journal of Statistics, 19, 271-279, 1992.*

Chapitre 5

Apprentissage de séquences non-indépendantes d'exemples. Identification de système, contrôle et stabilisation.

Résumé

Beaucoup de travaux récent considérent les applications pratiques des réseaux neuronaux pour le contrôle. Quelques papiers seulement (dont les résultats principaux sont rappelés ici) ont été consacrés aux applications de la partie théorique de l'apprentissage au contrôle. Ce papier fournit:

- Les notations et définitions pour l'identification de système, la stabilisation et le contrôle.
- Un état de l'art aussi complet que possible de résultats à propos des séries temporelles stationnaires, ergodiques ou chaotiques (théorie de l'apprentissage avec dépendance temporelle).
- Introductions historiques aux domaines scientifiques qui intersectent l'identification de système, la stabilisation ou le contrôle: physique statistique, dynamiques stochastiques de systèmes déterministes, processus empirique et VC-théorie, logique floue, réseaux neuronaux et outils d'apprentissage liés, modèles de Markov.
- Des résultats théoriques découlant de la réunion de ces différents domaines.
- Illustrations pratiques et benchmarks classiques, avec références à des algorithmes pratiques.
- Problèmes théoriques ouverts en identification de système, stabilisation et contrôle.

Ce chapitre est une version très très étendue de travaux préliminaires publiés dans les actes du VIIIème symposium Van Der Ziel sur les bruits en 1/f dans les phénomènes quantiques, d'EFTF 2001 et d'ICNF 2001 (JM Friedt, O. Teytaud, M. Planat, D. Gillet).

Table des matières

5.1 Introduction

La théorie de l'apprentissage est une grande aire de recherche, presque prête pour une extension à l'identification de système, au contrôle et à la stabilisation. De nombreuses bornes ont été fournies dans le cadre général de l'apprentissage d'une relation, puis améliorées. [Devroye, 1996, Vidyasagar, 1997] sont des états de l'art complets dans ces domaines. Des conditions nécessaires et suffisantes en classification (résumées dans [Vapnik, 1995]), et plus tard en régression (voir [Alon et al, 1997]) ont été fournies pour la convergence uniforme des moyennes empiriques vers les espérances, uniformes à la fois en la fonction et en la distribution. Des vitesses de convergence rapide ont été prouvées lorsque l'erreur optimale est nulle (voir [Devroye, 1996] pour un état de l'art) ou petite ([Vapnik, 1982]). Des résultats plus restreints, et peut-être plus efficaces, sont fournis sous certaines hpothèses: [Vidyasagar, 1997] liste des résultats quant à la sélection dans des familles de VC-dimension éventuellement infinie par ϵ-squelettes (utilisant les nombres de couverture). Alors que ces résultats viennent de la communauté de l'intelligence artificielle, les mathématiciens, dans la communauté du processus empirique, ont prouvées de nombreux théorèmes centraux généralisés, uniformes sur des espaces de fonctions et sur des espaces de distributions, qui sont résumés dans [Van der Vaart et al, 1996] et ont des applications en imagerie, en maximum de vraisemblance ou en inférence Bayésienne par exemple ([Van der Vaart et al, 1996, Teytaud et al, 2001, Teytaud et al, 2001]).

Après beaucoup d'améliorations, la théorie de l'apprentissage a été étendue dans d'autres directions: la distribution identique a été réduite dans l'aire de recherche maintenant vaste des erreurs malicieuses (l'occurence de bruit est indépendante identiquement distribuée comme dans [Decatur, 1995, Decatur, 1997], mais des exemples bruités sont choisis de la pire manière possible par un adversaire de puissance de calcul illimitée, voir [Kearns et al, 1993, Gentile et al, 1998]), ou peut être traitée par des théorèmes basés sur l'inégalité de Hoeffding. [Aldous et al, 1990, Gamarnik, 1999, Johansen et al, 1997] fournissent des essais d'extensions dans la direction Markovienne. Un but de ce travail est une extension dans la même direction. Bien que l'article soit axé sur la théorie, on rappelle les principaux algorithmes usuellement mis en œuvre en contrôle et des cas concrets sont présentés.

[Haykin et al, 1998, Hong, 1993] étudient la possibilité de prédire des systèmes chaotiques avec des outils de régression, éventuellement des réseaux neuronaux. [Mukherjee et al, 1997] vérifie la validité pratique de la théorie de l'apprentissage dans le cas de telles séries temporelles et conclut que la théorie VC est validée. Néanmoins, les hypothèses classiques ne sont pas vérifiées dans un tel cas: des points consécutifs fournis par

un système chaotique ne sont évidemment pas indépendants identiquement distribués. [Friedt et al, 2000, Teytaud et al, 2001, Friedt et al, 2001] ont souligné ce manque de résultat théorique et ont proposé une application de ces prédictions. L'idée consiste à transformer une prédiction en stabilisation, comme expliqué ci-dessous.

La partie 5.2.1 rappelle des notations & définitions. La partie 5.2.2 est un résume bref du cadre historique. La partie 5.3.1 résume l'apprentissage non-iid[1] (la partie 5.3.1 fournit de nouveaux théorèmes non-asymptotiques basés sur la VC théorie et des résultats classiques sur les chaines de Markov ergodiques; la partie 5.3.2 explique la conversion de la prédiction vers le contrôle). La partie 5.4 résume quelques faits expérimentaux utiles pour comprendre les objectifs théoriques (la partie 5.4.1 détaille l'application la plus simple, ie l'identification de système/la prédiction; la partie 5.4.2 discute certains résultats suggérant de possibles applications en contrôle et en stabilisation). En outre, des applications concrètes sont citées et des méthodes et benchmarks sont rappelés. La partie 5.5 est constituée de quelques remarques et souligne quelques problèmes ouverts en théorie du contrôle.

5.2 Identification de système, contrôle, stabilisation: définitions et état de l'art

5.2.1 Notations et définitions

Dans cette section on définit un problème, qui dans le même cadre comprend l'identification de système, la stabilisation et le contrôle. On travaille seulement dans le cadre du temps discret. Les notations ci-dessous sont (presque) consistances avec celles de [Sontag, 1993] et sont illustrées en figure 5.2.1, et nous définissons à la fin de cette section un modèle simplifié qui est (au moins théoriquement) équivalent au cas général ci-dessous sous des hypothèses raisonnables:

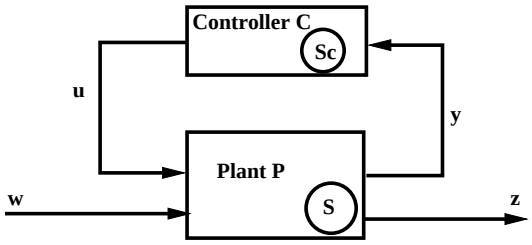

FIG. 5.1 – *Cadre général du contrôle, de la stabilisation, et de l'identification de système. Voir le texte.*

- S est l'espace des **états** du **système**, S_C est l'espace des **états** internes du **contrôleur**. W, Y, U sont respectivement des espaces d'entrée, de rétroaction, de control.
- Le **système** P est une application $W \times U \times S \to Z \times Y \times S$.
- Le **controleur** C est une application $Y \times S_C \times \mathbb{N} \to U \times S_C$. Comme montré par les équations et en figure 5.2.1, le controleur est une **rétroaction**.
- w_t pour $t \in \mathbb{N}$ est le **signal d'entrée**.
- z, **signal de sortie**, s **état du système**, et u, **signal de contrôle** sont définis comme suit:
  - $(z_{n+1},y_{n+1},s_{n+1}) = P(w_n,u_n,s_n)$.
  - $(u_{n+1},sc_{n+1}) = C(y_n,sc_n,n)$.

  Notez que ces deux fonctions peuvent être non-déterministes.
- Un signal cible r, une famille $\mathcal{F}$ de fonctions possibles C, et une fonction de coût L sont données, et l'objectif est la minimisation de $L(r,z)$ par le choix de $C \in \mathcal{F}$. Les algorithmes ont accès à une suite de signaux d'entrée/sortie en choisissant le signal de contrôle.

1. Iid=indépendant identiquement distribué.

Notez que w n'a pas été inclus dans les entrées de C, mais n si. Ainsi, par un choix ad hoc d'une famille $\mathcal{F}$, on a la possibilité d'inclure w comme entrée. Notez qu'en pratique, w_n sera souvent considéré comme une chaîne de Markov, et ainsi pourra être inclus dans le système lui-même. La même chose peut être dite du signal cible r_n. En outre, les états internes du controleur peuvent toujours être supprimés par une modification *ad hoc* de P et $\mathcal{F}$. Le problème est le choix de C parmi une famille $\mathcal{F}$. Notez que cette description est beaucoup plus générale qu'il ne semble à première vue. Pour autant que nous le sachions il n'existe pas de formalisation du contrôle qui n'entre pas dans ce cadre avec un choix *ad hoc* de $\mathcal{F}$.

Le contrôle est dit **adaptatif** quand le controleur est supposé varier. Formellement, il n'est pas adaptatif quand S_C peut être réduit à un simple élément.

Trois sortes d'applications des réseaux de neurones ou d'autres outils de régression pour des applications du contrôle peuvent être distingués:

- **Identification de système, ie modélisation du système à controller**. Le réseau de neurones, ou un autre outil de régression, est utilisé pour prédire le comportement du système, et éventuellement l'analyse (automatique ou non) du réseau mène à un controleur. Les aspects adaptatifs nécessitent ici seulement une adaptation (éventuellement) **en ligne** (ie la fonction de régression est mise à jour quand de nouveaux points sont fournis) et une adaptation **non-iid** de la phase d'apprentissage. Si le signal est iid, alors il s'agit de la théorie de l'apprentissage classique.
- Choix dynamique des paramètres d'un controleur classique (par exemple, un controleur PID).
- **Rétroaction**. Le réseau de neurones est utilisé directement comme un controleur. Ceci est très différent des utilisations classiques de la théorie de l'apprentissage. L'adaptation au contrôle adaptatif n'est pas immédiate, ni théoriquement ni intuitivement. Dans certains cas, appelés **apprentissage général**, la rétroaction est basée sur une identification de système "inversée" (l'entrée du réseau de neurones est la sortie du système, et la sortie du réseau de neurones est l'entrée du système); alors, le signal de contrôle est prédit en appliquant le réseau de neurones au signal-cible. Une autre solution, appelée *apprentissage spécifique*, est basée sur l'insertion d'un réseau de neurones dans la boucle de feedback, qui est optimisé par divers algorithmes, principalement comme les algorithmes génétiques.

Nous allons focaliser notre attention, dans la suite, sur le premier cas. Une autre distinction est proposée dans [Sontag, 1993]: alors qu'en contrôle adaptatif usuel, l'idée est d'adapter des paramètres d'apprentissage dynamiquement pour adapter le controleur à des variations lentes, l'ainsi dénommé **contrôle par apprentissage**, basé sur des grandes mémoires et des vastes espaces de fonction (typiquement, le contrôle neuronal) essaie de bénéficier des expériences précédentes du système dans une aire similaire.

La **stabilisation** peut être vue comme une forme particulière de contrôle: la cible est, par exemple, constante: $L(r,z) = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n}(z_i - r)^2$. Cependant, d'autres formes de stabilisation existent: la minimisation de la variance ou d'autres caractéristiques liées (e.g. $L(r,z) = Var(z)$).

L'**identification de système** est le problème de choisir f dans une famille donnée $\mathcal{F}$ telle que $f(X_n) \simeq X_{n+1}$ where (X_n) is a sequence. Dans certains cas, l'identification de systèmes est adaptative: ceci implique que la "fonction" est maintenant équipée d'un état interne Y. Formellement, cela amène à:

$$\begin{aligned} Y_{n+1} &= f_1(Y_n) \\ X_{n+1} &= f_2(X_n,Y_n) \end{aligned}$$

où (f_1,f_2) est extrait d'un sous-ensemble de $\mathcal{F}_1 \times \mathcal{F}_2$. Les filtres de Kalman (voir partie 5.2.2) sont faits pour résoudre de telles tâches. L'identification de systèmes, par un choix ad hoc de la famille $\mathcal{F}$, peut être replacé dans le cadre précédent; simplement utiliser P' au lieu de P et par exemple $L(r,z) = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} z_i^2$, avec

$$P'_{|1}(w_n,u_n,0) = P(w_n,0,0) - C(P_{|2}(w_{n-1},0,0),0,n)$$

(on utilise $f_{|1}(x) = a$, $f_{|2}(x) = b$, $f_{|3}(x) = c$ si $f(x) = (a,b,c,d,e)$ et ainsi de suite)

Nous définissons maintenant une version restreinte du contrôle, qui est théoriquement équivalente à celle ci-dessus sous de légères hypothèses, mais est beaucoup moins explicite pour des discussions avec les praticiens

(en particulier pour l'inclusion de connaissances a priori). L'intérêt de ce modèle simplifié est qu'il est plus facile à manier pour les preuves ou pour la compréhension intuitive lorsque l'équivalence est bien comprise.

Ainsi, on considère que:

- L'état interne (S_C) du controleur peut être supprimé par une adaptation de S, C et $\mathcal{F}$.
- Les signaux d'entrée comme de sortie w_n et z_n peuvent être supprimés par une adaptation de P et L.
- L peut être réécrit comme $L(r) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} L_i(r_i)$.
- Le controleur est remplacé lui-même par une simple rétroaction sur l'entrée, et la variable à optimiser est le système f pris dans une famille F.

Formellement, ceci est

$$(s_{n+1},z_{n+1}) = f(s_n,z_n) \tag{5.1}$$

qui peut à son tour être réécrit

$$X_{n+1} = f(X_n)$$

et on minimise

$$L(X_n) = \lim_{n\to\infty} \frac{1}{n} \sum_i L_i(X_i)$$

Un algorithme de contrôle doit minimiser $L(X)$, en ayant accès aux $\Pi(X_n)$ seulement ($\Pi(X_n)$ défini plus bas, incluant au moins $L_n(X_n)$. L'identification de système, la stabilisation et le contrôle peuvent être mis dans ce cadre. Le cas de l'identification de systèmes n'est pas le cas le plus intéressant, comme des simplifications supplémentaires pourraient être faites (pour utiliser directement les résultats comme ceux de la partie 5.3.1). Il est défini dans ce cadre seulement afin de garder l'unité de description.

- Identificatioon de système: $X_n = (s_n,z_n)$, et F est fait de fonctions f de la forme

$$f(s_n,z_n) = (g(s_n),g(s_n) - h(\pi(s_n)))$$
$$L(z_n) = z_n^2$$

  où g est le système (inconnu), et h est pris dans l'ensemble des fonctions de prédiction. $\Pi(X_n) = z_n$. Notez que la description donnée dans l'équation pourrait encore être simplifiée (sans perte de généralité) afin de montrer que l'identification de système peut être réduite à la convergence uniforme des moyennes empiriques vers leurs espérances pour des séquences de variables non-indépendantes, ainsi que décrit en section 5.3.1.
- Stabilisation: $X_n = (s_n,z_n)$, et F est fait de fonctions f de la forme

$$f(s_n,z_n) = (g(s_n,h(\pi(s_n))),\pi_2(s_n)) \tag{5.2}$$

  où g est la fonction (inconnue) qui combine l'état interne et le signal de contrôle. h est pris dans la famille des controleurs, $\pi(s)$ est l'information mesurée quand l'état interne est s, $\pi_2(s)$ est la mesure d'"instabilité" quand l'état interne est s. $\Pi(X_n) = (\pi(s_n),\pi_2(s_n))$.
- Contrôle: $X_n = (s_n,z_n)$, et F est fait de fonctions f de la forme

$$f(s_n,z_n) = (g(s_n,h(\pi(s_n))),\pi_2(s_n))$$

  où g est la fonction (inconnue) qui combine l'état interne et le signal de contrôle, h est pris dans la famille de controleurs, $\pi(s)$ est l'information mesurée quand l'état interne est s, $\pi_2(s)$ est la mesure du coût en terme de faute de contrôle quand l'état interne est s. $\Pi(X_n) = (\pi(s_n),\pi_2(s_n))$.

La nature très similaire du contrôle et de la stabilisation est renforcée dans les équations ci-dessus comme le signal cible a été mis dans l'état interne du système. D'un point de vue théorique, la stabilisation et le contrôle sont très liés.

Notez que le "contrôle par apprentissage" n'a pas été séparé du contrôl adaptatif. Formellement, ces contrôles sont *équivalents*, la seule différence (purement quantitative) résidant dans la taille relative des espaces de fonctions et dans les poids relatifs des points récent et des points anciens.

Le cas particulier du contrôl dans lequel la cible est dynamiquement fournie par un humain a parfois été appelée **neurointerface**.

Notez que le cadre défini ci-dessus peut être appliqué pour un apprentissage non supervisé aussi, tel que l'ACP temporelle.

Dans certaines cas, on considère la stabilisation ou le contrôle de *systèmes multiples*. Alors on considère le pire cas, ou la moyenne, parmi la famille de systèmes (éventuellement munie d'une distribution de probabilité). Le choix optimal d'un signal de contrôle est appelé **trajectoire optimale**. La stabilité sous petites variations du système amène à une trajectoire dite **trajectoire optimale de voisinage**.

Une introduction générale au contrôl peut être trouvée dans [Stengel, 1994].

Ces trois problèmes, identification de système, contrôle, stabilisation, amènent à (mais ne se restreignent pas à) des problèmes mathématiques mettant en jeu des exemples non indépendants identiquement distribués. Des solutions à ce problème sont résumées en section 5.3.1. D'autres problèmes apparaissent en contrôle et en stabilisation: il s'agit de transformer l'identification de système en contrôle ou stabilisation.

5.2.2 Bref historique et état de l'art

Les historiques et états de l'arts suivant sont inspirés de [RayChaudhuri et al, 1995, Harthong, 2000] (partie 5.2.2), [Viana, 2001] (partie 5.2.2), [Amzallag et al, 1978, Van der Vaart et al, 1996] (partie 5.2.2), [Baueri, 1996] (partie 5.2.2), [Spears et al, 1993] (partie 5.2.2), [Rosenthal, 2001] (partie 5.2.2). Les résultats mathématiques sur les exemples non-iid sont seulement cités et complètement décrits en section 5.3.1.

Théorie du contrôle et physique statistique

L'origine de la physique statistique peut être attribuée aux idées philosophiques de Démocrite et Epicure affirmant la nature corpusculaire des choses. Bien sûr, un cadre mathématique général basé sur des conséquences quantitatives de ces faits, ou plus généralement des applications des statistiques en physiques, apparait plus tard, disons quand Daniel Bernoulli écrit "hydrodynamica" (1738), et s'impose réellement dans les années 1850. Le travail de Boltzmann dans les années qui suivent être maintenant fameux, comme le travail de Clausius ou Maxwell. A. Blanc-Lapierre au 20^e siècle, est considéré comme le pionier dans le domaine des applications de la physique statistique en traitement du signal: ceci amène en particulier à la théorie du contrôle, qui est une vaste aire de recherche incluant l'informatique aussi bien que les mathématiques et la physique. L'idée général est l'utilisation de mesures d'une quantité afin de controler les valeurs futures, par exemple de manière à stabiliser. [RayChaudhuri et al, 1995] résume ce domaine, et nous suivons leur texte dans la suite de ctte section. Les premiers controleurs de l'histoire remontent à une horloge à eau de Ktesibios et une lampe à huile de Philon, basée sur des régulateurs à liquides, aussi anciens que -300. Un livre écrit pendant le premier siècle par Hero d'Alexandrie, "Pneumotica", explique les principes des régulateurs basés sur l'eau.

En Europe moderne, la régulation de température a été traitée par C. Drebbel (1572-1633), et James Watt utilise un régulateur pour son moteur à vapeur. Après un grand développement durant la seconde guerre mondiale en raison des applications militaires, le contrôle est maintenant souvent basé sur la logique floue ou les méthodes neuronales, ou des mélanges de ces deux techniques. Ces deux derniers outils sont sujets à controverse, puisque certains auteurs considères ces deux outils comme révolutionnaires alors que d'autres considèrent qu'il n'y a rien de vraiment neuf là-dedans. Notre point de vue est que ces deux outils sont simple d'autres outils de régression/prédiction/classification, avec, dans le cas des réseaux neuronaux, des avantages algorithmiques pour des grands espaces de fonctions, et, dans le cas de la logique floue, la possibilité aisée de prendre en compte les opinions d'experts. En outre, des bornes non-asymptotiques indépendantes de la distribution souvent associées aux réseaux de neurones sont utilisées plus bas comme une important contribution de la théorie du processus empirique.

Les filtres de Kalman sont un outil très classique en identification de système. La description ci-dessous est principalement basée sur [Moonen et al, 1998], [Tomasi, 1997]. Considérer que les y_k sont la suite de sortie, u_k le signal d'entrée, x_k le vecteur d'état, v_k et w_k des bruits (supposés gaussiens indépendants pour certaines dérivations). Le modèle est comme suit (les notations ne sont pas consistantes avec celles de la partie 5.2.1 à cause d'un conflit entre les notations usuelles en contrôle et en filtres de Kalman):

$$\begin{aligned} x_{k+1} &= A_k x_k + B_k u_k + v_k \\ y_k &= C_k x_k + D_k u_k + w_k \end{aligned}$$

Les matrices A_k,B_k,C_k,D_k sont supposées connues. Le problème est l'évaluation des états internes. Les équations suivantes, connues comme les équations conventionnelles du filtre de Kalman, sont une solution possible et classique (avec M^T transposée de la matrice M):

$$\begin{aligned} P_{k|k} &= P_{k|k-1} - P_{k|k-1}C_k^T(W_k + C_k P_{k|k-1} C_k^T)^{-1} C_k P_{k|k-1} \\ x_{k|k} &= x_{k|k-1} + P_{k|k} C_k^T W_k^{-1}(y_k - C_k x_{k|k-1} - D_k u_k) \\ P_{k+1|k} &= A_k P_{k|k} A_k^T + V_k \\ x_{k+1|k} &= A_k x_{k|k} + B_k u_k \end{aligned}$$

$P_{k|k-1}$ est un estimateur de $E(x_{k|k-1} - \overline{x}_{k|k-1}) \times (x_{k|k-1} - \overline{x}_{k|k-1})$ (supposé donné pour $k = 0$) et $P_{k|k}$ est un estimateur de $E(x_{k|k} - \overline{x}_{k|k}) \times (x_{k|k} - \overline{x}_{k|k})$, V_k variance de v_k et W_k variance de w_k. $x_{k+1|k}$ est un estimateur de x_{k+1} à l'étape k (supposé donné pour $k = -1$) et $x_{k|k}$ est un estimateur corrigé après l'observation y_k.

Les deux premières équations sont connues comme équations de **mise-à-jour** (elles modifient les prédictions de $x_{k|k-1}$ et $P_{k|k-1}$ de manière à inclure l'information y_k, définissant alors $x_{k|k}$ et $P_{k|k}$). Les deux suivantes sont les équations de **propagation**. Elles sont la réelle partie prédictive; à l'étape k, elles estiment $x_{k+1|k}$, prédiction de x_{k+1} à l'étape k (avec son estimateur de variance).

Comme expliqué dans [Michel, 1998], de nombreux modèles ont été prouvés efficaces (et parfois optimaux dans un sens donné sous certaines hypothèses) dans le cas de processus linéaires (modèles Arma), mais ces méthodes ne peuvent pas gérer des comportements complexes comme ceux de systèmes chaotiques.

Richard Bellman (1920-1984) a été le premier à voir que l'équation de Hamilton-Jacobi (de mécaniques classique) peut être appliqué en contrôle optimal; ceci fournit la célèbre équation de Hamilton-Jacobi-Bellman, résolvant analytiquement le problème du contrôle optimal pour des systèmes à temps continu pour des intervalles de temps fixés (voir [Stengel, 1994, p. 219].

Une forme particulièrement répandue de contrôleur est le **contrôleur PID**: un signal de contrôle est défini proportionnel à des multiples de l'erreur (définie comme étant la différence entre la sortie et la sortie souhaitée), de l'erreur intégrée sur un interval de temps donné, de la dérivée de l'erreur. Formellement:

$$u = K_p \times e + K_i \times \int e dt + K_d \times \frac{de}{dt}$$

K_d est appelé **gain dérivatif**, K_i **gain intégral**, K_p **gain proportionnel**.

Dynamiques stochastiques dans des systèmes déterministes

L'idée de dynamiques compliquées apparaissant spontanément dans des systèmes naturels provient de Landau-Lifschitz. Dans les années 60, Smale découvre que les flots réguliers et des transformations régulières peuvent entrainer une infinité de mouvements périodiques. Ces mouvement pouvaient provenir de perturbations arbitrairement petites. Smale a alors introduit la notion d'**hyperbolicité** ([Smale, 1967]). Dans le début des années 70, Ruelle-Takens a développé l'idée dans l'esprit de la présence d'attracteurs **"étranges"** dans l'espace des états. La notion d'hyperbolicité a été développée par un grand nombre de chercheurs dans le cadre de la théorie ergodique: [Sinai, 1972, Bowen et al, 1975, Ruelle, 1976]. [Mané, 1988, Hayashi, 1997] ont montré l'importance de l'hyperbolicité pour la **stabilité structurelle** (un système étant structurellement

stable si ses orbites sont en bijection avec les orbites des systèmes proches). D'un autre côté, [Newhouse, 1979] montre que dans de nombreux cas, les systèmes dynamiques étaient non hyperboliques - [Viana, 2001] fournit une liste d'exemples ([Lorentz, 1963, Henon, 1976, Coullet et al, 1978, Feigenbaum, 1978]). Lorenz a souligné l'importance de la sensibilité aux conditions initiales. Ceci a amené de nombreux chercheurs à étudier des systèmes *faiblement* hyperboliques.

Le théorème 1 donne des conditions générales sous lesquelles des systèmes déterministes se comportent "presque" comme des systèmes aléatoires.

Théorème 5.1 *On suppose qu'au moins l'un des ensembles d'hypothèses suivants est vérifié:*

1. *Applications uniformément étendantes.*
   - $g : M \to M$ *est* $C^{1+\mu_0}$ *pour un certain* $\mu_0 \in]0,1]$.
   - M *est une variété compacte connexe.*
   - g *est une application étendante, au sens où il existe* $\sigma > 1$ *tel que* $\| Dg(x).v \| \geq \sigma \| v \|$ *pour tout* x *et* v.
2. *Attracteurs uniformément hyperboliques.*
   - $g : M \to M$ *est un difféomorphisme sur la variété* M *et* $Q \subset M$ *est un certain ensemble ouvert positivement invariant, au sens où* $g(fermeture(Q)) \subset Q$.
   - $\mathcal{L} = \cap_{n \in \mathbb{N}} g^n(Q)$ *est* **transitif** *(ie contient des orbites denses) et* **hyperbolique** *pour* g*, ie il existe une séparation du faisceau tangent de* M $T_{\mathcal{L}}M = E^s_{\mathcal{L}} \oplus E^u_{\mathcal{L}}$ *et un certain* $\lambda_0 < 1$ *tel que:*
     - $Df(x)E^s_x = E^s_{g(x)}$ *et* $Dg^{-1}(x)E^u_x = E^u_{g^{-1}(x)}$.
     - $\| Dg(x)E^s_x \| \leq \lambda_0$ *et* $\| Dg^{-1}(x)|E^u_x \| \leq \lambda_0$ *pour tout* $x \in \mathcal{L}$.

Alors, on a:

1. *Il existe une unique* **mesure SRB** μ *de support sur* $\mathcal{L}$*. Cette mesure est* ergodique*, et son bassin a une mesure de Lebesgue* > 0*. Le fait que* μ *soit SRB (pour Sinai-Ruelle-Bowen) signifie que* μ *est* **invariant** *(ie* $\int f(t,P)d\mu(t) = \mu(P)$*) et il y a un ensemble de mesure positive* x *tel que pour tout* ϕ *continu, et* x *dans le bassin*

$$\int \phi d\mu = \lim_{n \to \infty} \frac{1}{n} \sum_{i=0}^{n-1} \phi \circ f^i(x)$$

2. *La chaîne est* **exponentiellement mixant** *et satisfait le* **théorème central limite** *dans l'espace de Banach des fonctions continues* μ*-Hölder, pour tout* $\mu \in]0,\mu_1]$.
   Le caractère "exponentiellement mixant" signifie qu'il existe $r < 1$ *tel que pour tout couple de fonctions* (ϕ,ψ) μ*-Hölder, il existe* C *tel que* $E((\phi \circ f^n(t,.) - E\phi \circ f^n(t,.)) \times (\psi \circ f^n(t,.) - E\psi \circ f^n(t,.))) \leq Cr^n$ *(décroissante exponentielle des corrélations). Le théorème central limite signifie que pour tout* ϕ μ*-Hölder, il y a* σ *tel que pour tout interval* A

$$\mu(\{x | \frac{1}{\sqrt{n}} \sum_{i=0}^{n-1} (\phi \circ f^j(x) - \int \phi d\mu) \in A\}) \to \frac{1}{\sqrt{2\pi\sigma}} \int_A \exp(-\frac{-t^2}{2\sigma^2})$$

3. *La chaîne est* **stochastiquement stable sous de petites perturbations aléatoires**. *Voir [Viana, 2001] pour plus d'informations sur cette notion.*

Ce théorème a été prouvé ailleurs sous diverses formes mais nous utilisons la référence [Viana, 2001] en tant que joli état de l'art de résultats, incluant de nombreuses améliorations dans un cadre commun. En particulier, les applications non-uniformément hyperboliques sont traitées dans [Viana, 2001, section 5].

Une classe importante (mais non-exhaustive) de systèmes déterministes est la classe des **systèmes chaotiques**. De nombreuses définitions du chaos existent. Une définition classique est la suivante ([Devaney, 1992, Holmgren, 1996]):

Définition 5.2 (chaos) *Soit X_n un système dynamique défini par $X_{n+1} = f(X_n)$ avec $X_0 \in D$ et $f \in D^D$. Il est dit* **chaotique** *si*

1. *Les points périodiques de f sont denses dans D (un point est dit* **périodique** *de période k si $\underbrace{f \circ f \circ f \ldots f}_{k \text{ times}}(x) = x$).*
2. *f est* **topologiquement transitif**. *Ceci signifie que pour tous ensembles ouverts U et V qui intersectent D, il y a $x \in U \cap D$ et un nombre entier n tel que $f^n(z)$ est dans V. Ceci est équivalent au fait que pour tout x et y dans D et $\epsilon > 0$, il y a $z \in D$ tel que $d(x,z) < \epsilon$, $d(\underbrace{f \circ f \circ f \circ \ldots f(z)}_{n \text{ times}}, y) < \epsilon$ pour un certain n.*
3. *f présente une* **dépendance sensible aux conditions initiales**. *Ceci signifie qu'il existe $\delta > 0$ tel que pour tout $x \in D$ et $\epsilon > 0$, il y a un $y \in D$ et un $n \in \mathbb{N}$ tel que $d(x,y) < \epsilon$ et*

$$d(\underbrace{f \circ f \circ \ldots f}_{n \text{ times}}(x), \underbrace{f \circ f \circ \ldots f}_{n \text{ times}}(y)) > \delta$$

Une classe particulière de systèmes chaotiques est issue d'équations différentielles grâce au fameux théorème de Takens. Considérons les équations différentielles suivantes:

$$\begin{aligned} \frac{dY}{dt} &= G(Y(t)) \text{ (évolution du système)} \\ x(t) &= H(Y(t)) + \epsilon(t) \text{ (mesure avec bruit indépendant } \epsilon) \\ x_n &= x(nT) \text{ (discrétisation)} \\ X_n &= (x_n, x_{n-\tau}, x_{n-2\tau}, \ldots, x_{n-(d-1)\tau}) \text{ (fenêtre)} \end{aligned}$$

Selon le théorème de Takens (voir [Takens, 1981] pour un énoncé précis), sous des hypothèses légères sur G et H, si $d \geq 2D + 1$, avec D la dimension de l'attracteur du système, alors il y a un difféomorphisme qui associe X_n et $Y(nT)$. Ceci implique, en particulier, le fait que $x_{n+1} = f(X_n)$ pour un certain f.

La prédiction de séries temporelles chaotiques est un exercice classique pour les algorithmes de prédiction. Dans le cas général, les systèmes chaotiques sont beaucoup trop compliqués pour des prédictions basées sur Arma, d'où le besoin d'algorithmes intensifs comme les réseaux de neurones, les plus proches voisins ou les réseaux RBF (RBF:fonction à base radiale). Voir [Michel, 1998] pour un résumé des points importants de la prédiction de systèmes chaotiques.

Processus empirique et VC-théorie

Les nombres de couverture et la VC-dimension, dont les définitions sont rappelées ci-dessous, sont bien connus en théorie statistique de l'apprentissage, où ils ont montré beaucoup de bornes non-asymptotique. [Vidyasagar, 1997] fournit des états de l'art clairs de tels résultats. Il est un peu intéressant de noter que les résultats de la VC-théorie, qui sont une partie de la théorie du processus empirique, sont bien connus de la communauté de la théorie de l'apprentissage, alors que la théorie du processus empirique reste typiquement étudié par des mathématiciens. [Van der Vaart et al, 1996] fournit un état de l'art utile de résultats hors VC-théorie.

Une suite de variables aléatoires X_i converge **faiblement** vers X si pour tout f continu et borné $E(f(X_n)) \to E(f(X))$. Une suite de variables aléatoires X_n est $O(u_n)$ **faiblement** si $\forall \epsilon \exists K /$ lim inf $P(|X_n| \leq K u_n) \geq 1 - \epsilon$. Un **processus stochastique** indexé par une famille T est une famille de variables aléatoires à valeurs dans $\mathbb{R}$ indexée par T et dépendant d'un même univers.

$\mathcal{F}_{B,k,d}$, avec $k > 0$ réel, est la famille des ensembles $\{(x,t)/f(x) < t\}$ pour f de $[0,1]^{d-1}$ dans $[0,1]$ tel que $|||f|||_k$ est bien défini et borné par B, avec

$$|||f|||_k = \max_{\sum k_i \leq \lfloor \alpha \rfloor} \sup_x |\frac{\partial^{\sum k_i} f(x)}{\partial x_1^{k_1} \partial x_2^{k_2} \ldots \partial x_d^{k_d}}| \tag{5.3}$$

$$+ \max_{\sum k_i = \lfloor \alpha \rfloor} \sup_{x \neq y} \frac{|\frac{\partial^{\sum k_i} f(x)}{\partial x_1^{k_1} \partial x_2^{k_2} \ldots \partial x_d^{k_d}} - \frac{\partial^{\sum k_i} f(y)}{\partial x_1^{k_1} \partial x_2^{k_2} \ldots \partial x_d^{k_d}}|}{|x-y|^{\alpha - \lfloor \alpha \rfloor}} \tag{5.4}$$

avec $\lfloor \alpha \rfloor$ le plus grand entier strictement plus petit que α (NB: égal à $\alpha - 1$ si α est entier). Cet espace de fonctions est appelé un espace de **Hölder**.

La notation $E(X\{F\})$ avec F un évènement et X une variable aléatoire note l'espérance de $X \times \chi_F$, avec χ_F la fonction caractéristique de F égale à 1 si F a lieu et 0 sinon. Ceci sera utile pour noter la condition de Lindeberg.

Avec $\mathcal{F}$ une famille de fonctions, et X une variable aléatoire, $\| X \|_{\mathcal{F}}$ est $\sup_{f \in F} |E(f(X))|$. Ceci peut être étendu aux mesures non-de probabilité.

Définition 5.3 (Réseaux, nombres de couverture & bracketing nombres de couverture) *Etant donné un espace F de fonctions, un ϵ-***réseau** *de F est un ensemble fini de fonctions $\{f_1, f_2, \ldots, f_N\}$ (pas nécessairement inclus dans F) tel que pour tout $f \in F$, il existe i tel que $\int |f - f_i| \leq \epsilon$.*

Quand N est minimal, il est appelé le **nombre de couverture** *$N(F,\epsilon,d)$ de F.*

*De même, pour F un espace de fonctions, un ϵ-***bracket** *$[l,u]$ avec l et u tels que $\int u - l \leq \epsilon$ est l'ensemble des $f \in F$ tel que $l \leq f \leq u$. L'ϵ-***bracketing nombre de** *F est le nombre minimal $N_{[\,]}(F,\epsilon,d)$ d'ϵ-brackets $[l_i,u_i]$ tels que $F \subset \cup_i [l_i,u_i]$.*

*L'***entropie** *ou la* **bracketing-entropie** *sont les logarithmes des nombres de couvertures ou bracketing-nombres de couverture respectivement.*

Définition 5.4 (VC-dimension) *Considérons C une classe de fonctions à valeurs dans $\{0,1\}$. Définissons $\Delta_n = \sup_{(x_1,\ldots,x_n) \in X^n} card\{c^{-1}(1) \cap \{x_1, \ldots, x_n\} | c \in C\}$ le n^e coefficient de pulvérisation $S(C,n)$ de C, et la* **VC-dimension** *de C est $VC(C) = \sup\{m | D_m = 2^m\}$.*

Le lemme de Sauer justifie l'intérêt de la VC-dimension:

Lemme 5.5 (le lemme de Sauer)

$$S(C,n) \leq \sum_{i=0}^{VC(C)} C(n,i) \leq (n+1)^{VC(C)}$$

Ceci est une borne sur les coefficients de pulvérisation, et peut être utilisé pour borner les nombres de couvertures (les nombres de couverture sont polynomiaux quand la VC-dimension est finie). Ci-dessous un ensemble de bornes usuelles en théorie de l'apprentissage:

$$P \leq 8(m+1)^V \exp(-\frac{1}{32} m\epsilon^2) \tag{5.5}$$

$$P \leq 4 \exp(4\epsilon + \epsilon^2 - 2m\epsilon^2) \times m^{2V} \text{ si } V \geq 2 \text{ et } m \geq 4V \tag{5.6}$$

$$P \leq 4 \exp(4\epsilon + \epsilon^2 - 2m\epsilon^2) \times (m+1)^2 \text{ si } V = 1 \tag{5.7}$$

$$P \leq 8m^V \exp(-\frac{1}{32} m\epsilon^2) \text{ si } V \geq 3 \text{ et } m \geq 2V \tag{5.8}$$

$$P \leq 8(m+1)^V \exp(-\frac{1}{32} m\epsilon^2) \tag{5.9}$$

$$P \leq 4 \exp(4\epsilon + \epsilon^2 - 2m\epsilon^2) \times \exp(m \times \mathcal{H}(2 \times V/m)) \text{ si } V \geq 2 \text{ et } m > 4V \tag{5.10}$$

$$P \leq 8 \exp(m \times \mathcal{H}(V/m)) \exp(-\frac{1}{32} m\epsilon^2) \text{ si } V \geq 3 \text{ et } m > 2V \tag{5.11}$$

$\mathcal{H}(x) = -x \log(x) - (1-x) \log(1-x)$, est une fonction entropie classique, définie par continuité en 0 et 1. Les deux premiers résultats proviennent de [Devroye, 1982], les deux suivants de [Vapnik, Chervonenkis, 1971].

Les derniers sont des optimisations détaillées dans [Devroye, 1996]. On peut remarquer qu'un résultat de [Alexander, 1984] donne une meilleure borne dans le cadre d'un très grand nombre d'exemples. Dans la vie de tous les jours, les formules 5.5 et 5.6 (selon le nombre m d'exemples) ou leurs équivalents moins-connus et moins simples 5.11 et 5.10 sont les plus pratiques.

Ces éqautions peuvent être réécrites en tant que bornes sur la précision ϵ, dépendant du risque δ et du nombre d'exemple m. Par exemple, l'équation 5.5 amène à

$$\epsilon \leq \sqrt{\frac{32 \log(8(m+1)^V/\delta)}{m}} \tag{5.12}$$

Une autre forme classique est la borne en **complexité d'échantillon**: on peut utiliser les équations ci-dessus pour calculer m tel que la précision est bornée par ϵ avec probabilité au moins $1-\delta$. Ceci donne une jolie formulation; on peut donner un nombre d'exemples suffisant pour apprendre dans une famille donnée, avec la précision et le risque requis.

On aura besoin dans la suite d'une autre forme de nombre de couverture ou de bracketing nombre de couverture. [Van der Vaart et al, 1996] fournit des résultats basés sur ces nombres, donnant comme référence [Koul, 1970, Shorack, 1973, Shorack, 1979, Van Zujilen, 1978, Shorack et al, 1986, Marcus et al, 1984, Shorack et al, 1986].

Définition 5.6 (Entropie "spéciale") *Etant donné $Z_{n,i}$ des variables aléatoires pour $1 \leq i \leq n$, $N'_{[\ ]}(\epsilon,\mathcal{F},n)$ est le cardinal (dépendant de n) de la plus petite famille (si elle est finie) de $(F_i)_{i\in I}$ tel que $\forall f \in \mathcal{F} \exists i \in I / f \in F_i$ et $\forall i / \sum_{i=1}^n E(\sup_{f,g\in F_i^2}(Z_{n,i}(f) - Z_{n,i}(g))^2) \leq \epsilon^2$.*

L'inégalité de Hoeffding, prouvée par Hoeffding en 1963 [Hoeffding, 1963], généralise la borne de Chernoff sous leur forme additive. Elle est bien connue en théorie de l'apprentissage. Notez qu'elle n'implique pas la forme multiplicative des bornes de Chernoff.

Théorème 5.7 (L'inégalité de Hoeffding) *Soient $X_1,\ldots,X_n$ des variables aléatoires indépendantes respectivement dans $[a_1,b_1],\ldots,[a_n,b_n]$.*

$$P(\frac{1}{n}\sum(X_i - EX_i) \geq \epsilon) \leq \exp\left(-\frac{2n\epsilon^2}{\frac{1}{n}\sum|a_i-b_i|^2}\right) \tag{5.13}$$

$$P(\frac{1}{n}\sum(X_i - EX_i) \leq -\epsilon) \leq \exp\left(-\frac{2n\epsilon^2}{\frac{1}{n}\sum|a_i-b_i|^2}\right)$$

$$P(\frac{1}{n}\sum|X_i - EX_i| \geq \epsilon) \leq 2\exp\left(-\frac{2n\epsilon^2}{\frac{1}{n}\sum|a_i-b_i|^2}\right) \tag{5.14}$$

Des théorèmes centraux limite généralisés (avec entropie uniforme) sont dus à [Dudley, 1978, Pollard, 1982, Kolcinski, 1981], et les équivalents bracketings sont dus à [Dudley, 1978, Dudley, 1984, Ossiander, 1987, Andersen et al, 1988]. Un état de l'art général de tels résultats peut être trouvé dans [Van der Vaart et al, 1996]. Le résultat suivant sera utile avec $Z_{n,i}(f) = f(X_{n,i})/\sqrt{n}$ principalement.

Théorème 5.8 (Convergence rapide) *Pour tout $n \in \mathbb{N}$, on considère $Z_{n,i}$, pour $i \in [1,n]$, des variables aléatoires indépendantes. On suppose que l'hypothèse suivante a lieu:*

$$\sum_{i=1}^n E \parallel Z_{n,i} \parallel_{\mathcal{F}}^2 \{\parallel Z_{n,i} \parallel_{\mathcal{F}} > \eta\} \to 0 \text{ pour tout } \eta > 0 \tag{5.15}$$

$$\sup_{(f,g)\in\mathcal{F},d(f,g)<\delta_n} \sum_{i=1}^n E(Z_{n,i}(f) - Z_{n,i}(g))^2 \to 0$$
$$\text{pour tout } \delta_n \text{ décroissant vers } 0 \tag{5.16}$$

$$\int_0^{\delta_n} \sqrt{\log N'_{[\ ]}(\epsilon,\mathcal{F},n)}d\epsilon \to 0$$
$$\textit{pour tout } \delta_n \textit{ décroissant vers } 0 \tag{5.17}$$

Alors $E_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n Z_{n,i} - E(Z_{n,i})$ *est asymptotiquement tight dans l'espace des fonctions totalement bornées de* $\mathcal{F}$ *dans* $\mathbb{R}$. *Ceci signifie que pour tout* $\epsilon > 0$ *il existe un ensemble compact* K *tel que* $\liminf P(E_n \in K^\delta) \geq 1 - \epsilon$ *pour tout* $\delta > 0$, *avec* $K^\delta = \{y/d(y,K) < \delta\}$ *le* δ*-agrandissment de* K.

Il converge en distribution, pourvu qu'il converge pour les lois marginales. Alors, le processus limite T *est centré (moyenne zéro), gaussien, avec covariance* $E(TfTg) = E(Tfg) - E(Tf) \times E(Tg)$.

La condition 5.16 peut être supprimée si la partition dans $N'_{[\ ]}$ *peut être choisie indépendamment de* n.

La finitude des bracketing-nombres sera suffisante pour un théorème plus faible.

Théorème 5.9 (Convergence lente) *On suppose que les* $N_{[\ ]}(\epsilon,\mathcal{F})$ *sont finis pour tout* ϵ, *et que tous les* f_i *et* g_i*'s can be chosen in* $\mathcal{F}'$, *such that almost sure convergence of* $\frac{1}{\sqrt{n}}E_n(f)$ *toward* 0 *is true for any* $f \in \mathcal{F}'$. *Then* $\frac{1}{\sqrt{n}}E_n(f)$ *converges towards* 0, *almost surely, uniformly in* $f \in \mathcal{F}$.

Le résultat suivant est adapté de [Vidyasagar, 1997, p188].

Théorème 5.10 (Bornes non-asymptotiques) *Soit* $F_{\epsilon_0/2}$ *une* $\frac{1}{2}\epsilon_0$*-couverture de* F'. *Alors, l'algorithme de minimisation du risque empirique appliqué à* $F_{\epsilon_0/2}$ *est PAC avec précision* $\epsilon > \epsilon_0$ *et confiance* $1 - \delta$ *dans* F, *pourvu que* n *(nombre d'exemples) est choisi plus grand que* $\frac{8}{\epsilon^2}\ln(\frac{Card\ F_\epsilon}{\delta})$. *En outre, avec confiance au moins* $1 - Card\ F_{\epsilon_0/2} exp(-n\epsilon^2)$, *la différence entre l'erreur en généralisation et l'erreur empirique, pour tout algorithme choisissant une fonction dans* $F_{\epsilon_0/2}$, *est bornée par* $e_0/2$.

Les résultats suivants justifient l'intérêt des théorèmes ci-dessus. Ils fournissent des classes pour lesquels les hypothèses sont vérifiées sous des conditions raisonnables.

Les nombres de couverture d'ensembles convexes ont été étudiés dans [Bronstein, 1976, Dudley, 1974]. Ici on suppose $d \geq 2$. Le cas $d = 1$ n'est pas intéressant. Une preuve du résultat qui suit peut être trouvée dans [Van der Vaart et al, 1996, p163]. $\mathcal{F}$ est la famille des convexes de $[0,1]^d$.

Théorème 5.11

$$\log N_{[\ ]}(\epsilon,\mathcal{F}) \leq K(\frac{1}{\epsilon})^{(d-1)}$$

Les espaces de fonctions régulières ont été pour la première fois étudiées, du point de vue du processus empirique, dans [Lorentz, 1966, Birman et al, 1967, Dudley, 1984]. On considère les sous-graphes d'applications de $[0,1]^{d-1}$ dans $[0,1]$, avec norme Hölderienne bornée. Une preuve du théorème qui suit peut être trouvée dans [Kolmogorov et al, 1961] ou dans [Van der Vaart et al, 1996, p157].

Théorème 5.12 $\mathcal{F}_{B,\alpha,d}$ *vérifie*

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

Ce résultat est connu optimal en termes de puissances de $\frac{1}{\epsilon}$.

D'autres belles propriétés des classes de Donsker (qui ne sont pas nécessairement vraies pour des VC-classes) sont la stabilité: la *fermeture point-à-point* d'une classe de Donsker est Donsker (pour la même distribution), l'*enveloppe symmétrique convexe* d'une classe de Donsker est Donsker (pour la même distribution). Notez que la notion de classe de Donsker est dépendante de la distribution.

Logique floue

La logique floue a émergie en contrôle appliqué, principalement au Japon et maintenant en Europe et aux Etats-Unis. Cela peut être brièvement résumé comme un système expert basé sur l'interpolation linéaire d'experts sur des ensembles "flous" (un ensemble par règle). La première référence concernant la logique floue est [Zadeh, 1965].

Alors qu'usuellement on considère des sous-ensembles de X comme éléments de $\{0,1\}^X$, la logique floue considère les ensembles flous comme éléments de $[0,1]^X$. L'ensemble flou associé à l'ensemble Y est l'application $x \mapsto \inf_{y \in Y} \max(0{,}1 - \lambda \parallel x - y \parallel)$. Alors, $A \cap B$ est (comme en logique classique) le min des fonctions associées à A et B, et $A \cup B$ et le max, et la négation de A est la fonction $1 - f_A$, avec f_A associé à l'ensemble flou A.

Si un système expert propose une règle quand "la situation" (l'état) est dans l'ensemble A et une règle quand la situation est dans l'ensemble B, alors ces règles peuvent être transformées en des règles floues $A \Rightarrow A'$ et $B \Rightarrow B'$. Alors, pour une nouvelle situation x, on peut utiliser une combinaison des règes floues, en utilisation une combinaison de ces règles basée sur les valeurs des fonctions caractéristiques floues. Ceci fournit une réponse floue, la plus simple solution pour prédire consistant à utiliser le centre de gravité de l'ensemble flou. Formellement, la réponse floue résultante pour une situation x pour les règles $A \Rightarrow A'$ et $B \Rightarrow B'$ est $\frac{A(x) \times A' + B(x) \times B'}{A(x) + B(x)}$ (en identifiant les ensembles et leurs fonctions caractéristiques).

Réseaux neuronaux, Support Vector Machines et outils d'apprentissage liés

Cette section est partiellement basée sur [Thiria et al, 1997]. En 1943, W. McCulloch et W. Pitts proposent un modèle formel de neurone. Hebb propose une règle pour l'évolution synaptique. De grands espoirs étaient basés sur l'idée de réseaux de petites unités de calcul. En 1969, Minsky et Papert ont montré de fortes limitations pour le perceptron initial avec seulement un neurone, tel que ceux de Rosenblatt et Widrow. En 1986, Rumelhart, Hinton et Williams dans un travail révolutionnaire ont fondé le mécanisme de rétropropagation (une descente de gradient astucieuse sur l'erreur empirique), permettant l'évolution des poids synaptiques dans les réseaux à couches cachées de neurones. Des problèmes demeurent, tels que l'existence de minima locaux, mais d'autres outils comme les Support Vector Machines, qui ont l'avantage d'une fonction objectif strictement, sont en fait beaucoup plus lentes et ont principalement été prouvées efficaces pour des petits benchmarks.

Dans un cadre non supervisé, le plus connu probablement des réseaux est les cartes de Kohonen. Alors que dans le cadre supervisé (VC-dimension, puis fat-shattering dimension, voir [Bartlett, 1998] ou [Alon et al, 1997]), de nombreux résultats théoriques supportent la capacité des réseaux de neurones à généraliser, un manque de fondations théoriques apparait dans ce cas.

Comme beaucoup d'applications sont gratuitement disponibles pour la rétropropagation ou d'autres réseaux de neurones, nous ne résumons pas ces algorithmes. Le lecteur intéressé est renvoyé à [Bishop, 1995] pour les réseaux de neurones classiques et des tutorials sur www pour les support vector machines. Le cas particulier d'apprentissage "en ligne" est facilement géré en rétropropagation par l'algorithme "batch" classique car il est naturellement "en ligne". Dans le cas des Support Vector Machines, voir [Ralaivola, 2001] pour une adaptation online.

[Edwards et al, 1995] résume les arguments suivants en faveur des réseaux de neurones pour le contrôle:

- dans beaucoup de cas pratiques, probablement grâce à la dépendanc non-linéaire de la fonction résultante par rapport aux paramètres, et bien que des contre-exemples peuvent être donnés, les réseaux de neurones entraînent un beaucoup plus petit nombre de paramètre que d'autres outils tels les RBFs ou les splines. Cet avantage est probablement encore vrai par rapport aux SVM.
- La sortie de la fonction résultante peut aisément être bornée, comme souvent requis en contrôle, comme par exemple dans le cas du problème du camion-remorque de la section 5.4.5.
- Les solutions bang-bang et les discontinuités peuvent être gérées.

Des résultats de praticiens quant à ces affirmations seraient intéressants.

Algorithmes génétiques

Les algorithmes génétiques sont un paradigme usuel pour l'optimisation de fonctionnelles. Dans le cas de l'apprentissage, des solutions analytiques telles que les descentes de gradient sont souvent préférées (comme en rétropropagation ou avec des Support Vector Machines), mais pour des applications dans lesquelles des couples entrées/sorties ne peuvent être aisément fournies comme en contrôle, ces approches échouent et les algorithmes génétiques (ainsi que le recuit simulé, qui peut être vu comme un cas particulier) apparait comme une solution naturelle.

Le principe général consiste en:

1. Sélectionner un ensemble initial de controleurs (en utilisant des experts ou une initialisation aléatoire). Cet ensemble est appelé "population".
2. Jusqu'à un critère d'arrêt donné:
   (a) Evaluer la "qualité" de chaque élément de la population
   (b) Sélectionner les meilleurs éléments de la population
   (c) Créer de nouveaux éléments en combinants les éléments sélectionnés
   (d) Opérer des mutations sur les nouveaux éléments (ou sur tous les éléments)
   (e) Remplacer l'ancienne population par les éléments nouveaux et les éléments sélectionnés

De telles approches permettent en particulier d'utiliser des controleurs symboliques ou des controleurs neuronaux. Des exemples d'algorithmes génétiques appliqués en contrôle sont [Schoenauer et al, 1994, Koza, 1992, Chellapilla, 1998]. Les algorithmes génétiques peuvent être utilisés pour optimiser les poids mais aussi l'architecture, en utilisant des codes génétiques pour les architectures ([1, 2, Eggenberger]).

Chaînes de Markov

Une chaîne de Markov est une suite de variables aléatoires X_n, définies par:

- Un état initial X_0, ou une distribution de probabilité pour X_0.
- Une probabilité de transition $P(X_n|X_{n-1})$, supposée constante pour $n \geq 1$. $P(X_n \in E|X_{n-1} = t) = f(t,E)$.

Des modèles de Markov de degré plus élevé peuvent être définis, avec des probabilités de transition $P(X_n|X_{n-1},X_{n-2},\ldots,X_{n-k})$, avec une distribution de probabilité pour $(X_0,X_1,\ldots,X_{k-1})$. On restreint notre attention au cas ci-dessus, qui peut inclure d'autres cas par simple adaptation. $f^n(t,E)$ est défini par induction par $f^1 = f$ et $f^{n+1}(t,E) = \int f^n(u,E)f(t,du)$.

Une aire de recherche importante à propos des chaînes de Markov est leur comportement asymptotique. De nombreux livres fournissent des résultats tels l'existence d'une distribution stationnaire et la convergence vers cette distribution, sous des hypothèses raisonnables, dans le cas d'espaces d'états finis. L'extension à des espaces d'état dénombrables existe, mais pour beaucoup d'applications, on a besoin de convergence rapide dans des chaines de Markov non dénombrables. [Doeblin, 1940] a prouvé un premier résultat dans cette direction, et [Rosenthal, 2001] survolle les résultats récents. Le résultat principal dans [Rosenthal, 2001] est donné ci-dessous:

Théorème 5.13 *On suppose que:*

- *f admet une* **distribution stationnaire** *π, ie une distribution π telle que*

$$\forall A \text{ mesurable } \pi(A) = \int f(y,A)d\pi(y)$$

- *f est* **apériodique**, *ie il n'existe pas une partition finie en $d \geq 2$ ensembles $\mathcal{X}_1,\ldots,\mathcal{X}_d$, telle que $\forall t \in X_i \ f(t,\mathcal{X}_{i+1}) = 1$, avec $\mathcal{X}_{d+1} = \mathcal{X}_1$.*
- *f est ϕ-***irréductible**, *ie $\exists \phi$ mesure non-trivialle telle que*

$$\forall A \text{ mesurable } \phi(A) > 0 \Rightarrow \exists n f^n(t,A) > 0 \text{ presque sûrement en } t \text{ pour } \pi$$

Alors, presque sûrement en t (pour π):

$$\lim_{n\to\infty} \sup_A |f(t,A) - \pi(A)| = 0 \tag{5.18}$$

On a besoin en fait, dans certaines applications ci-dessous, la convergence *uniforme* (en t), ou du moins des bornes explicites sur la dépendance en t. Ceci est géré dans le théorème (multiple) suivant ([Meyn, 1993] pour le premier, [Arcones, 1996, Theorem 4.1] pour le second):

Théorème 5.14 (Ergodicité uniforme) – *(convergence uniforme) (premier cas du théorème)*
Supposons qu'il existe $m \in \mathbb{N}$, μ une mesure de probabilité et $\delta > 0$ tel que

$$\forall t f^m(t,.) \geq \delta\mu \tag{5.19}$$

alors $\| f^n(t,.) - \pi \| \leq (1-\delta)^{\frac{n}{m}}$. La condition 5.19 est appelée la **condition de Doeblin**, *ou* ergodicité uniforme.

– *(un théorème central limit fonctionnel uniforme dans les classes de Donsker régulières pour les processus mixant) (deuxième cas du théorème)*
Supposons que

1. *les X_n sont un* **processus strictement stationnaire** *(ie la loi des X_n est indépendante de n).*
2. *$\mathcal{F}$ est un α-Hölder espace de fonctions sur X, sous-ensemble borné de $\mathbb{R}^d$.*
3. *Condition de mixage: avec $\alpha_k = \sup_{A,B\ measurable, l\geq 1}\{|Pr(AB) - Pr(A)Pr(B)|/A \in \sigma_1^l, B \in \sigma_{k+l}^{\infty}\}$, avec σ_a^b la σ-algèbre générée par $X_a,\ldots,X_{a+b-1}$, pour un $p > 2$, $\sum \alpha_n n^{2/(p-2)} < \infty$ et $\frac{d(p-1)}{p} < \alpha$.*

Alors

$$\{\frac{1}{\sqrt{n}} D_n f | f \in \mathcal{F}\} \to Df \ dans\ l^{\infty}(\mathcal{F})$$

avec D un processus gaussien centré avec covariances définies par

$$Dfg = Cov(f_1(X_1),f_2(X_{1+k})) + \sum_{k=1}^{\infty} Cov(f_1(X_1),f_2(X_{1+k})) + Cov(f_2(X_1),f_1(X_{1+k}))$$

Voir la partie B pour des résultats liés au second, et l'annexe A pour un résultat lié au premier dans le cas réversible. Notez que des comportements déterministes pour des applications uniformément dilatantes ou hyperboliques mènent à des meilleurs équivalent que ces chaînes de Markov.

Voir [Roberts et al, 1997] dans le cas réversible. [Meyn, 1994] fournit des résultats proches très intéressants, tels celui cité en appendice A: on peut avoir convergence géométrique avec dépendance en le point initial (dans le théorème ci-dessus, la première partie considère la convergence *uniforme en le point de départ*) et avec des bornes (presque ...) explicites sur les constantes.

5.3 Arrière-plan mathématique de l'apprentissage non-indépendant avec applications à l'identification de systèmes et au contrôle

5.3.1 Apprentissage d'exemples non-indépendants

Dans les applications ci-dessus, on nécessite un apprentissage avec des exemples qui ne sont pas iid. Dans l'esprit de la théorie de l'apprentissage, on demande l'uniformité de l'estimation de $L(f)$ pour $f \in \mathcal{F}$. La section 5.3.1 présente des résultats venus du processus empirique. La section 5.3.1 présente une adaptation de la VC-théorie dans le cas d'exemples distribués markoviennement.

Résultats asymptotiques: classes de Donsker pour des processus "ergodiques"

La présence de guillemets autour d' "ergodiques" est due au fait que nous travaillons sur différentes sortes de suites, dont certaines complètement déterministes, et nous ne demandons pas dans tous les cas une ergodicité *stricto sensu*. Comme expliqué ci-dessous, des dynamiques stochastiques peuvent apparaitre dans des systèmes complètement déterministes. Cette section est principalement basée sur [Viana, 2001] et [Arcones, 1996].

On considère dans cette partie des conditions sous lesquelles les exemples générés par un processus approximent une loi asymptotique. Ceci est fait dans deux étapes:

1. Les exemples X_n distribués par un système "ergodique" assurant la convergence des moyennes empiriques vers les espérances pour la distribution asymptotique, avec le même ordre de convergence que dans le théorème central limite. Ceci est vérifié par l'utilisation des théorèmes 1 (systèmes dynamiques déterministes) ou 14 (deuxième partie, chaînes de Markov stationnaires).
2. Généralisation en direction de l'uniformité. Ceci est fait grâce au théorème 15 (mettant en jeu des résultats sur le processus empirique).

On considère des X_n distribués par

1. un modèle de Markov qui mène à une probabilité conditionnelle $P(X_n|X_{n-1} = t) = p(t)$. Notez que des modèles de Markov de plus haut degré peuvent être utilisés de même.
2. une fonction déterministe g telle que $X_n = g(X_{n-1})$.

Le premier cas est plus facile sous certaines jolies hypothèses sur p. Le second cas requiert de très beaux résultats sur les systèmes dynamiques. On utilisera pour généraliser à l'uniformité le résultat suivant de [Arcones, 1996]:

Théorème 5.15 *Considérons $A_n(f)$ et $A(f)$ des processus stochastiques. Supposons que:*

1. $\sup_f |A_n(f)|$ *est presque sûrement fini pour tout n.*
2. $\sup_f |A(f)|$ *est fini presque sûrement.*
3. *Les distributions de dimension finie de $\{A_n(f) : f \in \mathcal{F}\}$ convergent vers celles de $\{A(f)|f \in \mathcal{F}\}$.*
4. *Pour tout entier positif q il existe une application $\pi_q : \mathcal{F} \mapsto \mathcal{F}$ telle que le cardinal de $\{\pi_q f : f \in \mathcal{F}\}$ est fini et pour tout η*

$$\lim_{q\to\infty} \lim_{n\to\infty} \sup Pr^*\{\sup_{f\in\mathcal{F}} |A_n(f) - A_n(\pi_q f)| \geq \eta\} = 0$$

Alors A_n converge faiblement vers A dans $l^\infty(\mathcal{F})$.

Ce résultat se trouve ailleurs dans la littérature. On utilise [Arcones, 1996] pour référence car beaucoup de résultats (nouveaux) liés à notre propos peuvent être trouvés dedans (voir appendice B), particulièrement dans le cas stationnaire. Le corollaire suivant sera utile par la suite:

Corollaire 5.16 *Considérons $A_n(f) = \frac{1}{\sqrt{n}} \sum_{i=1}^n (f(X_i) - Ef(X_i))$. Supposons que:*

1. *$\mathcal{F}$ est Donsker pour une mesure de probabilité π, avec $\pi fg - \pi f \pi g = Afg - AfAg$.*
2. $\sup_f |A_n(f)|$ *est fini presque sûrement.*
3. *Il existe des bracketing nombres de couverture pour tout ϵ,* avec *des brackets inclus dans $\mathcal{F}^2$.*
4. *Supposons que les distributions finies de $A_n(f)$ convergent vers celles de $A(f)$.*

Alors A_n converge faiblement vers A dans $l^\infty(\mathcal{F})$.

2. On a besoin de cette inclusion pour prouver la dernière hypothèse du théorème 15. Il est probable que cette hypothèse pourrait être réduite.

Proof:La condition 2 (aisément vérifiée dans beaucoup de cas pratiques) est la condition 1 du théorème 15. Le caractère Donsker de $\mathcal{F}$ assure que la condition 2 dans le théorème 15 est vérifiée, grâce au théorème 8. La condition 4 garantit la condition 3 du théorème 15. En utilisant π_q les projections successives sur les bornes supérieures dans les brackets garantit la condition 4 du théorème 15.□

Grâce à ce simple corollaire, on peut utiliser des résultats forts comme le théorème 8 pour prouver la finitude presque sûre de $\sup_f |Z_f|$, et alors utiliser Z_n simplement garantissant la convergence point à point, telle que les suites déterministes du théorème 1, ou les suites markoviennes avec théorèmes centraux limites (théorème 14, deuxième partie). Malheureusement, dans le dernier cas des suites markoviennes, de tels théorèmes sont seulement disponibles, pour autant que nous sachions, dans le cas de chaînes de Markov **réversibles** (ie pour tout A et B sous-ensembles de X $\int_{x\in A} P(x,B)d\pi(x) = \int_{y\in B} P(y,A)d\pi(y)$) ou des chaînes de Markov**stationnaires**. Cependant, des travaux sont en cours dans cette direction, selon [Roberts et al, 1997, remark 2.4]; on considère ici la stationarité.

[Viana, 2001] explique les mécanismes par lesquels des systèmes entièrement déterministes (pour autant que de tels systèmes existent dans un monde quantique ...). Dans le cadre déterministe, la probabilité est définie pour X_0, point initial du système dynamique. Le théorème 1 (voir [Viana, 2001]) justifie les comportements stochastiques observés en pratique.

Finalement, on peut résumer ces résultats (convergence de chaines de Markov ou dynamiques stochastiques dans des systèmes déterministes, plus uniformité de la convergence faible sous des conditions de bracketing-entropie) dans le théorème suivant:

Théorème 5.17 (Convergence uniforme pour des suites déterministes ou markoviennes) *Si l'un des faits suivants a lieu:*

- *les X_n sont distribués selon $X_{n+1} = g(X_n)$ et les hypothèses du théorème 1 sont vérifiées.*
- *les X_n sont distribués selon un modèle de Markov vérifiant les conditions du théorème 14 (second cas).*

Alors, une convergence en $O(1/\sqrt{n})$ a lieu uniformément en $f \in \mathcal{F}$. Formellement et plus précisément $D_n \to D$ faiblement dans $l^\infty(\mathcal{F})$, avec Df un processus gaussien centré comme défini en 1 (ie analogue déterministe du classique théorème central limite) ou 14 (second cas).

Bornes non-asymptotiques

Dans cette section, on rappelle que résultats classiques, et fournissant une nouvelle (pour autant que nous sachions) borne non-asymptotique basée sur la condition de Doeblin. Deux paradigmes d'apprentissage en sont déduits, l'un d'eux étant la minimisation du risque empirique et l'autre proche de la minimisation du risque empirique, et bénéficiant d'une borne un peu meilleure (nous prouvons seulement une meilleure borne mais ne prouvons pas une stricte supériorité). L'optimalité à des facteurs logarithmiques près est prouvée.

Espaces d'états finis ou dénombrables Pour autant que nous sachions, les résultats les plus généraux dans le cas d'espaces d'états finis ou dénombrables sont ceux de [Gamarnik, 1999]. Leur résultat principal est résumé comme suit: dans le cas d'espaces d'états finis (avec N états) et une distribution stationnaire uniforme, la complexité d'échantillon requise est bornée par $\frac{s}{1-\lambda_2}\ln(\frac{sN}{\delta})$, avec s la complexité d'échantillon dans le cas iid, δ le risque, λ_2 la seconde plus grande valeur propre de la matrice de transition. Des résultats liés peuvent être trouvés dans [Gamarnik, 1999].

Cas général On reformule le cadre PAC (Probablement Approximativement Correct) de Valiant de la façon suivante:

Définition 5.18 *Un algorithme d'apprentissage sur une famille F de fonctions et sur une suite markovienne avec probabilité conditionnelle $p(t) = U \mapsto P(X_n \in U|X_{n-1} = t)$ est PAC pour une complexité d'échantillon m avec précision ϵ et confiance $1-\delta$ si et seulement si pour tout $n \geq m$ et pour tout X_0 la probabilité d'une différence $> \epsilon$ entre l'erreur en généralisation pour la distribution asymptotique et l'erreur optimale dans F est $\leq \delta$ pour un échantillon de taille n. L'erreur en généralisation est définie comme l'erreur limite moyenne sur une suite Markovienne finie $X_0,\ldots,X_k$ (d'autres définitions sont possibles de même, prenant*

en compte la longueur de X_k; les résultats seraient similaires, à ceci près que k serait demandé plus grand qu'une quantité linéaire en $D(1-\ln(\delta))/\epsilon^2$ avec D défini plus loin).

Une famille F de fonctions est dite PAC s'il existe un algorithme choisissant $f \in F$ satisfaisant la condition ci-dessous pour m polynomial en ϵ, $\ln(1/\delta)$. L'algorithme est dit PAC.

La même définition vaut uniformément pour $p \in \mathcal{P}$ si la condition ci-dessus a lieu uniformément en $p \in \mathcal{P}$ (ie m doit être indépendant de $p \in \mathcal{P}$).

La nécessité de bornes non-asymptotiques apparait en contrôle adaptatif, ie quand l'environnement varie avec le temps (notez que ceci est peut-être moins fondamental en "contrôle par apprentissage" (voir section 5.2.1), comme les variations de l'environnement sont supposées rapides, ce qui implique que l'on utilise les passages précédents dans la même aire et donc qu'on peut utiliser des résultats asymptotiques). Alors, utiliser des milliers de points n'est plus possible, et le contrôle doit être dynamique.

On restreint notre attention à une suite $X_0,\ldots,X_n,\ldots$ de variables aléatoires, avec X_n dépendant seulement de X_{n-1}; en outre, $X_n|X_{n-1}$ est indépendant de n. Des chaînes de plus haut degré peuvent être considérées très similairement, simplement en considérant $Y_i = (X_i,X_{i+1},\ldots,X_{i+k})$. On considère par la suite la convergence uniforme de $\frac{1}{n}\sum_{i=0}^{n} f(X_i)$; uniformément en $f \in F$ et en X_0. Ceci peut aisément être étendu à la régression $X_i \mapsto X_{i+1}$ en considérant $g(Y_i) = |f(X_i) - X_{i+1}|$ par exemple. On suppose que les X_i et les $f \in F$ sont bornés et $\in [0,1]$. On suppose dans la suite que la loi μ_n de $X_n|X_0$ converge uniformément vers une mesure donnée μ_∞, dans le sens suivant:

$$\lim_{n\to\infty} \sup_{X_0} \int |\mu_n - \mu_\infty| \to 0 \tag{5.20}$$

La section 5.2.2 fournit des conditions suffisantes pour cela. On suppose, en outre, que $X_n|X_0 = t$ est continu par rapport à la mesure de Lebesgue. Ceci va seulement être utilisé en partie 5.3.1 pour prouver une extension exponentielle uniforme de l'équation 5.20; aussi cette hypothèse peut être relâchée, en utilisant le théorème 14 (première partie). L'intérêt est seulement de fournir ici une démonstration simple du résultat désiré, avec des prérequis aussi réduits que possible.

L'idée de la preuve, détaillée dans les sous-sections suivantes, est comme suit:

1. En un certain sens, les mesures dans les séquences markoviennes convergent uniformément et rapidement vers la mesure asymptotique. La vitesse de convergence est mesurée par une dimension entière.
2. Ceci implique que pour des sous-suites bien choisies de la suite initiale, les mesures sont "presque" indépendantes. Plus précisément, la loi de cette sous-suite est proche de la loi du produit indépendant de la mesure asymptotique.
3. La convergence uniforme des moyennes empiriques dans le cas indépendant implique la convergence uniforme des moyennes empiriques dans le cas original.

Première partie: convergence Markovienne Définissons $\Delta_n^{X_0}(P) = \mu_n^{X_0}(P) - \mu_\infty^{X_0}(P)$. $\mu_n^{X_0}$ est μ_n, conditionnellement à X_0. Alors:

$$\begin{aligned}
\Delta_n^{X_0}(P) &= \int_t \Delta_{n-k}^{X_0}(t) f_k(t,P) dt \\
\Delta_n^{X_0}(P) &= \int_t \Delta_{n-k}^{X_0}(t)(f_k(t,P) - f_\infty(t,P)) dt \qquad (5.21) \\
|\Delta_n^{X_0}(P)| &\le \int |\Delta_{n-k}^{X_0}|(t) |f_k(t,P) - f_\infty(t,P)| dt \\
|\Delta_n^{X_0}(P)| &\le \frac{1}{2} \int_t |\Delta_{n-k}^{X_0}|(t) dt \qquad (5.22)
\end{aligned}$$

La ligne (5.21) vient du fait que $\Delta_{n-k}^{X_0}$ a masse 0 et est indépendant de P. La ligne (5.22) est basée sur k suffisamment grand (en utilisant l'équation (5.20)) pour garantir $\| f_k(t,P) - f_\infty(t,P) \| < \frac{1}{2}$. Le plus petit tel k sera noté, par la suite, $D(X)$ (notation abusive pour $D(X_n|X_{n-1})$).

Définition 5.19 (Dimension d'une chaine markovienne) *On définit $D(X)$ la* **dimension** *d'une chaine de Markov. Le D est là pour "condition de Doeblin" ou "dimension", selon les préférences du lecteur.*

L'équation 5.22 implique ce qui suit:

Lemme 5.20 (Convergence rapide non-asymptotique dans les chaines de Markov ergodiques)

$$\Delta_n^{X_0}(P) \leq \frac{1}{2^{\lfloor \frac{n+1}{D(X)} \rfloor}} \Delta_1^{X_0}(P)$$

Seconde partie: apprentissage bruité La théorie de l'apprentissage avec bruit inclut de nombreux résultats pour de l'apprentissage perturbé; le "distribution shift", les erreurs malicieuses, le bruit CPCN, ([Aslam et al, 1996, Kearns et al, 1993, Decatur, 1995, Decatur, 1997, Gentile et al, 1998]). Ces résultats, pour autant que nous sachions, n'inclut pas le cas considéré. Cependant, il peut être traité directement.

Considérons X_N,X_{2N},...,X_{lN}, des sous-suites finies de X, avec $N = kD(X)$. Alors, $\mu_N^{X_0}$ conditionnellement à X_0 a une loi à distance $\leq \eta$ de μ_∞, $\mu_{X_{2N}}^{X_N}$ idem, et ainsi de suite, avec $\eta = \frac{1}{2^k} \sup_{X_0,P}$ masses de Dirac $\Delta_1^{X_0}(P) \leq \frac{1}{2^k}$ et avec la distance suivante:

$$d(\mu_1,\mu_2) = \sup_P |\mu_1 - \mu_2|(P)$$

(intuitivement, l'intégrale maximale d'une fonction bornée par 1, pour la loi $\mu_1 - \mu_2$)

On a alors besoin du lemme qui suit:

Lemme 5.21 (D'une loi à plusieurs) *Soient Z_1,...,Z_n des variables aléatoires (non nécessairement indépendantes!), chaque Z_{i+1} ayant une loi conditionnellement à Z_i à distance $< \eta$ de la loi de Z. Alors, la loi de $(Z_1,Z_2,\ldots,Z_n)$ est à distance $< n\eta$ de la loi du produit de n variables indépendantes avec la même loi que Z.*

Proof:Ceci se prouve par récurrence. La propriété pour $n = 1$ est claire. La récurrence est fait en intégrant la propriété au rang $n - 1$. □

Ceci implique que notre sous-suite a, à une précision explicite près, la loi du produit de n lois indépendantes.

Troisième partie: conclure Des résultats tels que l'équation 5.10 fournissent une borne sur la complexité d'échantillon dans le cadre iid en $O(V - \ln(\delta)/\epsilon^2)$ (ou $O(\frac{V\ln(1/\epsilon)-\ln(\delta)}{\epsilon})$, dans le cas d'un taux d'erreur minimal nul), avec ϵ la précision, $1 - \delta$ la confiance, V la VC-dimension (on considère ici le cas de la catégorisation deux-classes - on peut directement considérer la régression de même ou utiliser des fonctions d'égalié ϵ-insensibles en régression pour garder le même cadre qu'en classification). Considérons maintenant la complexité d'échantillon nécessaire pour garantir une différence bornée par ϵ entre les moyennes empiriques et les espérances; P est la probabilité sous l'hypothèse iid, alors que P_M est la probabilité dans le cadre markovien, pour la suite suite ci-dessus (la somme dans l'équation 5.23 est faite sur des copies indépendantes de μ_∞, qui sont distinguées par des indices (i)):

$$P(\exists f / |\frac{1}{k}\sum_{i=1}^{k} \mu_\infty^{(i)}(f) - \mu_\infty(f)| > \epsilon) \leq f(V,k,\epsilon) \tag{5.23}$$

$$P_M(\exists f,X_0 / |\frac{1}{k}\sum_{i=1}^{k} \mu_{X_{iN}}^{X_0}(f) - \mu_\infty(f)| > \epsilon) \leq f(V,k,\epsilon) + k\eta$$

Dans le théorème suivant on considère ERM, consistant à minimiser l'erreur empirique sur tous les exempes, et ERM_S consistant à minimiser l'erreur sur les points X_N,X_{2N},...,X_{kN} avec $N = D\ln(2m/\delta)$.

Théorème 5.22 *Quand $D(X)$ et V sont finis, et pour $\delta \leq \frac{1}{2}$, alors ERM_S a une complexité d'échantillon majorée par*

$$O\left(\frac{D(V + \ln(1/\delta))\ln(1/\delta)}{\epsilon^2}[\ln(DV) + \ln(\ln(1/\delta)) + \ln(1/\epsilon)]\right) \tag{5.24}$$

Quand le taux d'erreur minimal est nul, alors

$$O\left(\frac{D(V\ln(1/\epsilon) + \ln(1/\delta))\ln(1/\delta)}{\epsilon}[\ln(DV) + \ln(\ln(1/\delta)) + \ln(1/\epsilon)]\right) \tag{5.25}$$

En outre, l'équation (5.25) est valable pour ERM dans le cas d'une erreur minimale nulle, aussi. Pour ERM, dans le cas général, les mêmes bornes sont vraies, à des facteurs logarithmiques près.

[Gamarnik, 1999] fournit un résultat partiel dans le cas d'un espace d'état fini, pour apprendre avec une famille de fontions dont une qui a une erreur nulle en généralisation. En outre, [Gamarnik, 1999], comme rappelé plus tôt, fournit des bornes explicites de convergence uniforme de distributions vers la distribution asymptotiques.

Proof:Tout d'abord, considérons ERM_S. Grâce au lemme 21, les probabilités dans le cas Markovien et les probabilités dans le cas iid avec la loi μ_∞ sont à distance au plus

$$\delta_1 = O(\frac{k}{2^{\frac{N}{D}}}) \tag{5.26}$$

Dans le cas iid, avec confiance $1-\delta_2$, la précision est bornée par

$$\epsilon = O(\sqrt{\frac{V - \ln(\delta_2)}{k}}) \tag{5.27}$$

L'équation 5.26 satisfait $\delta_1 = O(\delta)$ si $N = D\ln(m/\delta)$. L'équation 5.27 avec $\delta_2 = Theta(\delta)$, conduit à $\epsilon = O(\sqrt{\frac{V-\ln(\delta)}{k}})$, vérifié avec probabilité $\Theta(\delta)$ dans le cas iid, et $\delta_1 + \delta_2 = \Theta(\delta)$ aussi dans le cas markovien.

Ceci conduit à la précision globale comme suit, avec confiance $1 - O(\delta)$:

$$\epsilon^2 = O(\frac{D(V + \ln(1/\delta))\ln(1/\delta)\ln(m)}{m}) \tag{5.28}$$

Ceci concuit à une complexité d'échantillon comme dans l'équation (5.24).

Le cas d'un taux d'erreur nul prend simplement en compte les bornes de complexité d'échantillon de la forme $(V\ln(1/e) + \ln(1/\delta))/\epsilon$.

Ceci montre qu'apprendre est possible avec complexité d'échantillon polynomiale en D, V, $1/\epsilon$, $\ln(1/\delta)$. Maintenant, considérons ERM. ERM en fait consiste en gros à utiliser N algorithmes ERM_S différents, qui ne sont pas indépendants. Dans le cas d'un taux d'erreur nul, ERM inclut ERM_S et donc est aussi efficace. Ainsi, l'équation 5.25 a lieu. Considérons maintenant le cas général.

Avec confiance $\geq 1-\delta' = 1-N\delta$, chacun des N apprentissages sur $(X_0, X_N, X_{2N}, \ldots)$, $(X_1, X_{N+1}, X_{2N+1}, \ldots)$,... a la précision ci-dessus (équation 5.28). Ainsi, on doit remplacer δ par $\delta' = O(\delta/N) = O(\frac{\delta}{D(\ln(m)+\ln(1/\delta))})$ dans l'équation 5.28.

Ainsi il nous faut

$$\delta = O(\frac{\delta'}{D\ln(m)\ln(1/\delta)})$$
$$\delta\ln(1/\delta) = O(\frac{\delta'}{D\ln(m)})$$
$$\frac{1}{\delta\ln(\frac{1}{\delta})} = \Omega(\frac{D\ln(m)}{\delta'})$$

Ceci est en particulier vérifié avec

$$\frac{1}{\delta} = (\frac{D\ln(m)}{\delta'} \times \ln(\frac{D\ln(m)}{\delta'}))$$

Remplacer $\frac{1}{\delta}$ par cette expression dans l'équation 5.28 conduit au résultat souhaité.□

Notez que nous avons proposé ci-dessus des bornes sur la différence entre l'erreur en généralisation et l'erreur empirique uniformément pour tout classifieur. Ceci n'est pas limité à la précision du classifieur empiriquement optimal.

Comme toujours en apprentissage ce théorème peut conduire à des algorithmes pratiques qui sont universellement consistent. Si V augmente suffisamment lentement (en tant que fonction de m), alors $\sqrt{DV/m}$ décroit vers 0. La minimisation du risque empirique est alors universellement consistente, pourvu que la séquence "emboitée" de classes de fonctions (chacune de VC-dimension finie), est un approximateur universel.

Optimalité On peut montrer que les dépendances linéaires en $\frac{D(X)V}{\epsilon^2}$ ne peut être supprimée, comme expliqué ci-dessous:

Théorème 5.23 *La dépendance linéaire en VD/ϵ^2 ne peut être supprimée. Précisément, pour tous V, D, δ, il existe une chaine de Markov de dimension D et une famille de VC-dimension D telle qu'avec probabilité au moins δ la précision ϵ est $\Omega(\sqrt{\frac{VD}{m}}) + O(D^{\frac{3}{2}}\ln(\frac{m}{D})/\sqrt{m} + D^3\ln(\frac{m}{D})^2/m)$.*

Proof:

Considérez une famille F de fonctions sur]0,1] avec VC-dimension V, une variable aléatoire μ telle que la complexité d'échantillon de F pour la distribution de μ soit la pire possible, Z_n chaine de Markov avec $Z_0 = 0$ et $Z_{n+1} = 1 - Z_n$ avec probabilité p et Z_n sinon, $X_n = Z_n \times \mu_{k(n)}$ avec les μ_n des copies indépendantes de μ et $k(n) = \sup([0,n] \cap \{i/Z_i \neq Z_{i-1}\})$.
Alors:

1. $D(X)$ est $\theta(1/p)$
2. La complexité d'échantillon $\Omega(VD(X)/\epsilon^2) + O(D^{\frac{3}{2}}\ln(\frac{m}{D})/\sqrt{m} + D^3\ln(\frac{m}{D})^2/m)$.

Le premier point est prouvé par l'évaluation de la loi de $(X_n - X_\infty)$, avec X_∞ la loi asymptotique. $X_n(0) - \frac{1}{2} = (1-2p) \times (X_{n-1}(0) - \frac{1}{2})$; la même relation est vraie pour tout sous-ensemble de]0,1].

Considérons maintenant l'évaluation empirique de $Eg(X_n)$ jusqu'à la $k+1^e$ occurence de $Z_n - Z_{n-1} = -1$. Ceci est une variable aléatoire N/P, avec $N = \sum_{i=1}^{k} \lambda_i g(A_i)$ et $P = \sum_{i=1}^{k}(\lambda_i + \lambda'_i)$, avec λ_i et λ'_i des variables aléatoires indépendantes égales à $k > 0$ avec probabilité $(1-p)^{k-1}p$, et A_i des variables aléatoires indépendantes avec loi commune μ.

• Tout d'abord, fixons les μ_i. Les espérances et probabilités ci-dessous sont calculées conditionnellement aux μ_i.

• La probabilité de $\lambda_i > K$ est bornée par $O((1-p)^K)$, et la probabilité d'avoir au moins un λ_i ou λ'_i plus grand que K est bornée par $O(k(1-p)^K)$. Pour un seuil de confiance fixé, on a $K = O(\ln(k)/p) = O(D\ln(k))$. La suite est faite conditionnellement à cela.

• Sous l'hypothèse $\forall i \max(\lambda_i, \lambda'_i) \leq K$, on peut conclure qu'avec un seuil de confiance, N et P sont tous deux en gros égaux à leurs espérances (toujours conditionnellement aux μ_i), avec précision $O(K\sqrt{k}) = O(D\ln(k)\sqrt{k})$, par l'inégalité de Hoeffding. N/P est alors, avec précision $O(D\ln(k)/\sqrt{k} + D^2\ln(k)^2/k) = O(D^{\frac{3}{2}}\ln(\frac{m}{D})/\sqrt{m} + D^3\ln(\frac{m}{D})^2/m)$, l'évaluation empirique de $Eg(A_n)$, qui est connu (grâce aux bornes inférieures de VC-dimension) $\theta\left(\sqrt{(VD/m)}\right)$ pour le pire choix a posteriori de g.

D'où le résultat souhaité.□

Notez que nous avons montré une un peu meilleure complexité d'échantillon de ERM_S, mais n'avons pas réussi à montrer qu'ERM_S était en un sens meilleur qu'ERM.

Une remarque important est le fait que ceci amène à un algorithme qui est universellement consistant: comme dans le cas d'échantillons iid, on peut utiliser ERM avec les modifications suivantes:

- accroissement (suffisamment lent) de la "taille" de la famille de fonction (en termes de VC-dimension), comme le nombre d'exemples augmente.
- Eventuellement, augmentation de l'"arité"[3], dans le cas de, disons, la prédiction (contrôle et stabilisation déduites de la prédiction de même). Pourvu que l'augmentation de VC-dimension résultante de cela est suffisamment lente (même combinée avec l'augmentation décrite ci-dessus).

Ceci est universellement consistant au sens où si la condition de Doeblin a lieu et si une arité finie est suffisante, alors l'erreur converge vers la plus petite possible. Néanmoins, une forte différence avec le cas iid persiste: l'erreur décroit comme dans le cas iid, mais si D est inconnu, alors *la validation est impossible.* On ne peut jamais être sûr que l'erreur en généralisation est ce qu'elle a l'air d'être sur un ensemble empirique d'exemples, sans borne sur D. Ainsi, des conditions générales sous lesquelles D peut être borné sont d'une importance cruciale. Des résultats liés peuvent être trouvés dans [Meyn, 1994].

3. On définit, pour la suite, l'arité comme le nombre de pas mémorisés - la largeur de la fenêtre. Ceci est usuellement la dimension de plongement, dans le cas de systèmes chaotiques.

5.3.2 Applications théoriques en identification de systèmes, contrôle et stabilisation

Identification de système et prédiction

Le problème d'identification de système est un vieux problème bien connu, qui bénéficie de vielles et claires définitions. Ainsi, les résultats les moins controversés en contrôle/stabilisation/identification de système sont dans cette aire. Comme dans le cas d'exemples iid, des algorithmes efficaces existent et bénéficient de preuves théoriques. Ceci est simplement une conséquence de théorème tels que 22 et 17. La partie 5.4.1 donne des références vers des expériences pratiques et des benchmarks classiques.

Comment convertir la prédiction en contrôle ou en stabilisation

Les sections 5.3.1 et 5.3.1 mettent en relief la possibilité de choisir f approximativement minimisant l'erreur de prédiction pour un système uniformément ergodique ou un système déterministe avec de bonnes propriétés stochastiques. Ceci peut être appliqué en identification de système. Des difficultés apparaissent pour des applications en théorie du contrôle. Pour les énoncer clairement, formalisons la solution intuitive, et étudions les difficultés.

Les paradigmes ci-dessous sont proposés pour le contrôle et la stabilisation.

Tout d'abord, supposons que l'identification de système (effectuée dans une phase séparée) a parfaitement réussi (ie dans le formalisme définie dans l'équation 5.2, on peut approximer g arbitrairement bien), et que l'état est entièrement connu (ce qui est le cas si la fenêtre est suffisamment large dans les systèmes chaotiques; ceci implique $h(x) = x$). Alors une idée intuitive consiste en minimiser $L(z)$ par une descente de gradient par rapport à $h(s_n)$ en équation 5.2.

Des tels outils analytiques de descente de gradient (et une analyse de convergence) a été développée dans [Hirasawa et al, 2001] par exemple.

Ceci peut approximativement être fait si la chaîne de Markov globale (avec feedback) vérifie une condition de Doeblin, et donc, a une dimension finie D. Alors, l'espérance de $L(z)$ peut être évaluée à l'étape n en évaluant (stochastiquement) la somme des $L_i(z_i)$ pour $i \in [n+1, n+K]$, considérant, grâce à la convergence exponentielle uniforme, que pour $K >> D$, les $L_i(z_i)$ ont (à peu près) la même espérance pour tout comportement initial. $EL(z)$ a à être évaluée pour tout choix de $h(s_n)$. La minimisation sur un nombre fini de pas mène à la notion d' **horizon**, par définition nombre de pas après lequel les variations sont négligées.

Problème ouvert 1
Y a-t-il une bonne raison pour que cette dimension soit finie, quand la dimension initiale (sans rétroaction) est finie? Si oui, comment l'évaluer?

Sous l'hypothèse ci-dessus (le système *sans feedback* a dimension finie D connue), une borne explicite sur K telle que minimiser $\frac{1}{K}\sum_{i=1}^{K} EL_{n+i}(z_{n+i})$ par rapport à $h(s_n)$ est optimal pour une précision et confiance donnés.

Notez que, sous cette hypothèse étrange (étrange car nous définissons le controleur par l'utilisation de K, qui est lui-même dépendant de D, qui est lui-même dépendant du controleur - mais pourvu que la controleur est extrait d'une famille donnée, on pourrait considérer une borne sur la dimension indépendante du controleur), et si l'identification de système est parfaite, on a prouvé (pour des valeurs arbitrairement petites de δ) que minimiser l'évaluation empirique de L comme défini ci-dessus mène à :

$$\begin{aligned} \forall n P(E(L(z)|h(s_n)) > \inf_{h(s_n)} E(L(z)+\epsilon|h(s_n))) < \delta \\ \text{and not} \\ P(E(L(z)) > \inf_{(h(s_n))_{n\in\mathbb{N}}} E(L(z)) + \epsilon) < \delta \end{aligned} \tag{5.29}$$

Décroitre δ à une vitesse exponentielle permet l'adaptation au second cas. Bien sûr, ceci n'est pas réaliste d'un point de vue du temps de calcul.

Ce paradigme sera dénommé, dans la suite, paradigme 1. Notez qu'une adaptation pratique consiste en minimiser ceci avec $K = 1$. De bons résultats expérimentaux sur des séries tests ont été reportés dans [Friedt et al, 2001] avec cette approche (en minimisant une fonction de coût égale à la valeur absolue de la fonction de retour d'une série de Feigenbaum, les auteurs allongent la durée avant déviation entre deux séries). Algorithmiquement, ils utilisent une descente de gradient.

Nous avons maintenant à considérer le problème d'identification de système *dans le cadre du contrôle*. L'identification est possible, comme expliqué ci-dessus, mais le problème est que quand une rétroaction est ajoutée, il n'y a absolument aucune raison pour l'état interne d'être dans la même aire que sans rétroaction. Cette difficulté a été soulignée et étudiée en pratique, dans le cadre des réseaux de neurones, dans [Hongping et al, 2001]. Une simplification consiste en simplement apprendre dans une phase séparée, et supposer que cet apprentissage sera robuste au changement de distribution opéré. Les expérimentations citées plus bas (partie 5.4.2) suggèrent que dans certains cas pratiques de systèmes chaotiques, cette approximation devrait être vraie. Ceci amène un second problème ouvert:

Problème ouvert 2
Y a-t-il une extension du théorème de Takens justifiant les résultats de la partie 5.4.2 ? Ie, la distribution asymptotique des X_n est-elle stable sous de petites variations des paramètres (ie, des signaux de contrôle régulier) ?

Ceci justifierait le paradigme 1.

Un autre paradigme, se débarassant de cette difficulté, consiste en remplacer la fonction de régression dynamiquement, par un apprentissage en ligne. Les difficultés suivantes apparaissent:

Problème ouvert 3
Comment pondérer les différents points (récents contre vieux) ? Une solution naturelle consiste en utiliser une pondération plus régulière à mesure que le nombre de points augmente. Comment quantifier cette régularisation? Y-a-t-il convergence du controleur résultant, donc de la distribution des X_n ? Notez qu'il ne s'agit pas, comme dans le cas de l'identification de système adaptatif, de simplement gérer les changements de régime du système, mais bien d'être capable de prendre en compte le changement de régime *du à la présence de la rétroaction*. Réapprendre au fur et à mesure de l'évolution de la rétroaction permet-il de stabiliser le processus?

5.4 Illustrations, benchmarks et expérimentations

5.4.1 Prediction / identification de système

[Friedt et al, 2000, Friedt et al, 2001, Haykin et al, 1998, Hong, 1993, Mukherjee et al, 1997, Teytaud et al, 2001] sont différents essais de prédiction de systèmes chaotiques (le plus célèbre étant sans doute la série de Mackey-Glass). Tous concludent à la possibilité pratique de prédire, et dans certains cas recommandent les réseaux de neurones, des réseaux à base radiale, ou des support vector machines. Dans [Mukherjee et al, 1997] les courbes d'apprentissage sont liées à la VC-théorie, sans détails formels sur la validité de la VC-théorie dans le cadre d'exemples non-iid. [Gamarnik, 1999] et ce papier donnent de nouveaux éléments en faveur de l'applicabilité de la VC-théorie dans ce cas. Dans beaucoup de cas, apprendre sur quelques centaines de points avec arité ≤ 6 était suffisant pour des précisions satisfaisantes. `http://neural.cs.nthu.edu.tw/`

`jang/benchmark` fournit une liste de benchmarks pour expérimentations sur séries temporelles chaotiques. `http://www.stern.nyu.edu/~churvich/Forecasting/Data/` fournit d'autres séries temporelles.

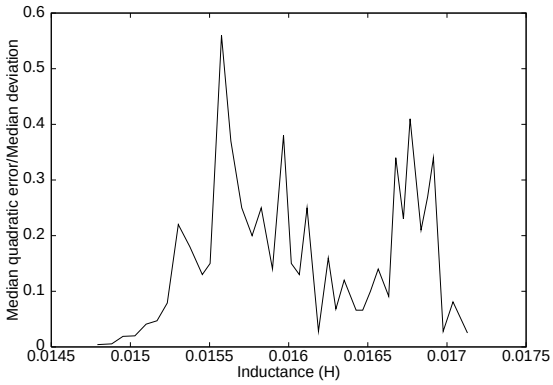

FIG. 5.2 – *Erreur dans la prédiction de fréquences instantannées en fonction de l'inductance.*

La figure 17 montre l'erreur en prédiction de fréquence instantanée, avec arité 3, dans le cas classique du circuit de Chua et autour des valeurs les plus difficiles du paramètre d'inductance. L'intérêt de ce circuit est partiellement dû à la similarité possible avec l'oscillateur de Colpitt ([Kennedy, 1995, Sarafian et al, 1995]).

Ces résultats expérimentaux ont été obtenus *sans apprendre spécifiquement des points ayant la même valeur de L*. En fait, les exemples dont les valeurs de L sont égales à celles de l'ensemble test on été *supprimés* de l'ensemble d'apprentissage. L'algorithme est local au sens où l'interval des valeurs de L a été coupé en 40 partie et a appris indépendamment sur chaque partie. Notez que 20 parties étaient en fait suffisantes pour une précision similaire.

5.4.2 Vers le contrôle?

Un problème pratique apparaissant dans les expérimentations est que tout l'apprentissage ci-dessus est basé sur l'idée que l'évolution du système est le même pendant l'application en contrôle et pendant l'apprentissage. Cependant, les conditions sont nécessairement différentes quand une rétroaction est ajoutée.

Dans de nombreux cas, l'existence d'une fonction déterministe de prédiction du comportement futur d'une série temporelle est une conséquence du théorème de Takens. Malheureusement, le théorème de Takens est seulement prouvé pour des valeurs fixés du paramètre. Cette partie teste l'hypothèse selon laquelle le théorème de Takens peut être utilisé quand le paramètre évolue lentement. Des résultats positifs expérimentaux sont reportés en figure 5.3. Notez que dans tous les cas l'erreur quadratique moyenne était environ 10^{-3} fois la variance.

5.4.3 Dimension d'une chaîne de Markov

On pourrait croire que dans les cas pratiques la condition de Doeblin est nécessairement vérifiée avec des constantes raisonnables, et que les contre-exemples théoriques évidents n'ont pas de sens pratique. Malheureusement, dans le cas de séries temporelles chaotiques extraites du circuit de Chua, nous voyons des exemples concret de grandes dimensions. La raison est que l'évolution continue du paramètre d'inductance du circuit de Chua entraîne une évolution discontinue de la fonction de retour. Quand le système est lancé près de la limite entre deux fonctions de retour, les constantes mises en œuvre augmentent à mesure que la valeur choisie est *moins* près de la limite. Ceci fournit un cas pratique qui est presque stable avec une fonction de retour, et qui soudain, après quelques secondes, utilise une autre courbe de retour. Cela est classique avec d'autres systèmes chaotiques munis de plusieurs attracteurs.

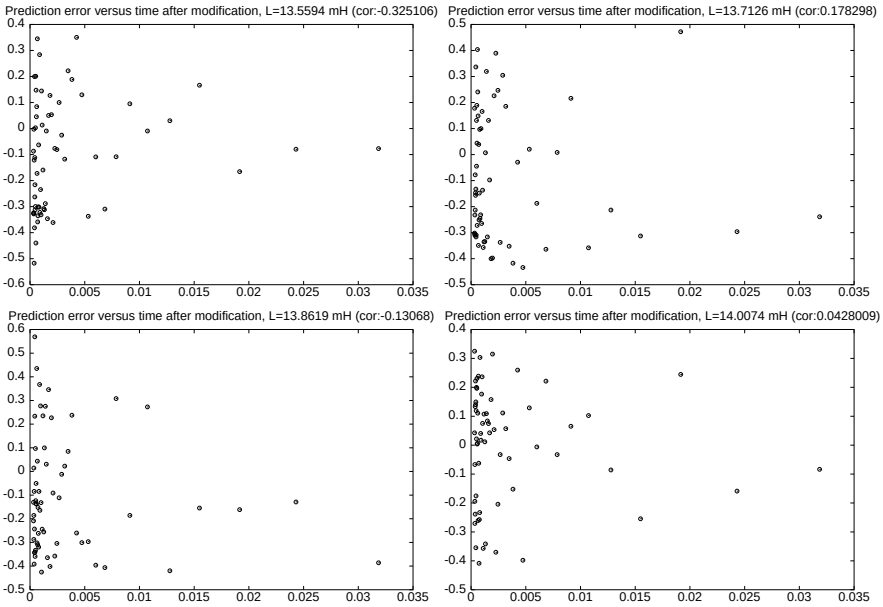

FIG. 5.3 – *Pour différentes valeurs de L, la corrélation entre le nombre de pas après modification du paramètre et l'erreur de prédiction, versus l'amplitude de la correction de L. Pour de grandes valeurs de la correction, la corrélation est négative. Pour les petites valeurs, la corrélation n'est pas clairement positive ni négative. Ceci montre que pour de petites modifications de L, le théorème de Takens reste vrai. Les corrélations sont évaluées sur les* 30 *premiers points après modification.*

5.4.4 Applications

Les applications en contrôle ont largement diffusé parmi la communauté réseaux de neurones. Par exemple, à Ijcnn 2001, les applications dans les aires suivantes ont été rapportées: soudure à l'arc à plasma, lithographie, double pendule ou pendule inversé, stabilisation de réacteur nucléaire, air conditionné, processus électrochimiques, mouvement de robot ou détection d'objets mobiles pendant l'autodéplacement, dispositifs thermoélectriques à pompage de chaleur, stabilisation de générateurs multiples sur la grille d'énergie, contrôle d'avion, applications financières, manœuvre d'hélicoptère Apache; une application amusante a été présentée par B. Widrow et M.M. Lamego: le contrôle d'un camion avec deux remorques avec un signal cible fourni dynamiquement par un utilisateur humain (le but est de permettre à un néophyte de guider dynamiquement un camion à reculons pour réaliser des huit ou des figures similaires). Ce dernier problème étend un problème maintenant classique consistant à faire reculer un camion-remorque dans un quai d'embarquement (voir plus bas). A Icann 2001 (proceedings chez Springer) d'autres problèmes étaient traités: annulation active de son, gestion de manipulateurs redondants, évitement d'obstacles, deux articles où le choix de chemin pour un robot était basé sur la prédiction mentale d'image après mouvement (d'inspiration biologique), la gestion de contrôleur PID par réseaux de neurones; plus divers outils de classification de séquences (ce qui peut être utile pour déterminer le régime d'un système) dont une nouvelle façon ([Trentin et al, 2001]) de mixer réseaux de neurones et modèles de Markov. L'identification de système/la prédiction n'étaient pas absents non plus et l'on proposait de ne pas utiliser de méthodes récurrentes lorsqu'un fenêtrage était suggéré par le théorème de Takens. Le traitement non-supervisé de séquences est très en vogue notamment avec le développement très important de l'analyse en composantes indépendantes, consistant en décomposer un signal supposé être un mélange bruité de plusieurs signaux indépendants (des résultats très élégants étant notamment obtenus par des méthodes bayésiennes).

Une comparaison entre différentes techniques neuronales de contrôle a été réalisée dans [De Jesus et al, 2001].

5.4.5 Un benchmark classique: le problème du camion-remorque. Exemples d'algorithmes.

Le maintenant classique problème de faire reculer un camion-remorque jusqu'à un quai d'embarquement a été initialement étudié dans [Nguyen et al, 1990]. Les équations ci-dessous sont extraites de [Koza, 1992, Chellapilla, 1998].

Variables (voir figure 5.4.5):

x,y: position du milieu de l'arrière de la remorque.
0,0: point cible.
y-axis: quai d'embarquement.
θ_t: angle entre la remorque et le quai d'embarquement.
θ_d: angle entre le tracteur et l'axe médian de la remorque.
$\theta_c = \theta_d + \theta_t$: angle entre le tracteur et le quai d'embarquement.
u: angle entre les roues et le tracteur.

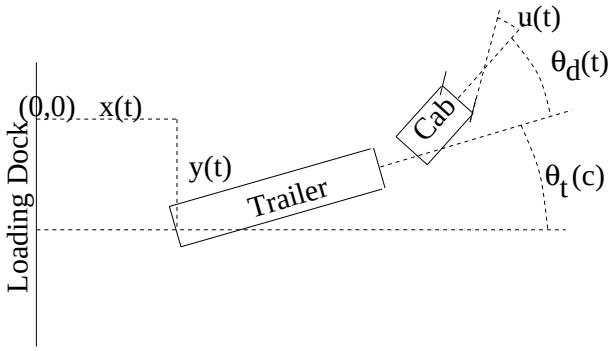

FIG. 5.4 – *Problème du camion-remorque.*

Equations du mouvement:

$$\begin{aligned} A &= r\cos(u(t)) \\ B &= A\cos(\theta_c(t) - \theta_t(t)) \\ C &= A\sin(\theta_c(t) - \theta_t(t)) \\ x(t+1) &= x(t) - B\cos(\theta_t) \\ y(t+1) &= y(t) - B\sin(\theta_t) \\ \theta_c(t+1) &= \tan^{-1}\left(\frac{d_c\sin(\theta_c(t)) - r\cos(\theta_c(t))\sin(u(t))}{d_c\cos(\theta_c(t)) + r\sin(\theta_c(t))\sin(u(t))}\right) \\ \theta_t(t+1) &= \tan^{-1}\left(\frac{d_s\sin(\theta_t(t)) - C\cos(\theta_t(t))}{d_s\cos(\theta_t(t)) + C\sin(\theta_t(t))}\right) \end{aligned}$$

Conditions initiales:

$$\begin{aligned} x(0) &\in [20m,40m] \\ y(0) &\in [-50m,50m] \\ \theta_t(0) &\in [-\pi/2,\pi/2] \\ \theta_d(0) &= 0 \end{aligned}$$

Conditions d'une trajectoire:

- Stop quand le temps est écoulé.
- Stop quand $x \leq 0$.

– Stop si succès: $|x| \leq 0.1, |y| \leq 0.42, |\theta_t(t)| \leq 0.12$

Fonction de réussite:

$$x(t)^2 + 2y(t)^2 + \frac{40}{\pi}\theta_t(t)^2$$

avec $t = 3000$, état final.

Le temps est mesuré en pas de 0.02 secondes. Les roues sont réorientées à chaque pas de temps. d_c, longueur du tracteur, vaut 6 mètres. d_s, longueur de la remorque, vaut 14 mètres. 60 secondes (3000 pas de temps) sont allouées. Le camion ne fait que reculer. L'angle des roues sature à -1 et $+1$ radian.

Des résultats expérimentaux ont été fournis par [Nguyen et al, 1990] (réseaux de neurones), [Kong et al, 1990, Kong et al, 1992] (flou et flou adaptatif), [Chellapilla, 1998] (programmation évolutive sur des parse-trees à longueur variable), [Tham et al, 1992] (apprentissage par renforcement pour réseaux de neurones avec différence temporelle et gradient stochastique), [Edwards et al, 1995] (réseaux de neurones avec temps continu entraîné par un algorithme direct basée sur une rétropropagation modifiée), [Schoenauer et al, 1994] (réseaux de neurones entrainés par algorithmes génétiques). Les avantages pratiques des réseaux de neurones résumés dans [Edwards et al, 1995] sont rappelés en partie 5.2.2. Les actes d'Icann 2001 ou d'Ijcnn 2001 fournissent de nombreux algorithmes pour de tels problèmes.

Une implémentation physique (en modèle réduit) du problème du camion à double remorque (un camion avec deux remorques doit suivre un signal cible dynamiquement fourni par un utilisateur - problème de *neurointerface*) a été réalisée et présentée à Ijcnn 2001 par Widrow et Lamego (pas de trace dans le proceedings).

5.5 Conclusion, remarques et problèmes ouverts

Cet état de l'art essaie de bénéficier de résultats provenant de différentes communautés des mathématiques (systèmes dynamiques, processus empirique)/physique statistique/informatique (VC-théorie) de manière à construire un cadre unifié pour le contrôle & l'identification de système & la stabilisation. Peut-être les questions suivantes pourraient elles être facilement résolues par des experts de ces différents domaines. Pour autant que nous sachions, le théorème 22, qui est une conséquence naturelle de résultats lointains les uns des autres réunis ici, est nouveau, et les solutions aux problèmes qui suivent, le seraient de même.

1. On montre dans le "beau" cas markovien une extension de la VC-théorie (bornes non-asymptotiques). On montre dans le cas déterministe, en utilisant des résultats de [Viana, 2001] et [Arcones, 1996], des résultats asymptotiques. Un progrès intéressant serait une extension déterministe de la VC-théorie, qui apparait comme le résultat manquant le plus important dans cet article.
2. L'identification de système peut bénéficier de nombreux résultats connus dans la littérature, tels que la convergence markovienne, les dynamiques stochastiques, ou la théorie de l'apprentissage non-iid. Par contre, des problèmes plus difficiles apparaissent en contrôl, tels qu'une sorte de "passage du local au global": on peut trouver (par descente de gradient) des optimisations "poussant" le système dans la bonne direction. Ceci n'assure pas une optimisation globale du critère, à moins que la condition de Doeblin du système avec rétroaction soit satisfaite, et que l'apprentissage fait sans rétroaction soit encore valable avec. Deux questions liées sont détaillées en partie 5.3.2.
3. Le théorème 17 est vrai sous deux conditions différentes. Le premier (cadre déterministe, issue, par exemple, des systèmes chaotiques) est bien controlé, grâce au théorème 1 ou des résultats proches qu'on trouvera dans [Viana, 2001]. Le second (cadre markovien) nécessite des hypothèses comme la condition de Doeblin, ou bien nécessite la réversibilité. Selon [Roberts et al, 1997, remarque 2.4] (qui sera intéressant pour un lecteur intéressé dans le cas réversible) de tels travaux sans réversibilité sont en cours. Des bornes non-asymptotiques existent grâce à la condition de Doeblin (comme expliqué en partie 5.3.1 et surtout au théorème 22) mais pourrait probablement être étendu grâce à des résultats basés sur des hypothèses plus faibles que la condition de Doeblin (équation 5.19).
4. Comment fournir des bornes générales sur la condition de Doeblin? Ceci semble un problème très important car dans les applications concrètes, cette constante est inconnue, et comme montré dans les résultats précédents, des bornes sur la dimension sont nécessaire pour construire des algorithmes

capables de décider s'ils ont assez d'exemples. Utiliser des mots tels que "quantique" ou "effets chaotiques sur la précision du système" semble parfois une sorte de sorcellerie; toutefois, on peut penser que chercher de tels arguments est un problème concret. Aussi nous formulons ce problème ouvert: l'incertitude quantique de physique statistique peut-elle justifier une condition générale sous laquelle la dimension ne peut être très haute? Voir [Balian, 1981] pour des considérations sur l'incertitude en physique quantique ou statistique.

5. Nous n'avons pas différencié "contrôle par apprentissage" et "contrôle adaptatif". Cela ne signifie absolument pas que la distinction n'ai pas de sens ou d'intérêt. Simplement nous travaillons sur l'aspect théorique et sans a priori sur la distribution, hormis la finitude de la VC-dimension ou de la "dimension markovienne", et donc n'avons pas intérêt à faire de différence. Dans un cadre concret, la distinction suivante apparait clairement:
   - Apprentissage sur les points récents seulement (nombre de points fonction de la VC-dimension et de la condition de Doeblin pour le régime en cours), et donc oubli total lors d'un changement de régime.
   - Apprentissage global, où l'ensemble, transition de régimes inclu, est modélisé par un réseau; tous les points sont utilisés dans l'apprentissage.

   Le cadre que l'on se fixe nous amène à l'apprentissage global. Le cas ou le signal d'entrée, de sortie, ou une autre source quelconque, permet de déceler un changement de régime pourrait donner lieu à d'autres formalisations. On pourra consulter [Aussem et al, 2001, Krishnamurthy et al, 1993, Holst et al, 1994] (segmentation de série temporelle par un modèle de Markov caché de réseaux neuronaux) pour une approche de ces problèmes.

Annexe A

Some interesting other results on Markov Chains

The following result is used in the context of learning in [Gamarnik, 1999]:

Théorème 5.24 *Let X_n be a* reversible *markov chain with asymptotic measure π. Then with $\pi_{\min}$ the minimum value of $\pi(x)$ for a state x, and λ the second largest eigenvalue of the transition matrix,*

$$|Pr(X_t = j|X_0 = i) - \pi(j)| \leq \frac{1}{\sqrt{\pi_{\min}}} \exp(-(1-\lambda)t$$

for any i and j states.

An extension to countable state spaces exists partially in [Gamarnik, 1999], based upon [Meyn, 1994].

The following theorem ([Meyn, 1994]) provides (in some cases) computable bounds for the geometric convergence rates of Markov chains.

Théorème 5.25 *Let X_n be a sequence of Markov chains. Assume that there exists a set C which is*

1. *a* **small set** *C, ie a set such that for some $m \in \mathbb{N}$, $\delta > 0$, μ measure of probability:*

$$P^m(x,A) \geq \delta\mu(A)$$

   for any A Borel set and $x \in C$.
2. *an* **atom**, *ie the transition function $P(X_{n+1}|X_n = x)$ is independent of $x \in C$.*
3. *a* **drift equation** *is verified, ie for a function $V \geq 1$*

$$\int P(x,dy)V(y) \leq \lambda V(x) + b\chi_C(x)$$

   with χ_C characteristic function of C, $\lambda < 1$, $b < \infty$.

Then for $n \in \mathbb{Z}^+$, $\sqsubseteq = \infty - \mathcal{M}_\alpha^{-\infty}$,

$$M_\alpha = \frac{1}{(1-\lambda)^2}(1 - \lambda + b + b^2 + \zeta_\alpha(b(1-\lambda) + b^2))$$

and

$$\zeta_\alpha = \sup_{|z|\leq 1} |\sum_{n=0}^{\infty}(P^n(\alpha,\alpha) - P^{n-1}(\alpha,\alpha))z^n|$$

(notice that ([Meyn, 1994, theorem 2.2]) if for some δ $P(C,C) > \delta$, then $\zeta_\alpha \leq \frac{32-8\delta^2}{\delta^2}(\frac{b}{1-\lambda})^2$) the following convergence holds:

$$\sup_{x\in X, f\leq V} \frac{\int f(y)P^n(x,dy) - f(y)\pi(dy)}{V(x)} \leq \frac{p}{p - \sqsubseteq}\rho^n$$

Some evaluations of ζ_α are made in [Meyn, 1994].

Annexe B

Weak convergence for convergence with temporal dependencies

The following theorems, related to theorem 14 (second part) are extracted from [Arcones, 1996] (theorems 2.1, 4.3, 4.5):

Théorème 5.26 (General result for weak convergence of stochastic processes indexed by smooth function *Let $\alpha > 0$ and F be the space of α-Hölder functions in $D \to \mathbb{R}$. Let G be a vector space of functions $D \to \mathbb{R}$ containing F and all the functions of the form $x \mapsto p(x)\chi_A(x)$ and $x \mapsto f(x)\chi_A(x)$ where p is polynomial, χ_A characteristic function of a Borel set $A \subset D$ and $f \in F$. Let Z_n be a sequence of stochastic processes indexed by G and let Z be another process indexed by the linear span of F. We assume that for a given sequence of $a_n \in \mathbb{R}$ and a given random variable X in D,*

1. *$Z_n(b_1 f_1 + b_2 f_2) = b_1 Z_n(f_1) + b_2 Z_n(f_2)$ for each $(f_1,f_2,b_1,b_2,n) \in G^2 \times \mathbb{R}^2 \times \mathbb{N}$.*
2. *$E|Z_n(f)| \leq c_0 \parallel f(X) \parallel_p$ for each $f \in G$.*
3. *$|Z_n(f)| \leq a_n \parallel f(X) \parallel_\infty$ almost surely for each $f \in G$.*
4. *$\frac{d(p-1)}{p} < \alpha$.*
5. *For each (f_1,f_2) in the linear span of F and each $(b_1,b_2) \in \mathbb{R}^2$,*

$$Z(b_1 f_1 + b_2 f_2) = b_1 Z(f_1) + b_2 Z(f_2)$$

6. *$Z_n(f) \xrightarrow{d} Z(f)$ for each f in the linear span of F.*

Then the sequence of Z_n converges weakly to Z in $l^\infty(F)$.

Théorème 5.27 *We consider X_n a strictly stationary sequence of random variables with values in $\mathbb{R}^d$. Moreover, we assume that* one of *the following hypothesis holds:*

– *with*

$$\rho(n) = \sup Ef_1(X_1,\ldots,X_k)f_2(X_{k+n+1}\ldots X_{k+n+j}) - Ef_1(X_1,\ldots,X_k)Ef_2(X_{k+n+1}\ldots X_{k+n+j})$$

the sup *being taken on $j,k \geq 1$ and functions f_1,f_2 such that $Ef_1^2(X_1,\ldots,X_k) \leq 1$, $Ef_2^2(X_{k+n+1},\ldots,X_{k+n+j}) \leq 1$, we assume that*

$$\sum_{n=1}^{\infty} \rho(2^n) < \infty$$

and $d/2 < \alpha$.

– *with* $r^{(p,q)} = E(X_r^{(p)} X_{r+k}^{(q)})$,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j,k=1}^{n} r^{(p,q)}(j-k) \text{ is finite, and}$$

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j,k=1}^{n} (r^{(p,q)}(j-k))^2 \text{ is finite}$$

for any $1 \leq p,q \leq d$, *and that* $d/2 < \alpha$.

Then

$$\{\frac{1}{\sqrt{n}} D_n f | f \in \mathcal{F}\} \to Df \text{ in } l^\infty(\mathcal{F})$$

with D *a centered gaussian process with covariances defined by*

$$Dfg = Cov(f_1(X_1),f_2(X_{1+k})) + \sum_{k=1}^{\infty} Cov(f_1(X_1),f_2(X_{1+k})) + Cov(f_2(X_1),f_1(X_{1+k}))$$

Les références gratuitement accessibles sur www sont préférés dans la bibliographie ci-dessous. Une grande partie d'entre elles peut être trouvée par une recherche avec www.citeseer.com.

Bibliographie

[Andersen et al, 1988] N.-T. Andersen, E. Giné, M. Ossiander, J. Zinn, *The central limit theorem and the law of iterated logarithm for empirical processes under local conditions, Probability theory and related fields 77, 1988.*

[Aldous et al, 1990] D. Aldous, U. Vazirani, *A Markovian extension of Valiant's learning model, Proc 31st Annual IEEE Symp. on the Foundations of Comp. Sci., p392-396, 1990.*

[Alon et al, 1997] N. Alon, S. Ben-David, N. Cesa-Bianci, D. Haussler, *Scale-sensitive dimensions, uniform convergence and learnability. Journal of the ACM, 44(4):615-631,1997.*

[Alexander, 1984] K. Alexander, *Probability inequalities for empirical processes and a law of the iterated logarithm. Annals of Probability, 4:1041-1067, 1984.*

[Amzallag et al, 1978] E. Amzallag, N. Piccioli, F. Bry, *Introduction à la statistique, 1978.*

[Arcones, 1996] M.A. Arcones, *Weak Convergence of stochastic processes indexed by smooth functions, Stochastic Processes and their Applications 62 115-138, 1996.*

[Aslam et al, 1996] J.-A. Aslam, S.-E. Decatur, *On the sample complexity of noise tolerant learning, information processing letters, 1996.*

[Aussem et al, 2001] A. Aussem, M. Fuentes, *Segmentation de série temporelle par une chaîne de Markov Cachée d'experts neuronaux, proceedings of CAP, 2001.*

[Balian, 1981] R. Balian, *Un principe d'incertitude fort en théorie du signal ou en mécanique quantique, CRAS Paris II, Vol 292, No 20, pp 1357-1362, 1981.*

[Bartlett, 1998] P.-L. Bartlett, *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, 44:525-536, 1998.*

[Baueri, 1996] P. Baueri, S. Nouak, R. Winkler, *A brief course in Fuzzy Logic and Fuzzy Control, version 1.2, 1996.*

[Birman et al, 1967] M.-S. Birman, M.-Z. Solomjak, *Piecewise-polynomial approximation of functions of the classes W_p, Mathematics if tge USSR Sbornik 73, 295-317, 1967.*

[Bishop, 1995] Bishop C.M. 1995, *Neural Networks for Pattern Recognition, Oxford*

[Bowen et al, 1975] R. Bowen, D. Ruelle, *The ergodic theory of Axiom A flows, Invent. Math. 29, pp181-202, 1975.*

[Bronstein, 1976] E.-M. Bronstein, *Epsilon-entropy of convex sets and functions, Siberian Mathematics Journal 17,393-398, 1976.*

[Chellapilla, 1998] K. Chellapilla, *Evolving nonlinear controllers for backing up a Truck-and-Trailer using evolutionary Programming, EPS 1998*

[Coullet et al, 1978] P. Coullet, C. Tresser, *Iterations d'endomorphismes et groupes de renormalisation, C.R. Acad. Sci. Paris 287, 577-580, 1978.*

[Decatur, 1995] S.-E. Decatur, *Efficient learning from faulty data, Thesis, 1995.*

[Decatur, 1997] S.-E. Decatur, *PAC learning with Constant-Partition Classification Noise and Applications to Decision Tree induction, Proceedings of ICML 1997.*

[De Jesus et al, 2001] O. De Jesus, M. Hagan, A. Pukrittayakamee, *A comparison of Neural Network Control Algorithm. Proceedings of Ijcnn 2001.*

[Devaney, 1992] R. Devaney, *Introduction to Chaotic Dynamical Systems: Theory and Experiments. Addison-Wesley, 1992.*

[Devroye, 1982] L. Devroye, *Bounds for the uniform deviation of empirical measures. Journal of Multivariate Analysis, 12:72-79, 1982.*

[Devroye, 1996] L. Devroye, L. Gyorfi, G. Lugosi, *A Probabilistic Theory of Pattern Recognition, Springer, 1996.*

[Doeblin, 1940] W. Doeblin, *Eléments d'une théorie générale des chaînes simples constantes de Markov. Annales scientifiques de l'École Normale Supérieure 57 (III) 61-111, 1940.*

[Dudley, 1974] R.-M. Dudley, *Metric entropy of some classes of sets with differentiable boundaries. Journal of Approximation Theory 10, 227-236, 1974. Correction; Journal of Approximation Theory 26, 192-193, 1979.*

[Dudley, 1978] R.-M. Dudley, *Central limit theorems for empirical measures, Annals of probability 6, 1978. Correction: Annals of probability 7, 1978.*

[Dudley, 1984] R.-M. Dudley, *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.*

[Edwards et al, 1995] N.J. Edwards, C.J. Goh, *Direct training method for continuous-time nonlinear optimal feedback controller, Journal of Optimization Theory and Applications, 84(3):509-528, 1995..*

[Eggenberger] P. Eggenberger, *Creation of Neural Networks based on Developmental and Evolutionary Principles, à paraître dans International Conference on Artificial Neural Networks ICANN'97, Lausanne, Suisse, 8-10 Octobre 1997*

[Feigenbaum, 1978] M. Feigenbaum, *Qualitative universality for a class of nonlinear transformations, J. Stat. Phys. 19,25-52, 1978.*

[Friedt et al, 2000] J.-M. Friedt, O. Teytaud, D. Gillet, M. Planat, *Simultaneous amplitude and frequency noise analysis in Chua's circuit - neural network based prediction and analysis, VIII Van Der Ziel Symposium on Quantum 1/f Noise and Other Low Frequency Fluctuations in Electronic Devices, 2000.*

[Friedt et al, 2001] J.-M. Friedt, O. Teytaud, M. Planat, *Learning from noise in Chua's oscillator, accepted in ICNF, 2001.*

[Gamarnik, 1999] D. Gamarnik, *Extension of the PAC Framework to Finite and Countable Markov Chains, Proceedings of the Workshop on Computational Learning Theory, Morgan Kaufmann Publishers, 1999.*

[1] F. Gruau et D. Whitley, *Adding Learning to the cellular development of neural networks: Evolution and the Baldwin Effect, Evolutionary computation, 1, 3:213-234. 1993*

[Gentile et al, 1998] C. Gentile, D.-P. Hambold, *Improved Lower Bounds for Learning from Noisy Examples: an information-theoretic approach, Proceedings of the 11th Annual Conference on Computational Learning Theory, ACM Press, 1998.*

[Harthong, 2000] J. Harthong, *Les origines de la physique statistique, http://moiret.u-strasbg.fr/hist/gaz.htm,2000.*

[Hayashi, 1997] S. Hayashi, *Connecting invariant manifolds and the solution of C^1 stability and Ω-stability conjectures for flows, Annals of Math. 145, 81-137, 1997.*

[Haykin et al, 1998] S. Haykin, J.Principle, *Using Neural Networks to Dynamically model Chaotic events such as sea clutter; making sense of a complex world, IEEE Signal Processing Magazine 66:81, 1998.*

[Henon, 1976] M. Hénon, *A two-dimensional mapping with a strange attractor, Comm. Math. Phys. 50, 69-77, 1976.*

[Hirasawa et al, 2001] K. Hirasawa, Y. Yu, J. Hu, J. Murato, *Stability analysis of nonlinear systems using high order derivatives of universal learning networks, proceedings of Ijcnn 2001.*

[Hoeffding, 1963] W. Hoeffding, *Probability inequalities for sums of bounded random variables. J. Amer. Statist. Assoc. 58, pp13-30, 1963.*

[Holmgren, 1996] R.A. Holmgren, *A first course in discrete dynamical systems. Springer, 1996.*

[Holst et al, 1994] U. Holst, G. Lindgen, J. Holst, M. Thuvesholmen, *Recursive estimation in switching autoregressions with a Markov regime, Journal of Time Series Analysis, Vol. 77, 257-287, 1994.*

[Hong, 1993] W. Hong, *The application of Radial Basis Function Networks to the prediction of Chaotic Time Series, Term Project, Course 2.160, Intelligent Control and Sensing, 1993.*

[Hongping et al, 2001] C. Hongping, K. Hirasawa, J. Hu, J. Murata, *A new robust neural network controller designing method for nonlinear systems. Proceedings of Ijcnn 2001.*

[Johansen et al, 1997] T.A. Johansen, E. Weyer, *On convergence Proofs in System Identification - A general principle using ideas from Learning Theory. Preprint, 1997: submitted to Elsevier Preprint.*

[Kearns et al, 1993] M. Kearns, M. Li, *Learning with malicious errors, SIAM Journal on Computing, 22, 807–837, 1993.*

[Kennedy, 1995] M.-P. Kennedy, *On the relationship between the chaotic Colpitts oscillator and Chua's oscillator, IEEE transaction on Circuits and Systems 42 (6), 1995.*

[Kolcinski, 1981] V.-I. Kolcinski *On the central limit theorem for empirical measures. Theory of probability and mathematical statistics 24, 1981.*

[Kolmogorov et al, 1961] A.-N. Kolmogorov, V.-M. Tikhomirov, *ϵ-entropy and ϵ-capacity of sets in functional spaces, Amer. Math. Soc. Transl. 17, pp 277-364, 1961.*

[Kong et al, 1990] S. Kong, B. Kosko, *Comparison of Fuzzy and Neural Truck Backer-Upper Control Systems, proceedings of Ijcnn 1990, 349-358, vol 3, NY:IEEE Press, USA.*

[Kong et al, 1992] S. Kong, B. Kosko, *Adaptive Fuzzy Systems for Backing up a Truck-and-trailer, IEEE transactions on Neural Networks, pp 211-223, vol 3, n2, 1992.*

[Koul, 1970] H. Koul, *Some convergence theorems for ranks and weighted empirical cumulatives. Annals of Mathematical Statistics 41, 1768-1773, 1970.*

[Koza, 1992] J.R. Koza, *Genetic Programming: on the programming of computers by means of natural selection. Cambridge, MA: MIT Press, 1992.*

[Krishnamurthy et al, 1993] V. Krishnamurthy, J.B. Moore, *On-line estimation of hidden Markov model parameters based on the Kullbach-Leibler information measure, IEEE Transactions on Signal Processing, Vol. 41, N8, pp 2557-2573, 1993.*

[Lorentz, 1963] E.N. Lorenz, *Deterministic nonperiodic flow, J. Atmosph. Sci. 20, 130-141, 1963.*

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

[Mané, 1988] R. Mané, *A proof of the C^1-stability conjecture, publ. math. IHES 66 (161-210), 1988.*

Chapitre 6

Classification avec VC-dimension infinie

Résumé

Pour obtenir un bon comportement en apprentissage, indépendamment de la distribution, on ne peut se limiter à la VC-dimension finie. Divers paradigmes permettent d'obtenir des résultats de convergence dans le cas général, que nous présentons et comparons.

On ne cherchera pas dans cette partie, qui se veut une sorte de regard d'ensemble sur l'apprentissage théorique, à rendre les résultats aisément utilisables par un lecteur peu au courant de certains des domaines évoqués. Ainsi, les classes de Donsker, la théorie VC, et de nombreuses notions seront à peine définies avant leur utilisation. Il est donc souhaitable, pour un lecteur peu au courant de ces notions, de lire au préalable le reste de cette thèse. La technicité elle-même des résultats montre bien qu'il n'existe pas de solution magique en apprentissage; il n'y a pas de cadre général simple et naturel dans lequel un algorithme s'impose comme le meilleur, à moins de disposer de connaissances a priori très nettes. Se replier derrière l'estimation d'un expert peut raisonnablement justifier une méthode Bayésienne, comme expliqué dans [Roberts, 2001] (à l'inverse d'une lecture invitée précédente dans la même conférence selon laquelle la seule approche scientifiquement valable était la minimisation structurelle du risque).

Table des matières

6.1 Introduction

Dans la suite, on considère un ensemble fini $D = \{(X_1,Y_1),(X_2,Y_2),\ldots,(X_m,Y_m)\}$ tirés iid (indépendamment identiquement distribués) dans $\mathcal{X} \times \{0,1\}$ et une famille $\mathcal{F}$ d'applications de $\mathcal{X}$ vers $\{0,1\}$ à VC-dimension V (éventuellement infini) et coefficient de pulvérisation $S(\mathcal{F},m)$ ou une famille $\mathcal{F}$ d'applications de $\mathcal{X}$ vers $[0,1]$. On appelle L^* l'erreur Bayésienne $\inf_{f\in\{0,1\}^{\mathcal{X}}} P(Y \neq f(X))$, et $L_{\mathcal{F}} = \inf_{f\in\mathcal{F}} P(Y \neq f(X))$. L_m est l'espérance de l'erreur pour f sélectionné par un algorithme d'apprentissage donné (éventuellement stochastique), pour la fonction de coût L. Les conditions pour la convergence uniforme (en la distribution des (X_i,Y_i)) vers les espérances sont bien connues: ceci est équivalent à la finitude de la VC-dimension. En outre, la convergence de EL_m est en gros linéaire $L_{\mathcal{F}} \times \sqrt{\frac{V}{m}}$.

Pour autant que nous le sachions, les meilleures bornes connues sont les suivantes, extraites de [Long, 1998] et [Vapnik, 1982]:

Théorème 6.1 *Pour tout $\varepsilon \geq 0, \delta \geq 0$, avec probabilité au moins $1-\delta$, si le nombre d'exemples est $\geq \frac{576 \ln 41}{\varepsilon^2}\left(4d + \ln\frac{1}{\delta}\right)$, alors la différence entre les moyennes empiriques et leurs espérances est plus petite que ϵ.*

Si $\mathcal{F}$ est tel qu'il existe $f \in \mathcal{F}$ tel que $L(f) \leq L < \frac{1}{2}$, alors le risque δ d'une déviation ϵ entre une moyenne empirique et son espérance est borné par $8S(\mathcal{F},m)\exp(-\frac{m\epsilon^2}{8(L+\epsilon)})$.

Ceci n'implique pas que la convergence est impossible dans un cadre plus général, mais seulement que l'on doit relâcher la clause d'uniformité (en la distribution). En outre, la finitude de la VC-dimension est une condition très restrictive, comme expliqué par le théorème ci-dessous, qui est prouvé de différentes manières par Benedek, Itai ([Benedek et al, 1994]), Devroye, Gyorfi, Lugosi ([Devroye et al, 1997], preuve basée sur des propriétés de la minimisation structurelle du risque face aux résultats négatifs généraux):

Théorème 6.2 (Conséquences de la finitude de la VC-dimension) *Supposons que $V < \infty$. Alors pour tout ϵ il existe une distribution telle que $L_{\mathcal{F}} > L^* + \frac{1}{2} - \epsilon$.*

Un résultat négatif général (prouvé par Devroye, Gyorfi, Lugosi dans [Devroye et al, 1997, p239], utilisant des résultats de Vapnik, Chervonenkis ([Vapnik et al, 1979]), Haussler, Littlestone, Warmuth ([Haussler, 1988])) sur la VC-dimension infinie est énoncé dans le théorème ci-dessous:

Théorème 6.3 (Résultat négatif à propos de la VC-dimension infinie) *Supposons que $V = \infty$. Alors pour tout $\eta > 0$, tout $m \in \mathbb{N}$, tout algorithme A, il existe une distribution avec une erreur optimale $L_{\mathcal{F}} = 0$, telle que $EL_m > \frac{1}{2e} - \eta$.*

Ceci interdit tout résultat positif de complexité d'échantillon de la forme "on peut trouver le classifieur optimal dans $\mathcal{F}$ avec précision ϵ pourvu que le nombre d'exemples est plus grand que $m(\epsilon)$". Mais des résultats dépendants de la distribution peuvent exister. Le résultat négatif plus fort ci-dessous (dans [Devroye et al, 1997], basé sur un résultat prouvé par Devroye dans [Devroye, 1982]), est basé sur des hypothèses plus fortes:

Théorème 6.4 *Pour toute suite décroissante $a_1,\ldots,a_m,\ldots$ de nombres positifs converge vers zéro avec $\frac{1}{16} \geq a_1 \geq a_2 \ldots$, si $\mathcal{F}$ pulvérise un ensemble infini, alors quel que soit l'algorithme d'apprentissage, il existe une distribution de (X_i,Y_i) telle que $EL_m \geq a_m$, bien que $L_{\mathcal{F}} = 0$.*

Ceci montre que s'il existe un ensemble infini pulvérisé par $\mathcal{F}$, alors l'apprentissage peut être arbitrairement lent. la différence avec le résultat précédent est que maintenant la distribution ne dépend pas de m. Ceci signifie que dans ce cas (ensemble infini pulvérisé), aucune vitesse asymptotique de convergence ne peut être donnée. Ainsi de telles vitesses asymptotiques peuvent seulement être vraies avec VC-dimension infinie (on verra que de tels résultats existent, fournissant des vitesses asymptotiques) mais sans pulvérisation d'un ensemble infini.

Ce papier survole et compare les résultats de l'état de l'art à propos de la convergence avec une VC-dimension infinie. Beaucoup de résultats sont basés sur la fonction de coût la plus usuelle: $L(f)$ est l'espérance de $Y \neq f(X)$ avec f à valeurs dans {0,1}, mais dans certaines classes, on considère une fonction de coût *probabiliste*: $L'(f)$ est l'espérance de $-Y\ln(f(X)) - (1-Y)\ln(1-f(X))$, avec F à valeurs dans [0,1]. L'^*, $L'_{\mathcal{F}}$ et L'_m ont leur sens intuitif: respectivement minimum global de l'erreur pour L', minimum relatif pour L', ou erreur après apprentissage sur un ensemble de taille m pour L'.

Nous ne travaillons pas sur des fonctions non-mesurables. Pour des études incluant ce cas, voir [Van Der Vaart et al, 199 par exemple. Beaucoup de résultats peuvent être étendus dans cette direction.

6.2 Théorie de l'approximation

La théorie de l'approximation permet de savoir quelles familles de fonctions permettent d'approximer telle ou telle fonction, et notamment quelles familles de fonctions permettent d'approximer toute fonction. Ceci est le fondement de plusieurs des paradigmes qui suivent. Ainsi, la minimisation structurelle du risque comme la minimisation de l'erreur empirique sur des familles de fonctions de taille croissante avec le nombre d'exemples bénéficient de résultats de consistence universelle grâce aux propriétés d'approximation des familles de fonctions considérées. Cette partie est ainsi une base fondamentale pour rendre non triviaux de nombreux théorèmes qui suivent, qui supposent que diverses propriétés d'approximation sont vérifiées.

Cette partie est essentiellement basée sur [Carothers, 1998] et survole quelques résultats utiles en théorie de l'approximation. D'abord, considérons certains résultats généraux pour l'approximation de par $y \in Y \subset X$, du point de vue de l'existence/l'unicité de y minimisant $\| x - y \|$.

- Pour Y un sous-espace de dimension finie d'un espace vectoriel normé X, et pour tout $x \in X$, il existe $y \in Y$ (non nécessairement unique) tel que $\| x - y \|$ est minimal. Ceci implique en particulier le fait que toute fonction continue f sur un sous-espace compact de $\mathbb{R}^d$ et tout $n \in \mathbb{N}$, il y a un polynôme P avec degré n minimisant la norme de $f - P$ parmi les polynômes de degré n (pour toute norme sur l'espace linéaire de fonctions continues sur X).
- Si Y a une dimension finie et est inclus dans l'espace linéaire X et *si* pour tout x, ce y est unique, alors l'application $x \mapsto y$ est continue.
- Si X est un espace linéaire et a une norme strictement convexe (ie tout point distinct de a et b dans [a,b] pour (a,b) $\in X^2$ a une norme plus petite que a et plus petite que b - ceci est le cas pour la norme usuelle dans $\mathbb{R}^d$ comme pour les normes L^p dans $\mathbb{R}^d$, mais pas pour les normes usuelles sur les sous-espaces de fonctions continues), alors pour tout Y sous-espace de X, *si* il existe un meilleur approximateur y de x, alors il est unique.
- Si Y est un sous-espace d'un espace linéaire normé X, alors l'ensemble des meilleurs approximateurs de tout $x \in X$ est borné et convexe.

Maintenant quelques résultats d'approximation:

- le théorème d'approximation de Bernstein: si X est l'espace des fonctions continues sur $[a,b] \subset \mathbb{R}$ et si $\epsilon > 0$, alors pour tout $f \in X$ il y a un polynome tel que $\| f - p \| \leq \epsilon$. En outre, un algorithme fournit une suite de polynomes convergeants vers f.
- Si T_n est une application linéaire de X vers X tel que $T_n(f) \to f$ avec X ensemble des fonctions continuues de [0,1] vers $\mathbb{R}$, pour chaque $f \in \{x \mapsto x, x \mapsto 1, x \mapsto x^2\}$ pour la norme infinie $\| f \| = \sup_{t \in [0,1]} |f(t)|$ alors $T_n(f) \to f$ pour tout $f \in X$ et pour la même norme (théorème de Bohman et Korovkin).

Maintenant voyons quelques couples X,Y d'espaces de fonctions telles que $X \subset Y$ et tout élément de Y est approximé arbitrairement bien par des éléments de X pour une norme donnée:

- Y ensemble des indicatrices d'ensembles mesurables de mesure de Lebesgue finie dans $\mathbb{R}^d$, X ensemble des fonctions indicatrices d'ensembles de la forme $f^{-1}(]0,\infty])$ pour f C^∞, pour la norme L^1.
- Y algèbre des fonctions continues de K dans $\mathbb{R}$, avec K compact, pour la norme du supremum. X sous-algèbre unitaire de Y. Supposons que X sépare les points, ie que pour tous a,b dans K il existe $f \in X$ tel que $f(a) \neq f(b)$. Alors le résultat d'approximation est vrai. Eg, les polynomes de K compact de $\mathbb{R}$ dans $\mathbb{R}$ sont des approximateurs pour la norme du supremum des fonctions continues de K dans $\mathbb{R}$.
- Théorème de Lusin: Y ensemble des applications mesurables d'un sous-ensemble mesurable de $\mathbb{R}^d$ de mesure de Lebesgue finie. Alors X, sous-ensemble des fonctions continues de Y, approxime Y pour la norme suivante:
$$\| f \| = \mu(\{x / f(x) \neq 0\})$$
- Y, ensemble des fonctions indicatrices de sous-ensembles fermés de $\mathbb{R}^d$, et X, sous-ensemble de Y des fonctions indicatrices des zéros de fonctions C^∞, pour la même norme que ci-dessus.
- X sous-ensemble des fonctions de Y C^∞ à support compact, avec Y ensemble des fonctions C^k de $\mathbb{R}^d$ vers $\mathbb{R}$, pour la norme de Holder dans les espaces C^k.
- X sous-ensemble des fonctions C^∞ à support compact de Y, ensemble des fonctions L^p de $\mathbb{R}^d$ dans $\mathbb{R}$, pour la norme L^p, pour $p < \infty$.

D'autres approximateurs universels important pour certaines normes sont les séries de Fourier, les polynomes trigonométrique.

6.3 Minimisation du Risque Empirique et estimateurs par squelettes

La minimisation du risque empirique (ERM) est connue efficace seulement dans le cas de la VC-dimension finie. Toutefois, le résultat suivant de [Devroye et al, 1997, p290], basé sur un choix judicieux de la taille de la famille de fonctions, justifie la minimisation du risque empirique sur une union dénombrable de familles de VC-dimension finie.

Théorème 6.5 *Soient $\mathcal{F}_1 \subset , \ldots, \subset \mathcal{F}_k$ ... de VC-dimensions finies V_1, ..., V_k, ... Soit $\mathcal{F} = \cup_n \mathcal{F}_n$. Supposons que chaque distribution conduise à $L_{\mathcal{F}} = L^*$ (de telles classes de fonctions existent!).*

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 $\mathcal{F}^{k_m}$ lorsque le nombre d'exemples est m a une erreur asymptotique égale à L^* avec probabilité 1.

Ce paradigme sera appelé, par la suite, ERM "incrémentale" - MRE^I en abrégé. Aucune vitesse de convergence n'est donnée, et aucune vitesse de convergence ne peut être donnée uniformément en la 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, [Benedek et al, 1994]), qui établit que les classes VC peuvent être de mauvais approximateurs:

Théorème 6.6 *Soient $\mathcal{F}_1$, ..., $\mathcal{F}_k$ ... ayant VC-dimension finie V_1, ..., V_k, ... Soient a_1,a_2,... une suite de nombres réels décroissant vers zéro. Alors il existe une distribution telle que pour k suffisamment grand,* $\inf L_{\mathcal{F}_k} - L^* > a_k$.

Ainsi, la convergence fournie par le théorème 5 peut être arbitrairement lente. Considérons maintenant des résultats (non-uniformes!) à propos de convergence rapides. Premièrement, certaisns résultats positifs sont possibles, au-delà de la VC-dimension finie, supposant que la loi de X est connue. Considérons un espace totalement borné de fonctions à valeurs dans [0,1]. Alors[1]:

Théorème 6.7 *Considérons $\mathcal{F}$ 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 $\mathcal{F}_\epsilon$ un ensemble couvrant fini pour ϵ de taille $N(\epsilon)$. Considérons un algorithme minimizant le risque empirique[2] parmi $\mathcal{F}_{\epsilon_m}$, 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 $\mathcal{F}$, alors

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

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

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

Ceci sera appelé, par la suite, ERM_C^I (minimisation du risque empirique incrémental avec couvertures). Ce résultat provient de [Devroye et al, 1997], où on peut trouver de nombreuses références sur des travaux liés. Utiliser des nombres de couverture pour L_∞ au lieu de L_1 amène à une borne uniforme en la distribution sous-jacente. La seule condition nécessaire est que $P(Y = 1|X)$ appartienne à $\mathcal{F}$:

Corollaire 6.8 *Considérons $\mathcal{F}$ une famille de fonctions à valeurs dans* [0,1]*, avec des nombres de couverture pour L_∞ $N(\epsilon)$. Considérons $\mathcal{F}_\epsilon$ une couverture finie pour ϵ de taille $N(\epsilon)$. Considérons un algorithme minimisant le risque empirique[3] parmi $\mathcal{F}_{\epsilon_m}$, 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 $\mathcal{F}$, alors

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

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

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

Ceci permet de travailler sur tout ensemble totalement borné de fonctions, sous une hypothèse qui est plus forte que l'hypothèse faite dans la SRM ci-dessous (où seule la fonction de *décision* a à appartenir à une famille donnée). D'autre part, la famille de fonctions sous considérations a seulement à avoir des ϵ-nombres de couverture finis.

La différence entre ces deux résultats 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 [Vidyasagar, 1997, p188] peuvent être établis de même, utilisant ce savoir a priori:

Théorème 6.9 *Considérons $\mathcal{F}$ 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 $\mathcal{F}_\epsilon$ un ensemble fini de couverture pour ϵ de taille $N(\epsilon)$. Considérons un algorithme minimisant le risque empirique[4] parmi $\mathcal{F}_{\epsilon_m}$, quand on apprend sur un ensemble d'apprentissage de taille m, avec ϵ_m choisi minimal tel que $m \geq \frac{2}{\epsilon^2}\ln(\frac{N(\epsilon)}{\delta})$.

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

1. Notez que nous pouvons considérer les nombres de couverture et les réseaux seulement dépendant de la distribution marginale en X.
2. Pour la fonction de coût usuelle L.
3. Pour la fonction de coût usuelle L.
4. Pour la fonction de coût usuel L.

Les théorèmes ci-dessous, extraits de [Van Der Vaart et al, 1996], sont des développements de résultats prouvés initialement par Donsker dans un cas restreint. Ils sont essentiellement basés sur des théorèmes centraux limites généralisés.

Théorème 6.10 (Entropie uniforme) *Considérons $\mathcal{F}$ une classe de fonctions mesurables qui satisfont la condition suffisante:*

$$\int_0^\infty \sup_Q \sqrt{\log N(\epsilon \parallel F \parallel_{Q,2}, \mathcal{F}, L_2(Q))} d\epsilon < \infty$$

où le supremum est pris parmi toutes les mesures de probabilités discrètes finies sur $\mathcal{X}$, F est une fonction enveloppe de $\mathcal{F}$ supposée L^2, $\parallel f \parallel_{Q,2} = \sqrt{\int f^2 dQ} > 0$. Supposons que les ensembles $\{f-g/(f,g) \in \mathcal{F}^2, \parallel f-g \parallel_{P,2} < \delta\}$ pour tout $\delta > 0$ et l'ensemble $\{(f-g)^2/(f,g) \in \mathcal{F}^2, \parallel f-g \parallel_{P,2} < \infty\}$ sont mesurables pour P.

Alors, $\mathcal{F}$ est Donsker pour P. Ceci signifie que

$$\frac{1}{\sqrt{m}}(\sum_{i=1}^{m}(f(X_i) - E(f)))$$

(appelé processus empirique*) converge faiblement vers un Borélien tight dans $l^\infty(\mathcal{F})$. En particulier, le supremum de cette différence (en valeur absolue) est faiblement $O(1/\sqrt{m})$.*

$L_{2,\infty}$, utilisé ci-dessous, est $\parallel f \parallel_{2,\infty} = \sqrt{\sup_{x>0} x^2 \times P(|f| > x)}$.

Théorème 6.11 (Bracketing entropie plus entropie) *Supposons que F, enveloppe de $\mathcal{F}$, a un second moment fini, et que*

$$\int_0^\infty \sqrt{\log N_{[\,]}(\epsilon, \mathcal{F}, L_{2,\infty}(P))} d\epsilon < \infty \tag{6.1}$$

$$\int_0^\infty \sqrt{\log N(\epsilon, \mathcal{F}, L_2(P))} d\epsilon < \infty \tag{6.2}$$

Alors $\mathcal{F}$ est P-Donsker.

Notez que les conditions (6.1) et (6.2) peuvent être remplacées par

$$\int_0^\infty \sqrt{\log N_{[\,]}(\epsilon, \mathcal{F}, L_2(P))} d\epsilon < \infty$$

Notez qu'aucun de ces théorèmes n'impliquent la consistance universelle. Notez, aussi, que les nombres de couverture peuvent être rendus uniformes en la distribution conditionnelle de Y sachant X, ce qui permet les résultats cités en conclusion (avec la terminologie de la conclusion, il s'agit de rendre les nombres de couverture uniformes en B et η).

Théorème 6.12 *Si, pour tout ϵ, $N_{[]}(\epsilon, \mathcal{F}, L_1(P)) < \infty$, alors $\mathcal{F}$ est P-Glivenko-Cantelli (pour la convergence presque sûre), ce qui signifie la convergence uniformes des moyennes empiriques vers les espérances.*

D'un point de vue pratique, la nécessité d'exhiber une ϵ-couverture a probablement empêché le développement des ϵ-squelettes. Les avantages théoriques suivants sont néanmoins intéressants:

- Possibilité de prendre en compte un savoir a priori sur la distribution marginale de X.
- Consistance universelle possible
- Possibilité de travailler sur des familles telles que les espaces de Holder de fonctions.

ERM^I est très aisément utilisé dans des cas pratiques. Il conduit à la consistance universelle, mais à des problèmes NP-complets dans beaucoup de cas.

6.4 Minimisation structurelle du risque

Beaucoup de résultats existent sous le nom de "minimisation structurelle du risque" (ou minimisation du risque structurel). Le théorème ci-dessous est extrait de [Devroye et al, 1997, p294] et a été initialement prouvé par Lugosi et Zeger.

Théorème 6.13 *Soient $\mathcal{F}_1$, ..., $\mathcal{F}_k$ ... ayant des dimensions finies V_1, ..., V_k, ... Soit $\mathcal{F} = \cup_n \mathcal{F}_n$. Supposons que toute distribution conduise à $L_{\mathcal{F}} = L^*$ (de telles classes de fonctions existent!). Considérons l'algorithme consistant à choisir $f \in \mathcal{F}$ 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 $f \in \mathcal{F}_k$. Alors:*

- *Si une règle de Bayes (une fonction conduisant à une erreur L^*) appartient à $\mathcal{F}_k$, 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.*

Le second résultat était déjà vrai pour la minimisation du risque empirique "amélioré" comme dans le théorème 5. 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*, and on assure *convergence asymptotique en* $O(1/\sqrt{m})$ (les facteurs logarithmiques peuvent être supprimé grâce à un résultat de Alexander, voir [Alexander, 1984]). Ceci sonne intuitivement comme un résultat de résistance au bruit; perturber la probabilité conditionnelle de $Y|X$ sans changer les fonctions bayésiennes perturbe peu l'apprentissage (avec la terminologie de la conclusion, la convergence est rapide pour des mauvais choix de η). Ce point pourrait toutefois être débattu par comparaison avec d'autres paradigmes dans le même cadre.

La minimisation du risque structurel est NP-complète dans beaucoup de cas intéressants de familles de fonctions. Les Support Vector Machines sont souvent présentées comme des minimiseurs du risque structurel, mais ceci ne nous apparait pas plus justifié que dans le cas de la rétropropagation avec régularisation. Notez qu'une approximation est faite dans les Support Vector Machines, de manière à transformer le problème NP-complet de la minimisation du risque structurel dans un problème quadratique qui est strictement convexe avec un domaine convexe. Les équivalents linéaires des Support Vector Machines sont parfois dites plus efficaces que les Support Vector Machines. Il n'y a pas de raison théorique à ça dans le cadre général, mais il est joli d'avoir un algorithme d'apprentissage traduisant un problème d'apprentissage en un problème de simplexe, pour lequel beaucoup d'algorithmes classiques existent.

6.5 k-plus proches voisins

Les k plus proches voisins sont le plus simple et le plus intuitif algorithme pour l'apprentissage. Le principe est un vote à majorité parmi les k voisins les plus proches. L'égalité parmi le nombre de voisins dans chaque classes peut être évitée en choisissant k impair. Les ambiguités dans le choix de l'ordre des distances peut être évitée soit en considérant, en cas d'égalité, que le point de plus petit index est plus près, où en ajoutant une composante, uniformément distribuée sur un intervalle, de telle manière que ce cas n'arrive presque jamais. Pour des raisons détaillées dans [Devroye et al, 1997], on considère le dernier cas. Dans le suite, k est choisi tel que $k(m) \to \infty$ et $k/m \to 0$. γ_d notant le nombres de cones d'angle $\Pi/6$ nécessaire pour couvrir $\mathcal{X} = \mathbb{R}^{d+1}$. Le résultat suivant a été prouvé par Devroye et Gyorfi, et Zhao. Une preuve claire peut être trouvée dans [Devroye et al, 1997]. Stone avait prouvé auparavant la convergence de EL_m vers L^*.

Théorème 6.14 (Consistance universelle forte) *Pour tout ϵ, pour n suffisamment grand (dépendant de*

la distribution),

$$P(L_m - L^* > \epsilon) \leq 2\exp(-\frac{m\epsilon^2}{72\gamma_d^2})$$

Ceci implique la convergence de L_m vers L^* avec probabilité 1.

Des algorithmes pratiques ont été établis pour les k plus proches voisins, la plupart d'entre eux basés sur la notion de séparations (KD-trees). Voir [Devroye et al, 1997] pour un rappel complet sur les plus proches voisins et pour des résultats de complexité.

6.6 Maximum de vraisemblance

Prouver des résultats négatifs pour le maximum de vraisemblance avec L pour fonction de coût est un exercice facile. Le premier résultat positif énoncé ci-dessous provient de [Devroye et al, 1997]:

Théorème 6.15 *Soit $\mathcal{F}$ une famille de fonctions à valeurs dans* [0,1]. *Supposons que les nombres* bracketing *$N_{[\,]}(\epsilon)$ de $\mathcal{F}$ sont finis. Alors, si $P(Y=1|X)$ appartient à $\mathcal{F}$, alors $L_m \to L^*$.*

En fait, L' est le cadre adapté pour des résultats positifs. Il n'y a finalement pas de raison pour que minimiser L' soit moins utile que minimiser L.

Maximiser la vraisemblance pour $f \in \mathcal{F}$ est maximiser le produit des probabilités $P(Y_i|X_i,w)$. Ceci est équivalent à maximiser la somme des $\ln(P(Y_i|X_i,w))$. Supposons par la suite qu'à la fois $P(Y=1|X,w)$ et $P(Y=0|X,w)$ sont minorés par c. Alors, on peut noter, comme remarque préliminaire, que $N(\epsilon/C,\mathcal{F}) \geq N(\epsilon,\{(u,v) \mapsto \ln(v\times(1-f(u))+(1-v)\times f(u))\})$ pour un C donné (dépendant de c). Rappelons maintenant deux théorèmes bien connus en processus empirique (voir [Van Der Vaart et al, 1996]).

Théorème 6.16 (Entropie uniforme) *Considérons $\mathcal{F}$ une classe de fonctions mesurables qui satisfont la condition suivante:*

$$\int_0^\infty \sup_Q \sqrt{\log N(\epsilon \parallel F \parallel_{Q,2},\mathcal{F},L_2(Q))}d\epsilon < \infty$$

où le supremum est pris parmi toutes les mesures de probabilités discrètes finies sur $\mathcal{X}$, F est une fonction enveloppe de $\mathcal{F}$ supposée L^2, $\parallel f \parallel_{Q,2} = \sqrt{\int f^2 dQ} > 0$. Supposons que les ensembles $\{f-g/(f,g) \in \mathcal{F}^2, \parallel f-g \parallel_{P,2} < \delta\}$ pour tout $\delta > 0$ et l'ensemble $\{(f-g)^2/(f,g) \in \mathcal{F}^2, \parallel f-g \parallel_{P,2} < \infty\}$ sont mesurables pour P.

Alors, $\mathcal{F}$ est Donsker pour P. Ceci signifie que

$$\frac{1}{\sqrt{m}}(\sum_{i=1}^m (f(X_i) - E(f)))$$

(appelé processus empirique*) converge faiblement vers un Borélien tight dans $l^\infty(\mathcal{F})$.*

$L_{2,\infty}$, utilisé ci-dessous, est $\parallel f \parallel_{2,\infty} = \sqrt{\sup_{x>0} x^2 \times P(|f| > x)}$.

Théorème 6.17 (Bracketing entropie plus entropie) *Supposons que F, enveloppe de $\mathcal{F}$, a un second moment fini, et que*

$$\int_0^\infty \sqrt{\log N_{[\,]}(\epsilon,\mathcal{F},L_{2,\infty}(P))}d\epsilon < \infty \tag{6.3}$$

$$\int_0^\infty \sqrt{\log N(\epsilon,\mathcal{F},L_2(P))}d\epsilon < \infty \tag{6.4}$$

Alors $\mathcal{F}$ est P-Donsker.

Notez que les conditions (6.3) et (6.4) peuvent être remplacées par

$$\int_0^\infty \sqrt{\log N_{[\,]}(\epsilon,\mathcal{F},L_2(P))}d\epsilon < \infty$$

Les applications sont les suivantes:

Corollaire 6.18 (Maximum de vraisemblance dans des classes de Donsker) *Supposons que $\mathcal{F}$ est P-Donsker et que $c \leq \mathcal{F} \leq 1-c$ avec $c > 0$. Alors L'_m converge faiblement en $O(1/\sqrt{m})$ vers L'^*.*

Des versions étendues peuvent être prouvées en direction de l'uniformité (partielle) en la distribution.

6.7 Inférence Bayésienne

Le résultat suivant a été établi dans [Teytaud et al, 2001]:

Théorème 6.19 *Si $\mathcal{F}$ est Donsker, alors L'_m (pour l'inférence Bayésienne) converge vers L'^* faiblement en $O(1/\sqrt{m})$, sous une condition d'intégrabilité uniforme au sens de Riemann dans $\mathcal{F}$. En outre, la différence entre l'erreur en généralisation et l'erreur empirique décroit uniformément pour tout f qui peut être choisi par l'algorithme[5] faiblement en $O(1/\sqrt{m})$. En outre, ces deux convergences sont non-asymptotiques et uniformes en la distribution sous-jacente P si la VC-dimension de $\mathcal{F}$ est finie.*

Ce résultat peut être prouvé en utilisant le fait que la limit simple de l'enveloppe convexe symétrique d'une classe de Donsker est Donsker (voir [Teytaud et al, 2001] pour des résultats sur l'approximation de l'inférence Bayésienne par les algorithmes de Gibbs).

La Bayes Point Machine est présentée par ses auteurs comme une approximation de l'inférence Bayésienne, ainsi que les Support Vector Machines étaient présentées comme minimisant le risque structurel. Selon les auteurs de la Bayes Point Machine, les Support Vector Machines essentiellement fonctionnent comme des approximateurs d'inférence Bayésienne.

6.8 Minimisation du risque empirique régularisé

Des travaux récents tels [Radulovic et al, 2000] étudient de nouveslles propriétés du processus empirique en utilisant différentes hypothèses d régularité sur la distribution d'exemples. Il serait certainement intéressant de considérer des résultats antérieurs de ce point de vue. Le principe est d'utiliser, comme approximation de la distribution, non pas la somme des masses de Dirac masses en les X_i's, mais $\frac{1}{nh}\sum_{i=1}^{n} K(\frac{x-X_i}{h})$ avec K tel que $\int xK(x)dx = 1$, $\int x^2K(x)dx$ et $\int K^2(x)$ sont finis. h est choisi tel que $nh^2 \to \infty$ et $nh^4 \to 0$. Ceci peut être réécrit comme une convolution. Dans le cas général, sans régularité, Yukich [Yukich, 1992] et Van der Vaart [Van Der Vaart et al, 1996] ont montré que cet estimateur est environ aussi bon que l'estimateur initial (ie, avec masses ponctuelles), en utilisant, en fait, des bornes sur la différence entre ces estimateurs. Mais Radulovic et Wegkamp [Radulovic et al, 2000] vont plus loins, montrant que cet estimateur est meilleur dans certains cas, car il fournit des résultats positifs dans le cas de familles prégaussiennes, non nécessairement Donsker. D'autre part, des conditions supplémentaires sont requises au niveau de la régularité; dans ce cadre, ils montrent le caractère nécessaire et suffisant du caractère prégaussien pour les familles indicatrices d'ensembles. En outre, ils montrent que l'estimateur classique peut échouter sur des familles prégaussiennes non-Donsker. Cet estimateur est donc *meilleur*, et non seulement *équivalent*, à l'estimateur classique. Quelques questions restent ouvertes après ce résultat. Y-a-t-il consistence universelle? Les familles prégaussiennes peuvent elles être des approximateurs universels? Un résultat positif de consistence universelle a été fourni il y a longtemps par Devroye et Krzyzak. Supposons que K est positif, et qu'avec B une boule de rayon r centré en 0, et avec $b > 0$, $K(x) \geq b$ si $x \in B$, et $\int \sup_{y \in x+B} K(y)dx < \infty$. Alors, si $h(n) \to 0$ et $nh^d \to \infty$, pour toute distribution, avec $f(x) = 1$ si et seulement si $\sum_{Y_i=1} K(\frac{X_i-x}{h}) > \sum_{Y_i=0} K(\frac{X_i-x}{h})$ (ceci est en gros - à part dans le cas d'égalité - équivalent au paragraphe ci-dessus dans le cas de la famille de *toutes* les fonctions caractéristiques), pour tout $\epsilon > 0$, il existe n_0 tel que pour tout $n \geq n_0$, $P(L_n - L_* > \epsilon) \leq 2\exp(-\frac{n\epsilon^2}{32\rho^2})$, avec ρ dépendant de K et d seulement.

5. Notez que f choisi par l'inférence Bayésienne dans une famille $\mathcal{F}$ n'est pas nécessairement dans $\mathcal{F}$.

6.9 Conclusion

Définissons $\epsilon(\delta,m,X,\eta,B)$ (resp. $\epsilon'(\delta,m,X,\eta,B)$) avec $\eta \geq \frac{1}{2}$ la différence entre l'erreur en généralisation et L^* (resp. $L_{\mathcal{F}}$) guaranti avec probabilité $1-\delta$, sur la distribution (X,Y), avec $P(Y=1|X)=\eta(X)$ si $B(X)=1$ et $P(Y=1|X)=1-\eta(X)$ si $B(X)=0$, après apprentissage sur m exemples. Intuitivement, B est un classifieur Bayésien (toutefois, il n'y a pas un unique classifieur Bayésien).

Comparons les différents paradigmes résumés ci-dessus, dans le cas de la fonction de coût L: (notez que ERM_S comporte les résultats positifs de ERM)

- ERM_C^I, ERM^I, SRM, $k-NN$, ERM_S assurent la consistence universelle (converge simple presque sûre de ϵ vers 0).
- ERM garantit une convergence simple en $O(1/\sqrt{m})$ de ϵ' vers 0 pour tout classe X-Donsker de fonctions $\mathcal{F}$. Ceci est uniforme en B et η, et peut être rendu uniforme sur des classes restreintes de X (eg, densité bornée par rapport à la distribution initiale X).
- ERM^I, ERM_S, SRM et k-NN garantissent une décroissante asymptotiquement exponentielle $P(\epsilon' > s)$ pour tout $s>0$, sur des familles de VC-dimension infinie de B.
- ERM^I et SRM tous deux garantissent la convergence asymptotique de ϵ en $O(1/\sqrt{m})$ sur une famille de VC-dimension infinie de fonctions B vers 0, uniformément en X,η pour un B donné.
- ERM_C^I comme dans le théorème 7 garantit la convergence asymptotique de ϵ en $O(\epsilon(m)+1/\sqrt{m})$, avec $\epsilon(m)=O(\log N(\epsilon(m))/m)$.
- ERM_C^I garantit la convergence uniforme de η, B de ϵ' vers 0 sur tout ensemble de fonctions totalement borné pour X. Ceci peut être rendu uniforme sur des familles restreintes de X (eg, densité bornée par rapport à la distribution initiale X).
- ERM_S garantit une convergence faible en $O(1/\sqrt{m})$ sur toute famille prégaussienne de fonctions indicatrices sous certaines hypothèses de régularité.

Quel est alors le meilleur paradigme, s'il y en a un?

- Point de vue **consistence universelle**: $k-NN$, SRM, ERM^I, ERM_S, ERM_C^I.
- Point de vue **complexité d'échantillon** dans une famille donnée de classifieurs (étant donnée une famille de fonctions, combien d'exemples me faut-il pour être sûr d'être près, avec confiance $1-\delta$, avec précision ϵ, du meilleur classifieur dans la famille considérée?) : ERM_C^I sur toute famille totalement bornée[6], ERM sur toute famille avec VC-dimension finie. SRM échoue ici, comme tout paradigme gérant des familles de VC-dimension infinie, à moins qu'un a priori sur la distribution (par exemple, sa distribution marginale en X, a priori qu'on ne prend pas en compte avec SRM) puisse être fourni.
- Point de vue **vitesse de convergence asymptotique de** ϵ': ERM garantit la convergence asymptotique de ϵ' en $1/\sqrt{m}$, pour toute classe de Donsker de fonctions pour la loi X. Ceci est uniforme en η et B et peut être rendu uniforme en des familles restreintes de X. Ceci est étendu à des familles prégaussiennes de fonctions indicatrice dans le cas d'ERM_S, sous hypothèse de régularité.
- Point de vue **vitesse de convergence asymptotique de** ϵ: ERM^I et SRM garantisse une convergence asymptotique en $O(1/\sqrt{m})$, uniformément en η et X, *pour un B donné dans une famille de VC-dimension infinie.*
- Point de vue **avec connaissance a priori sur la distribution**: si la distribution de (η,B) est connue, Bayes est bien sûr optimal. D'autres formes de connaissances a priori peuvent mener à différentes variantes de ERM^I, ERM_S (distributions régulières) ou SRM.
- Point de vue **avec régularité** (régularité des fonctions et de la distribution): ERM_S semble être très efficace dans ce cas. Améliorer les résultats de propriétés d'approximations des familles prégaussiennes est nécessaire pour une comparaison complète.

Finalement, un argument classique en faveur de la minimisation du risque structurel doit être analysé. On peut lire que la minimisation du risque structurel minimise une borne sur l'erreur en généralisation, et fournit des garanties sur l'erreur en généralisation, alors que d'autres paradigmes non. Un premier point

6. Pour la norme L^∞.

qui a à être souligné est qu'une borne *a posteriori* sur l'erreur en généralisation minimisée sur une union dénombrable de familles de VC-dimension finie, devient une borne *a priori*; le minimum atteint n'est pas garanti plus grand que l'erreur en généralisation, quelle que soit la confiance. De ce point de vue, ERM^I réussit mieux à fournir une borne a priori. Malheureusement, séparer l'apprentissage comme dans le *hold out*, fournit une borne sur l'erreur en généralisation bien plus précise. Cette conclusion selon laquelle de simples outils non-asymptotiques (le hold out fournit un ensemble de test qu'on peut tester non-asymptotiquement avec des coefficients de pulvérisations linéaires en m) peuvent remplacer la VC-théorie est renforcée par le fait que beaucoup d'articles illustrant les algorithmes basés sur la VC-théorie comme les support vector machines testent l'erreur en généralisation sur un échantillon séparé. Seuls quelques articles comparent les bornes en généralisations au lieu de l'erreur en généralisation testée sur un échantillon séparée (voir, toutefois, [Herbrich et al, 2000]).

Finalement, notre conclusion personnelle est que ni la VC-théorie ne peut réussir à fournir des algorithmes plus efficace que les algorithmes classiques pour des échantillons finis, du point de vue de la vitesse en fonction du nombre d'exemples. Bien sûr, ceci n'exclut pas des avantages pratiques possibles de certains paradigmes, les distributions usuelles étant quelque peu régulières. Nous avons une préférence philosophique pour Bayes ou ERM_S mais ceci n'est absolument pas un résultat formel. Si vous disposez d'un échantillon fini et n'avez aucun a priori, vous pouvez juste choisir une famille de fonctions (de VC-dimension adaptée) et valider le modèle extrait. Néanmoins, si vous pouvez utiliser un oracle fournissant un nombre infini d'exemples, alors une solution pour garantir une précision ϵ avec confiance $1 - \delta$ consiste simplement en choisir le meilleur classifieur parmi une première famille, essayer de le valider avec confiance $1 - \delta/2$. Si ça marche, vous avez réussi; sinon, vous pouvez choisir une autre famille pour apprendre avec un autre échantillon (et éventuellement d'autres entrées aussi, en augmentant la dimension d'entrée!) et essayer de la valider avec confiance $1 - \delta/4$, et ainsi de suite. Cet algorithme n'est bien sûr pas garanti - peut-être l'erreur Bayésienne est-elle plus grande que ϵ. Mais pourvu que l'erreur Bayésienne est strictement plus petite que ϵ et si votre famille est bien choisie, alors l'algorithme va réussir avec probabilité un.

Bibliographie

[Alexander, 1984] K. Alexander, Probability inequalities for empirical processes and a law of the iterated logarithm. Annals of probability, 4:1041-1067, 1984.

[Benedek et al, 1994] G. Benedek, A. Itai, Nonuniform Learnability. Journal of Computer and Systems Sciences, 48:311-323. 1994.

[Carothers, 1998] N.L. Carothers, A short course on Approximation theory, summer course on approximation theory, Bowling Green State University, 1998.

[Devroye et al, 1997] L. Devroye, L. Györfi, G. Lugosi, A probabilistic Theory of Pattern Recognition, Springer, 1997.

[Devroye, 1982] L. Devroye, Necessary and sufficient conditions for the almost everywhere convergence of nearest neighbor regression function estimates, Zeitschrift fur Wahrscheinlichkeitstheorie un verwandte Gebiete, 61:467-481, 1982.

[Haussler, 1988] D. Haussler, N. Littlestone, M. Warmuth, Predicting {0,1} functions from randomly drawn points. In Proceedings of the 29th IEEE symposium on the Foundations of Computer Science, pp. 100-109, IEEE Computer Society Press, Los Alamitos, CA. 1988.

[Herbrich et al, 2000] R. Herbrich, T. Graepel, J. Shawe-Taylor, *Sparsity vs. Large Margins for Linear Classifiers. . In Proceedings of the Thirteenth Annual Conference on Computational Learning Theory, pages 304-308, 2000.*

[Long, 1998] P.M. Long, The Complexity of Learning According to Two Models of a Drifting Environment, Proceedings of the 11th Annual Conference on Computational Learning Theory, 116-125, ACM press, 1998.

[Radulovic et al, 2000] D. Radulovic, M. Weckamp, *Weak convergence of smoothed empirical processes: beyond Donsker classes. High Dimensional Probability II, E. Gine, D. Mason and J. Wellner, Editors. Birkhauser, 2000.*

[Roberts, 2001] S. Roberts, Independent Component Analysis: a Bayesian perspective, invited lecture at Icann 2001.

[Teytaud et al, 2001] O. Teytaud, H. Paugam-Moisy, Bounds on the generalization ability of Bayesian Inference and Gibbs algorithms, Proceedings of Icann 2001.

[Van Der Vaart et al, 1996] A.-W. van der Vaart, J.-A. Wellner, Weak convergence and Empirical Processes, Springer, 1996.

[Vapnik et al, 1979] V. Vapnik, A. Chervonenkis, Theory of pattern recognition. Nauka, Moscow (in Russian). German translation: Theorie der Zeichenerkennung, Akademie Verlag, Berlin, 1979.

[Vapnik, 1982] V. Vapnik, *Estimation of Dependencies Based on empirical data. Springer-Verlag, New York, 1982*

[Vidyasagar, 1997] M. Vidyasagar, *A theory of learning and generalization, Springer 1997.*

[Yukich, 1992] J.E. Yukich, *Weak convergence of smoothed empirical processes. Scandinavian Journal of Statistics, 19, 271-279, 1992.*