Introduction à la théorie de l'apprentissage statistique

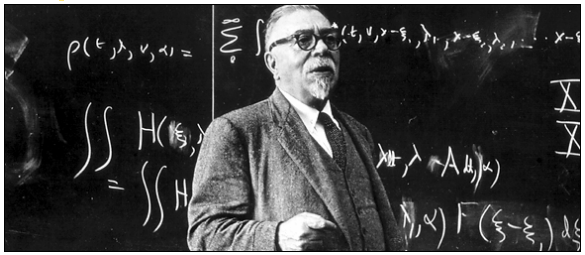
Gilbert Saporta
Chaire de Statistique Appliquée & CEDRIC
CNAM
292 rue Saint Martin, F-75003 Paris

gilbert.saporta@cnam.fr
http://cedric.cnam.fr/~saporta

Décembre 2012

- Guillaume d’Ockham 1320

- Norbert Wiener 1948



- Frank Rosenblatt 1962

- Vladimir Vapnik 1982

- Une introduction aux théories de V.Vapnik
  (rédigée en collaboration avec Michel Béra, chaire de modélisation statistique du risque)

Le problème de la boîte noire et l'apprentissage supervisé

- Etant donnée une entrée x, un système non déterministe renvoie une variable $y = f(x)+e$. On dispose de n paires (x_i,y_i) et on cherche une fonction qui approxime la fonction inconnue f.
- Deux conceptions:
  - Une bonne approximation est une fonction proche de f
  - Une bonne approximation est une fonction qui donne un taux d'erreur voisin de celui de la boîte noire

Risque d'apprentissage

- Apprentissage "supervisé"
- Y réponse à prédire, X prédicteurs
  - Y numérique régression ; binaire (-1;+1) discrimination
- Un modèle calcule un prédicteur

- où: $$\hat{y} = f(X, w)$$
  - f classe de fonction
  - w est un paramètre qui définit le modèle, estimé sur l'ensemble d'apprentissage

- Fonction de perte $L(y;f(\mathbf{x},\mathbf{w}))$
  - *Régression* $L(y;f(\mathbf{x},\mathbf{w}))=(y-f(\mathbf{x}))^2$
  - *Discrimination : taux (ou coût) d'erreur de classement*
    - *y et* $\hat{y}$ *à valeurs dans* {-1 ;+1}

$$L(y;\hat{y}) = \frac{1}{2}\left|y-\hat{y}\right| = \frac{1}{4}\left(y-\hat{y}\right)^2$$

- Risque (erreur de généralisation sur de nouvelles données $z = (X, y)$)

$$R = E(L) = \int L(z,w)dP(z)$$

- Objectif impossible: minimiser sur w le Risque

  - $P(z)$ probabilité inconnue

- On dispose seulement de n cas d'apprentissage $(z_1, .. , z_n)$ tirés suivant la loi $P(z)$, au lieu de minimiser R, on minimise le Risque Empirique :

$$R_{emp} = \frac{1}{n} \sum_{i=1}^{n} L(y_i ; f(\mathbf{x}_i ; \mathbf{w}))$$

- Problème central en théorie de l'apprentissage:

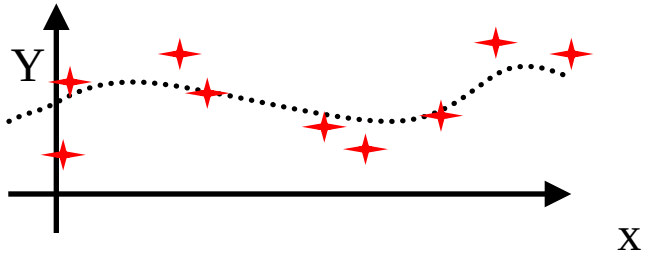
  Quelle est la relation entre le Risque R et le risque empirique R_{emp} ?

- Quelle est la capacité de généralisation de ce genre de modèle?

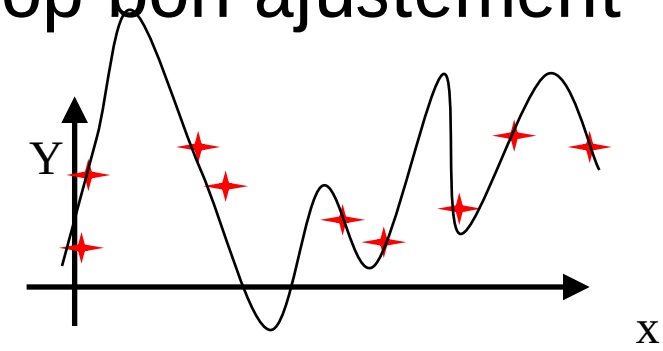
Robustesse

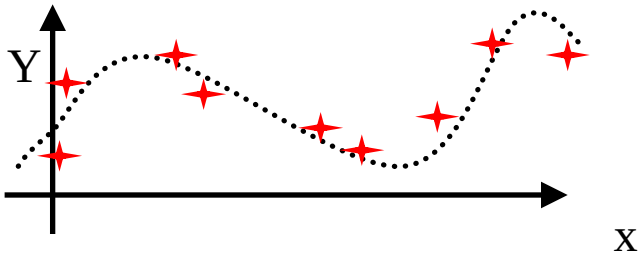
- Modèle robuste: erreurs en apprentissage et en généralisation du même ordre de grandeur

- Modele robuste



trop bon ajustement



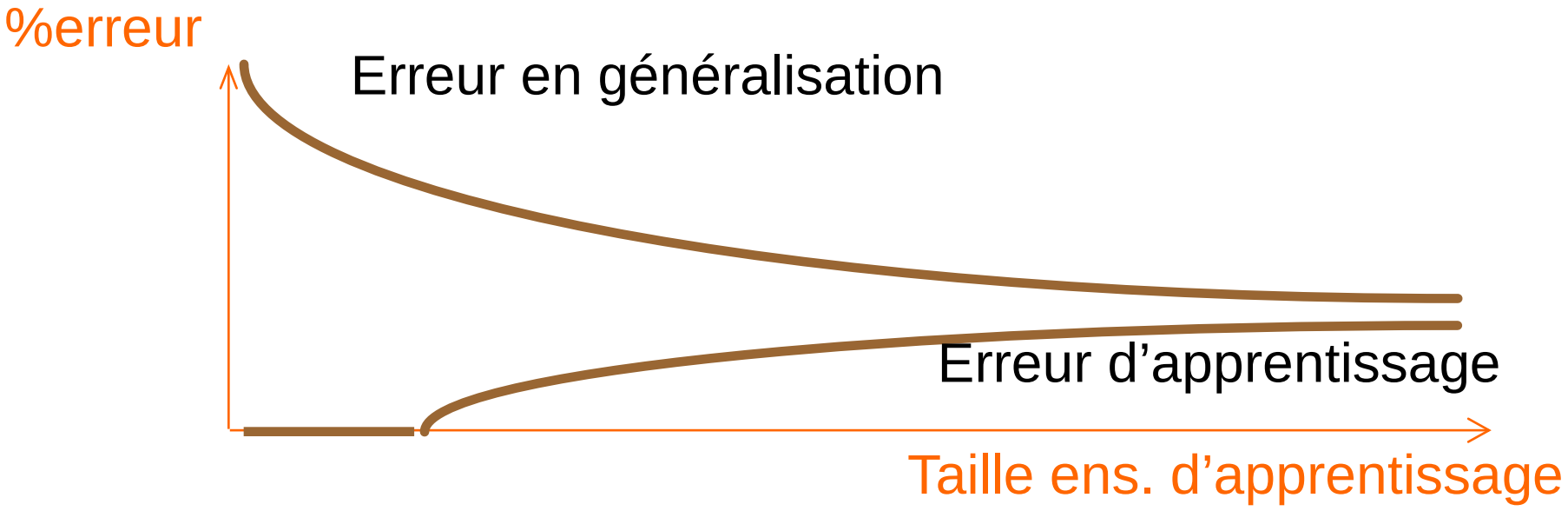


Compromis

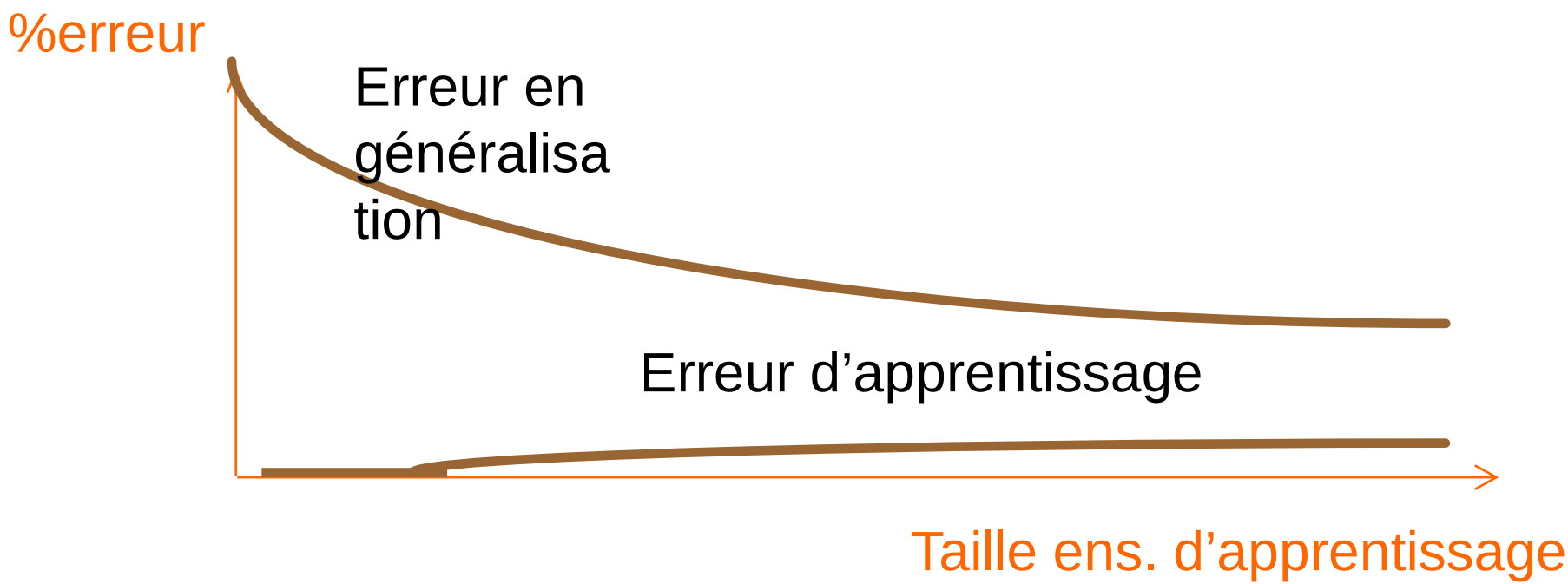
Consistence

- Un processus d'apprentissage est consistent si l'erreur sur l'ensemble d'apprentissage converge, lorsque la taille de cet ensemble augmente, vers l'erreur en généralisation.

Apprentissage consistent



Apprentissage non consistent



Les quatre piliers de la théorie de l'apprentissage

- **1** Consistence (garantit la généralisation)
  - Sous quelles conditions un modèle peut-il généraliser?

- **2** Vitesse de convergence en fonction du nombre d'exemples (mesure de la généralisation)
  - Comment s'améliore la généralisation lorsque le nombre d'exemples augmente ?

Quatre piliers de la théorie de l'apprentissage

- **3** Contrôle de la capacité de généralisation
  - Comment contrôler efficacement la généralisation à partir de l'information contenue dans un ensemble d'apprentissage de taille finie ?

- **4** Construire des algorithmes d'apprentissage
  - Existe-t-il une stratégie pour construire des algorithmes qui garantissent, mesurent et contrôlent la capacité de généralisation de modèles d'apprentissage ?

La VC dimension

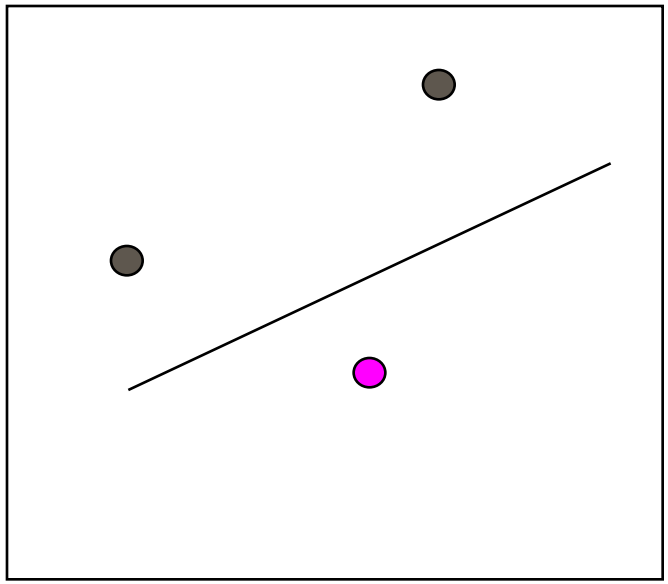
- Dimension de Vapnik-Cervonenkis: une mesure du pouvoir séparateur (complexité) d'une famille de fonctions $f(X,w):\mathbb{R}^p \rightarrow \mathbb{R}$
- VC dimension : un nombre entier attaché à une famille F de fonctions
- Chaque f de F – c'est-à-dire, pour un w donné – peut-être utilisé pour de la classification :
  - $f(X,w) >= 0$: X classé en 1
  - $f(X,w) < 0$: X classé en -1

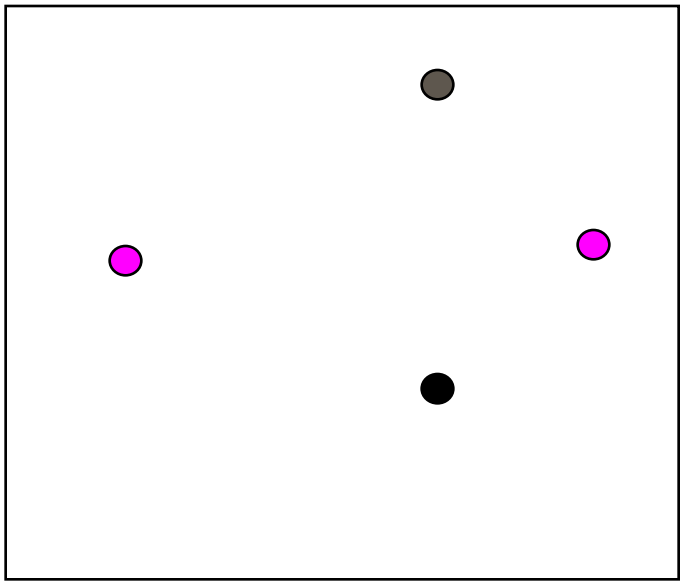
VC dimension suite

- Pour un échantillon de n points $(x_1, .. , x_n)$ de $\mathbb{R}^p$ Il existe 2^n manières différentes de séparer cet échantillon en deux sous-échantillons
- Un ensemble F de fonctions $f(X,w)$ “hache” *(shatters)* l’échantillon si les 2^n séparations peuvent être faites par des $f(X,w)$ différentes de la famille F

Exemple

- En 2-D, les fonctions linéaires (droites) peuvent “hacher” 3 points, mais pas 4





Aucune ligne droite ne peut séparer les points noirs des points roses

Un ensemble de fonctions de $\mathbb{R}^p \to \mathbb{R}$ a la dimension h si :

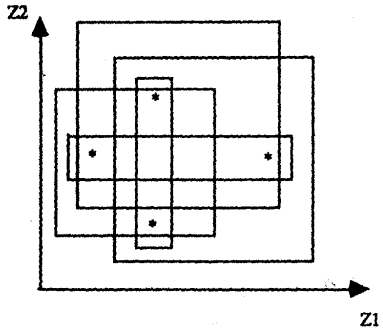
- Il existe un jeu de h points de $\mathbb{R}^p$ qui peut être “haché”, quel que soit l’étiquetage des points
- Aucun ensemble de $h+1$ points ne peut être haché par cet ensemble de fonctions.

Quelques exemples

- La VC dimension de l'ensemble des hyperplans de $\mathbb{R}^p$ est $p+1$

- Hyper-rectangles de $\mathbb{R}^p$ parallèles aux axes

  h=2p



(V.Cherkassky, F.Mulier, 1998)

- Sphères de $\mathbb{R}^p$

  h=p+1

Mais les VC dimensions ne sont PAS égales au nombre de paramètres libres

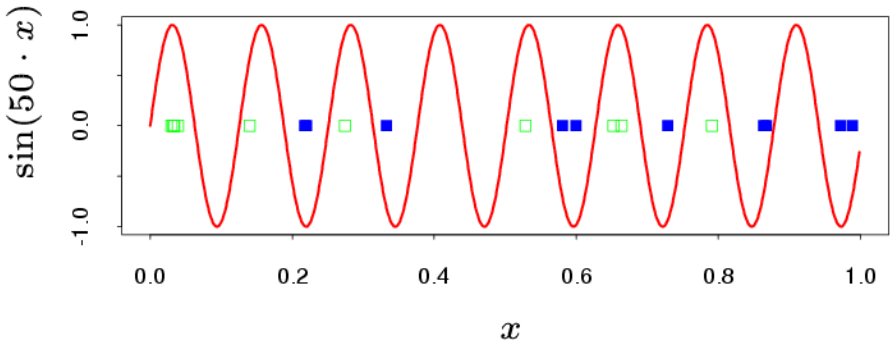
- La VC dimension de l'ensemble de fonctions

  $f(x,w) = sign\,(sin\,(w.x)\,)$,

  $c < x < 1$, $c>0$,

  avec un paramètre libre w est infinie.

Hastie et al. 2001



Deux cas importants:

a) régression ridge

- La VC dimension de l'ensemble des indicatrices linéaires $f(X,\mathbf{w}) = sign\left(\sum_{i=1}^{p}(w_i x_i)+1\right)$

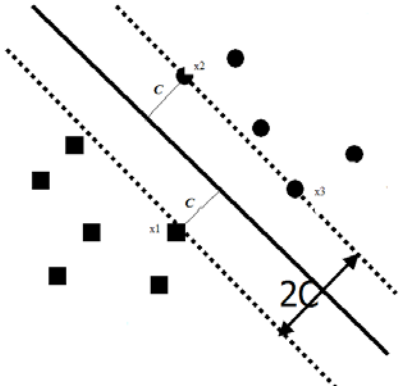
$$\|X\| \leq R$$

satisfaisant à la condition : $\|W\|^2 = \sum_{i=1}^{p} w_i^2 \leq \frac{1}{C}$

dépend de C et peut prendre toute valeur de *0* à *p+1*.

$$h \leq \min\left[ent\left(\frac{R^2}{C^2}\right); p\right]+1$$

b) L'hyperplan de marge maximale



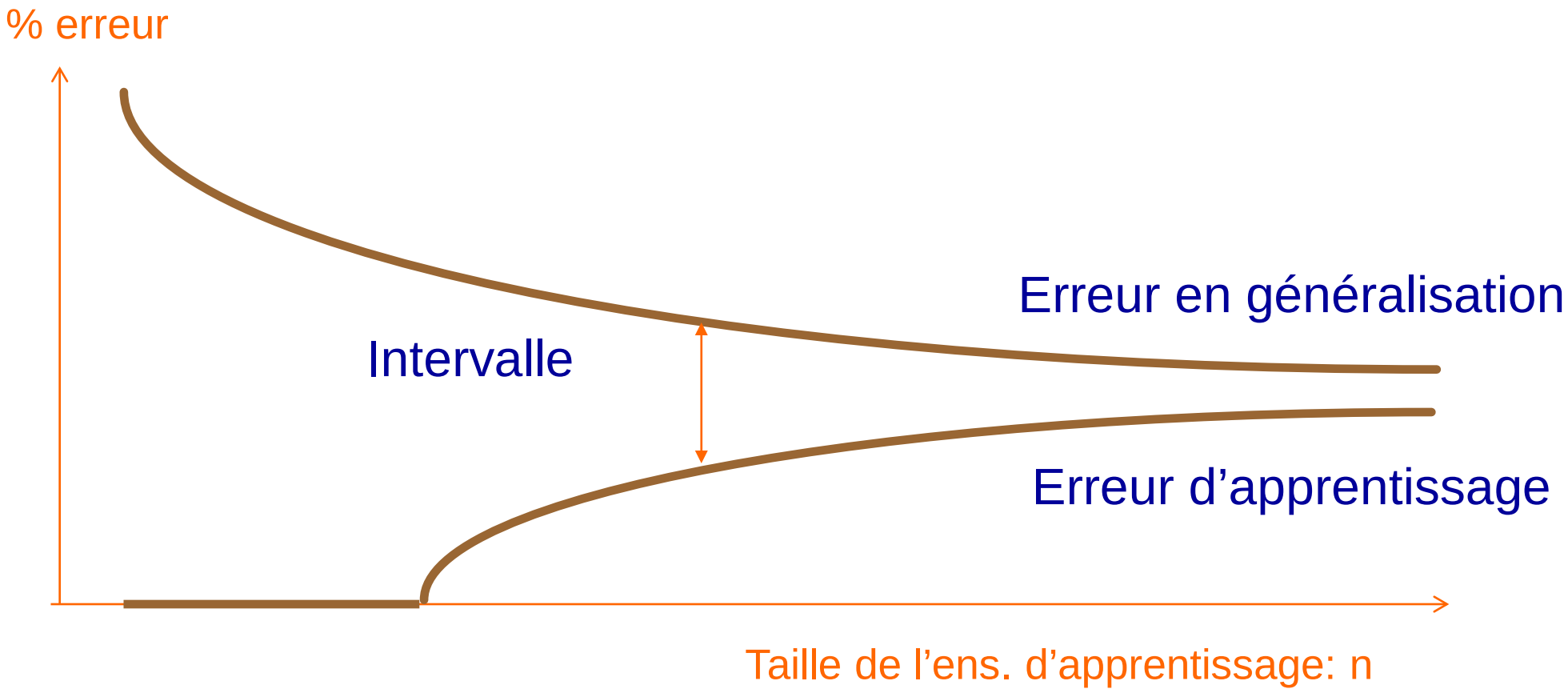
- Même résultat:

$$h \leq \min\left[ent\left(\frac{R^2}{C^2}\right); p \right] + 1$$

Théorème de Vapnik :

- Q : Quelles sont les conditions nécessaires et suffisantes pour assurer la consistence ?
- R : Le processus d'apprentissage est consistent si et seulement si la famille de modèles a une VC dimension finie h
- La VC dimension finie ne garantit pas seulement la généralisation, mais c'est LA SEULE MANIERE qui permet à la généralisation de se produire.

Vitesse de convergence



Vitesse de convergence (2)

- Q : Quelle est la différence entre les erreurs d'apprentissage et de test pour une taille donnée de l'ensemble d'apprentissage ?
- R : La différence entre les erreurs d'apprentissage et de test dépend du rapport entre la VC dimension, h, et la taille de l'ensemble d'apprentissage, n.

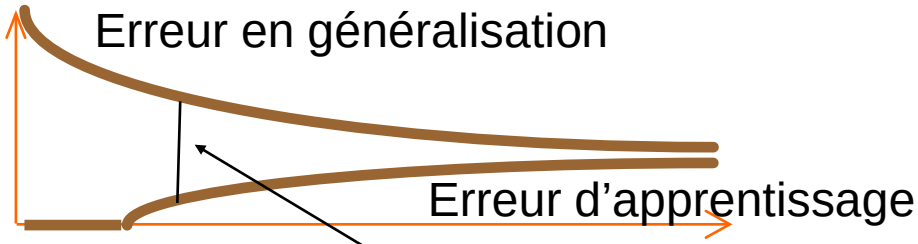
Inégalité de Vapnik

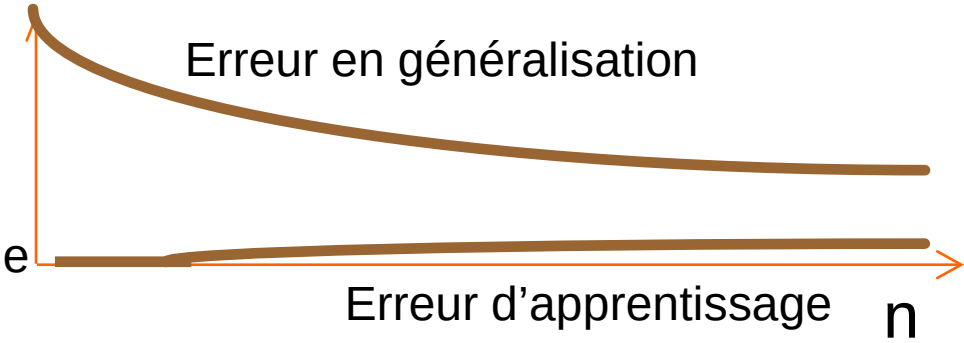
- Avec la probabilité 1- α :

$$R < R_{\text{emp}} + \sqrt{\frac{h\left(\ln\left(2n/h\right)+1\right)-\ln\left(\alpha/4\right)}{n}}$$

  - ne fait pas intervenir p mais la VC dimension h
  - Ne fait pas intervenir la distribution de probabilité P

- Apprentissage consistent
- Apprentissage non consistent



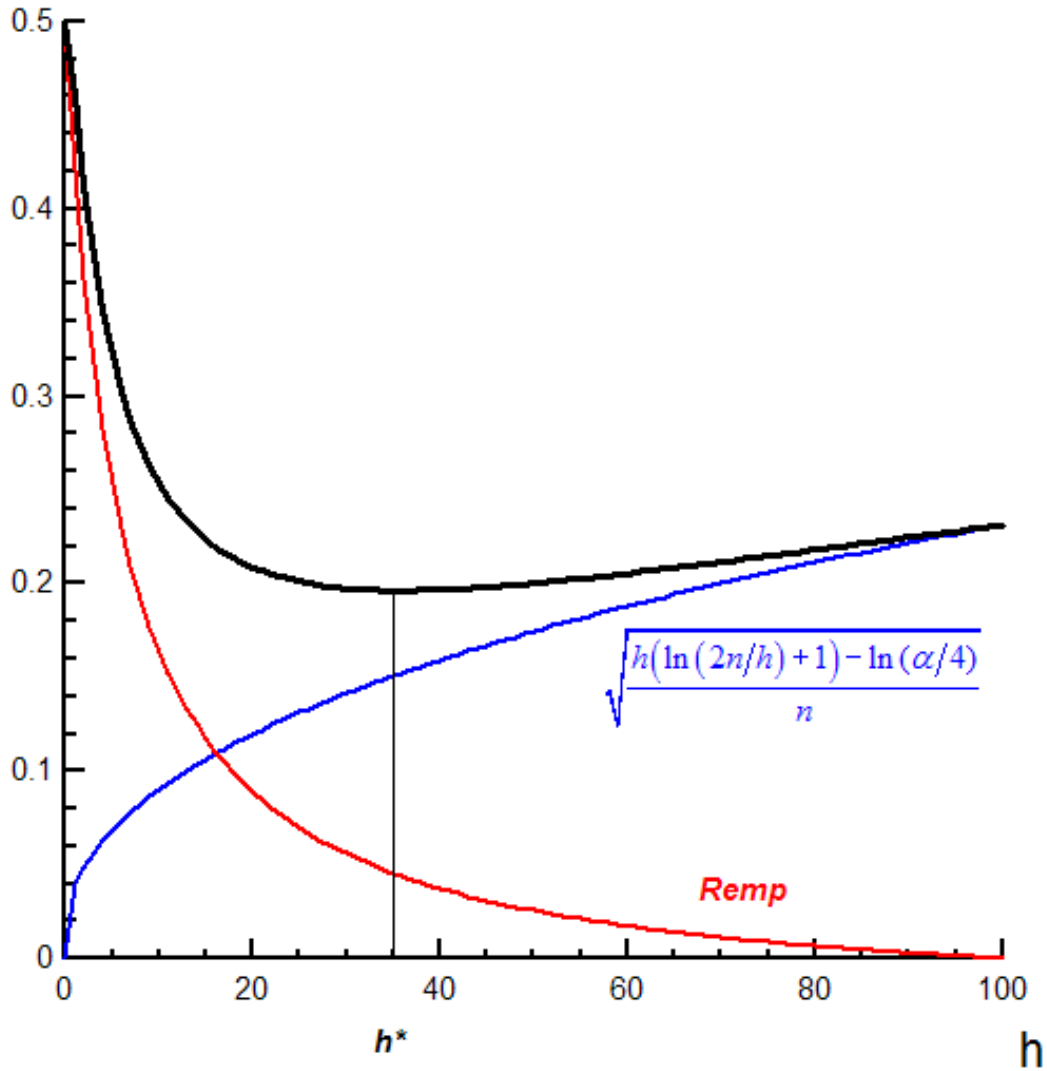


h doit être fini

Inégalité de Vapnik

$$R < R_{\text{emp}} + \sqrt{\frac{h\left(\ln\left(2n/h\right)+1\right)-\ln\left(\alpha/4\right)}{n}}$$

n fixé


0.5
0.4
0.3
0.2
0.1
0
0
20
40
60
80
100
h*
h
$\sqrt{\frac{h(\ln(2n/h)+1)-\ln(\alpha/4)}{n}}$
Remp

De Guillaume d'Ockham à Vapnik

Si deux familles de modèles expliquent les données avec une qualité égale, alors la famille qui a la plus faible VC dimension doit être préférée.

1re découverte: La VC (Vapnik-Chervonenkis) dimension mesure la complexité d'une famille de modèles.

De Guillaume d'Ockham à Vapnik

Si deux modèles expliquent les données avec une qualité égale, alors celui qui provient d'une famille à plus faible VC dimension a une meilleure performance en généralisation.

2ème découverte: La VC dimension peut être reliée à des résultats de généralisation (résultats sur de nouvelles données).

De Guillaume d'Ockham à Vapnik

Pour construire le meilleur modèle à partir de données, il faut tenter d'optimiser à la fois sa performance sur l'ensemble d'apprentissage,
et sa performance de généralisation tirée de la VC dimension : pour ce faire, il faut parcourir une suite de familles d'applications pour y construire ce modèle

3ème découverte: Au lieu d'observer des différences entre des modèles, mieux vaut les contrôler..

Contrôle de la Capacité de Généralisation

$$R < R_{\text{emp}} + \sqrt{\frac{h\left(\ln\left(2n/h\right)+1\right)-\ln(\alpha/4)}{n}}$$

- Risque = Risque d'Apprentissage + Intervalle de Confiance
- Minimiser la seule erreur d'apprentissage ne donnera pas une espérance d'erreur faible (une bonne généralisation)

→ minimiser la <u>somme</u> de l'erreur d'apprentissage et de l'intervalle de confiance.

Principe de minimisation structurée du risque (SRM) (1)

$$R < R_{\text{emp}} + \sqrt{\frac{h\left(\ln\left(2n/h\right)+1\right)-\ln(\alpha/4)}{n}}$$

$$R(w) < E(w) + \sqrt{\frac{h\left(\ln\left(2L/h\right)+1\right)-\ln q}{L}}$$

- lorsque n/h est faible (h trop grand), le deuxième terme est grand
- L'idée générale du SRM est de minimiser la somme des deux termes à la droite de l'inéquation.

Principe de minimisation structurée du risque (SRM)(2)

- Considérons une structure $S_1 \subset S_2 \subset .. \subset S_L$ sur l'ensemble des fonctions vérifiant la propriété

  $h_1 < h_2 < .. < h_L$

- Pour chaque élément S_i de la structure, l'inégalité est valide

$$R < R_{\text{emp}} + \sqrt{\frac{h_i\left(\ln\left(2n/h_i\right)+1\right)-\ln(\alpha/4)}{n}}$$

SRM : Trouver i tel que la somme devienne minimale,

Avec/sans l'approche SRM de Vapnik

- Sans le SRM:
  - Hypothèses sur la distribution statistique (inconnue) des données
  - Un grand nombre de dimensions signifie un modèle à grand nombre de paramètres, ce qui pose des problèmes de généralisation
  - Modéliser revient à chercher le meilleur ajustement
- Avec le SRM:
  - On étudie la famille de modèles, contrôlant sa VC dimension h
  - Le nombre de paramètres peut être très grand, car on contrôle par définition la généralisation
  - Modéliser c'est rechercher le meilleur compromis entre ajustement et robustesse

Borne supérieure trop grande**,**

mais:

Théorème (Devroye, Vapnik) :

Pour toute distribution le SRM fournit la meilleure solution possible avec probabilité 1

(universally strongly consistent)

Contrôle de h

- h doit être fini
- h/n doit être petit: si n augmente, on peut augmenter la complexité du modèle (opposé de BIC)
- h décroit avec:
  - Réduction de dimension (cf. Disqual)
  - La marge (SVM)
  - k en régression ridge
- Mais h difficile à obtenir

Les 3 échantillons:

- Apprentissage: pour estimer les paramètres des modèles
- Test : pour choisir le meilleur modèle
- Validation : pour estimer la performance sur des données futures

- Rééchantillonner: validation croisée, bootstrap

Modèle final: avec toutes les données disponibles

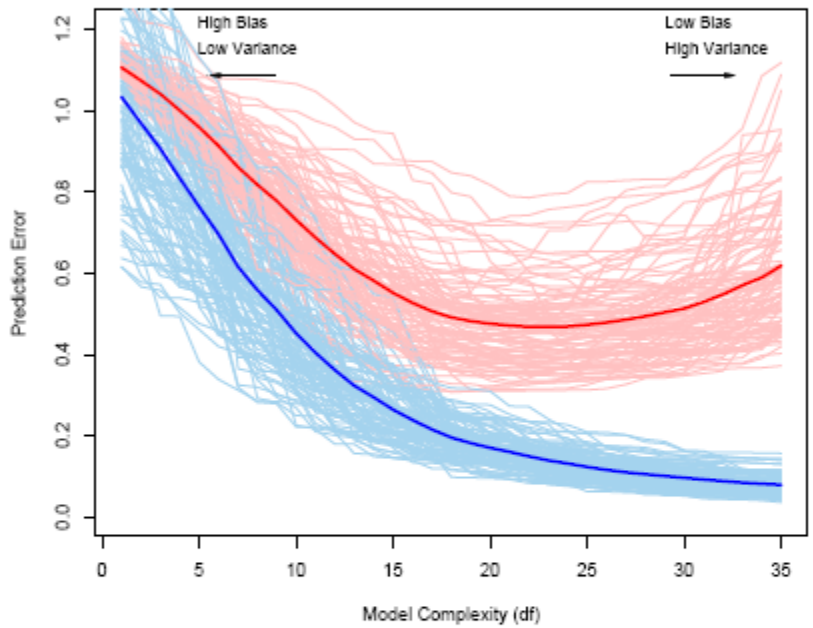


FIGURE 7.1. *Behavior of test sample and training sample error as the model complexity is varied. The light blue curves show the training error* $\overline{\mathrm{err}}$*, while the light red curves show the conditional test error* $\mathrm{Err}_{\mathcal{T}}$ *for* 100 *training sets of size* 50 *each, as the model complexity is increased. The solid curves show the expected test error* Err *and the expected training error* $\mathrm{E}[\overline{\mathrm{err}}]$.

Principes d'induction

- Ne pas chercher à résoudre un problème plus général que nécessaire
  - Ne pas estimer une densité si on veut estimer une fonction
  - Ne pas estimer une fonction si on veut seulement estimer une valeur en un point

Retour sur le rôle des modèles

- Les modèles pour comprendre correspondent à la partie de la statistique considérée comme auxiliaire de la science. Les modèles pour prédire relèvent de la statistique comme méthodologie d'aide à la décision.
- Doit-on opposer science et action ? Une technique qui prédit bien améliore notre connaissance. La modélisation prédictive fait partie de l'empirisme (une théorie de la connaissance à ne pas confondre avec le pragmatisme).