

# Cours IFT6266,

## Introduction à la théorie de l'apprentissage

- Formalisons la notion de **généralisation**:

- Nous avons un ensemble (peut-être ordonné) de données,

$$D_l = (z_1, z_2, \dots, z_l)$$

- Quand l'ordre n'est pas important (ce qui exclut en principe mais pas nécessairement l'apprentissage sur des séries chronologiques), on suppose généralement que ces points sont tirés de manière indépendante de la même **distribution inconnue**  $P(Z)$ . C'est l'hypothèse que les données sont *iid*: identiquement et indépendamment distribuées.
- Nous avons choisi un ensemble de fonctions (i.e., de solutions possibles)  $\mathcal{F}$ .
- On peut mesurer la qualité (ou le coût) d'une solution sur un  $Z = z$  particulier avec une fonction de perte  $L(z, f)$  (où  $f \in \mathcal{F}$ ), que l'on peut appeler notre perte ou *critère local*.
- Le **risque empirique** mesure l'erreur ou perte ou ou coût moyen pour un  $f$  particulier sur un ensemble d'exemples  $D_l$ :

$$\hat{R}(f, D_l) = \frac{1}{l} \sum_{i=1}^l L(z_i, f)$$

- Le **risque espéré** ou **erreur de généralisation** mesure l'erreur espérée avec  $f$  sur un exemple tiré au hasard de  $P(Z)$ :

$$R(f) = \int L(z, f)P(z)dz$$

On ne peut calculer cette quantité, mais c'est celle qui nous intéresse vraiment.

- Le problème de l'**induction** est de trouver  $f \in \mathcal{F}$  qui minimise le risque espéré  $R(f)$ . Il ne s'agit pas simplement d'un problème d'optimisation, car on ne peut mesurer  $R(f)$ ,  $P$  étant inconnu. Cependant, on peut mesurer  $\hat{R}(f, D_l)$ . Notez que  $\hat{R}(f, D_l)$  est une approximation bruitée mais *non-biaisée* de  $R(f)$

(pour un  $f$  quelconque choisi indépendamment de  $D_l$ ), ce qui veut dire que:

$$E_{D_l}[\hat{R}(f, D_l)] = R(f)$$

- Le principe de **minimisation du risque empirique** est à la base d’une grande quantité d’algorithmes d’apprentissage. On cherche

$$f^*(D_l) = \operatorname{argmin}_{f \in \mathcal{F}} \hat{R}(f, D_l).$$

Dans une variante très commune, on minimise l’erreur empirique ainsi qu’une *pénalité* ou terme de régularisation, une fonctionnelle  $H(f)$  qui induit une préférence sur les solutions:

$$f^*(D_l) = \operatorname{argmin}_{f \in \mathcal{F}} \hat{R}(f, D_l) + H(f).$$

- On appelle **erreur d’apprentissage** la plus petite valeur moyenne de perte  $L$  sur les exemples d’apprentissage  $D_l$  obtenue en choisissant  $f^*(D_l)$  dans  $\mathcal{F}$ :

$$\text{erreur d'apprentissage} = \hat{R}(f^*(D_l), D_l) = \min_{f \in \mathcal{F}} \hat{R}(f, D_l).$$

- Est-ce que  $\hat{R}(f^*(D_l), D_l)$  est un estimé non-biaisé de  $R(f^*(D_l))$ ? NON! C’est un estimé OPTIMISTE:

$$E[R(f^*(D_l)) - \hat{R}(f^*(D_l), D_l)] \geq 0$$

Pourquoi? pour la même raison que l’espérance du  $\min_i Z_i$  est systématiquement plus petite que l’espérance d’un  $Z_i$  quelconque. La solution  $f^*(D_l)$  que l’on trouve en minimisant  $\hat{R}(f, D_l)$  est meilleure sur  $D_l$  que sur un autre ensemble tiré au hasard de  $P(Z)$ . En général, quand on choisit un autre ensemble d’exemples  $D'_l$ , on obtient une autre fonction  $f^*(D'_l)$ ; comme  $f^*(D_l)$  et  $f^*(D'_l)$  ne peuvent pas en général être les deux simultanément “correctes” partout où elles diffèrent, elles sont aussi différentes de la fonction “optimale”, qui minimise l’erreur de généralisation dans  $\mathcal{F}$ .

- Est-ce qu'on peut borner la différence entre l'erreur de généralisation et l'erreur d'apprentissage

$$|R(f^*(D_l)) - \hat{R}(f^*(D_l), D_l)|?$$

Oui, sous certaines conditions sur  $\mathcal{F}$ . C'est un des sujets principaux de la théorie de l'apprentissage.

- Cette théorie nous apprend que ces bornes dépendent de ce qu'on appelle la **capacité**  $h(\mathcal{F})$  de  $\mathcal{F}$ , qui est une mesure de sa taille, ou de sa "complexité".
- Dans le cas de problèmes de classification, la capacité  $h(\mathcal{F})$  de  $\mathcal{F}$  est le plus grand  $l$  tel qu'il existe un ensemble d'exemples  $D_l$  tel qu'on puisse toujours trouver un  $f \in \mathcal{F}$  qui donne la bonne réponse sur tous les exemples de  $D_l$ , quel que soit l'étiquetage de ces exemples.
- Est-ce qu'on peut estimer l'erreur de généralisation  $R(f^*(D_l))$ ? Oui, en se basant sur le principe que  $\hat{R}(f^*(D_l), D')$  est un estimé non-biaisé de  $R(f^*(D_l))$  quand  $D'$  est tiré de  $P(Z)$  indépendamment de  $D_l$ . Par exemple, on peut séparer nos données originales en deux sous-ensembles disjoint,  $D_l$  et  $D'$ . On appelle  $D_l$  l'**ensemble d'apprentissage**, et  $D'$  un **ensemble de test**.
- Par la convergence d'une somme de variables aléatoires *iid* vers une Normale, on obtient assez facilement un estimé de la variance de notre estimé d'erreur de généralisation  $\hat{R}(f^*(D_l), D')$  pour  $f = f^*(D_l)$ :

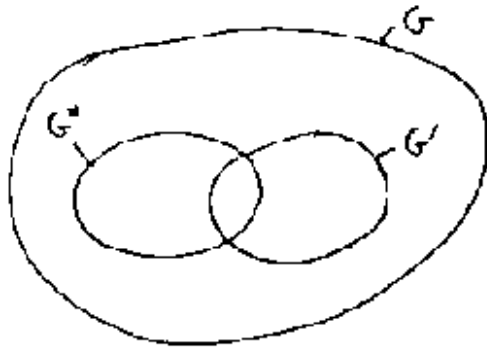
$$\frac{1}{l'} \hat{\sigma}_L^2 \rightarrow \text{Var}[\hat{R}(f, D') | D_l]$$

quand  $l' \rightarrow \infty$ , où  $l' = |D'|$  et  $\hat{\sigma}_L^2$  est un estimé empirique de la variance des pertes:

$$\hat{\sigma}_L^2 = \frac{1}{(l' - 1)} \sum_{z_i \in D'} (L(z_i, f) - \hat{R}(f, D'))^2$$

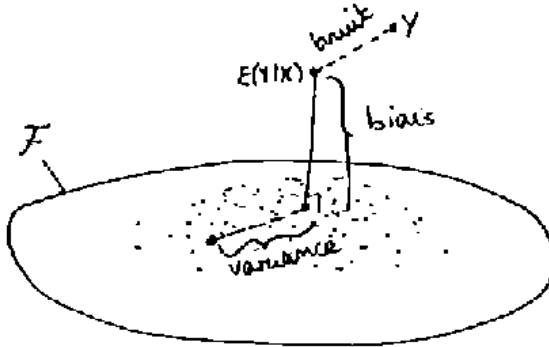
Notez qu'on a écrit que la variance est conditionnelle à  $D_l$  car on suppose  $f$  fixe (alors qu'en fait on a choisi  $f$  en fonction de  $D_l$  et  $D_l$  est une variable aléatoire). On sait que la limite converge très rapidement si la fonction de perte est bornée ou bien si sa distribution a des queues qui décroissent rapidement. Tenir compte de la variabilité due au choix des exemples d'entraînement est plus difficile et sera traité ailleurs. Un autre problème est que beaucoup de données ne sont pas réellement *iid* (mais certaines formes de correction sont alors possibles).

- L’approche classique en statistique est généralement basée sur des hypothèses assez précises sur la véritable distribution des données ( $P$ ), et cette approche est souvent **asymptotique**: on étudie ce qui arrive quand  $l \rightarrow \infty$ . Dans ce cours, nous étudieront des approches où on ne suppose pas une forme pour  $P$  pour faire des inférences (on peut par contre chercher  $f$  dans une classe de distributions, sans supposer qu’elle contient la vraie). Puisqu’on ne peut pas faire d’inférences de type “test d’hypothèse” basées sur une hypothèse sur  $P$ , on va faire nos inférences de manière empirique, en faisant des estimés empiriques de l’erreur de généralisation.
- Références fondamentales dans le cas où  $l$  est fini: les trois livres de Vladimir Vapnik [Vapnik, 1982, Vapnik, 1995, Vapnik, 1998].
- Exemples de fonctions de coût:
  - Apprentissage supervisé:  $f$  est une fonction de  $\mathcal{X}$  vers  $\mathcal{Y}$ ,  $Z = (X, Y) \in \mathcal{X} \times \mathcal{Y}$ .
    - \* Classification:  $Y \in$  un ensemble fini, et  $L(f, z) = 0$  si  $f(x) = y$ , et 1 sinon.
    - \* Régression:  $Y \in \mathcal{R}^n$ , et par exemple  $L(f, z) = (f(x) - y)^2$ . On peut montrer que  $f^*(D)(x)$  estime  $E[Y|x]$ .
  - Estimation de densité:  $f(z) > 0$  et  $\int f(z)dz = 1$ ,  $L(f, z) = -\log f(z)$ . Il n’y a pas d’entrée et de sortie désirée: apprentissage non-supervisé.
- Combinaison de contraintes provenant des exemples et des connaissances a priori:



- Considérons un grand ensemble de solutions  $\mathcal{G}$ .
- Un sous-ensemble  $\mathcal{G}' \subset \mathcal{G}$  est “compatible” avec les exemples  $D_l$  (ses éléments donnent une erreur  $\hat{R}(f, D_l)$  “petite”).
- Un sous-ensemble  $\mathcal{G}'' \subset \mathcal{G}$  est “compatible” avec les connaissances a priori (ses éléments sont plausibles selon ces connaissances).
- Nous combinons les deux contraintes en choisissant une solution dans  $\mathcal{G}' \cap \mathcal{G}''$ .

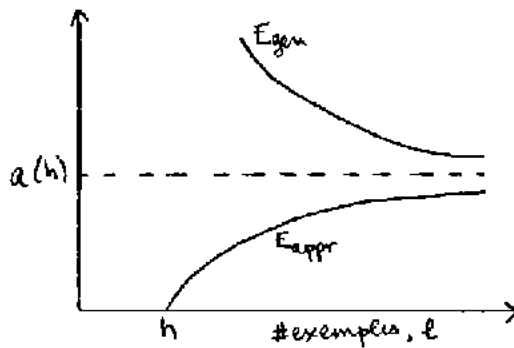
- Des connaissances a priori nous permettent donc de réduire l'ensemble des solutions considérées: elles sont équivalentes à une certaine quantité d'exemples (et vice-versa).
- On peut décomposer l'erreur d'apprentissage en 3 éléments:



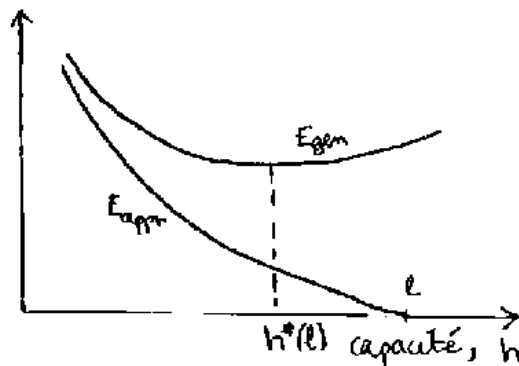
1. l'erreur due au fait que l'ensemble de fonctions choisi,  $\mathcal{F}$ , ne contient pas nécessairement LA solution optimale au problème; on appelle parfois cette partie de l'erreur **le biais**, car l'espérance de la fonction choisie sur tous les ensembles d'apprentissage ne sera pas la fonction globalement optimale.
  2. l'erreur due au fait que  $l < \infty$  et donc  $f^*$  n'est pas nécessairement la meilleure fonction dans  $\mathcal{F}$  (elle minimise  $\hat{R}(f, D_l)$  mais pas nécessairement  $R(f)$ ); on appelle parfois cette partie de l'erreur **la variance**, car elle provient de la variabilité entre différents ensembles d'apprentissage possibles tirés de  $P$ .
  3. l'erreur due au "bruit", i.e., au fait que même la solution optimale se "trompe" parfois. Par exemple, si on fait de l'apprentissage supervisé et qu'on cherche une fonction de  $X$  vers  $Y$ , il n'existe pas de solution "parfaite" quand la relation entre  $X$  et  $Y$  n'est pas une fonction (i.e., pour chaque valeur de  $X$  il y a plusieurs valeurs de  $Y$  possibles).
- Nous ne pouvons rien pour réduire l'erreur due au bruit intrinsèque. Le **dilemme biais-variance** exprime le conflit entre le désir de réduire l'erreur de biais et celui de réduire l'erreur de variance. Quand on augmente la capacité  $h(\mathcal{F})$  de  $\mathcal{F}$ , l'erreur de biais diminue, mais l'erreur de variance augmente. Cette décomposition biais + variance + bruit a un sens mathématique clair dans le cas de la régression, où on peut effectivement décomposer l'erreur quadratique espérée ainsi (à un  $X = x$  particulier):

$$\begin{aligned}
 E_D[(f^*(D)(x) - Y)^2] = & (E[Y|x] - Y)^2 && \text{"bruit"} \\
 & + (E_D[f^*(D)(x)] - E[Y|x])^2 && \text{"biais"} \\
 & + E_D[(f^*(D)(x) - E_D[f^*(D)(x)])^2] && \text{"variance"}
 \end{aligned}$$

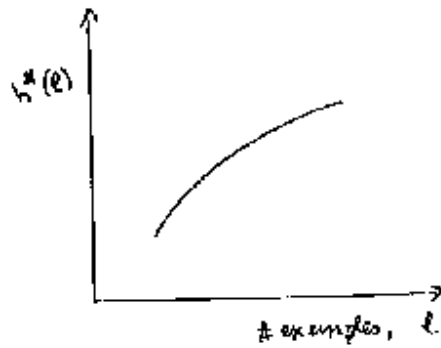
- Courbes d'apprentissage (analyse qualitative):



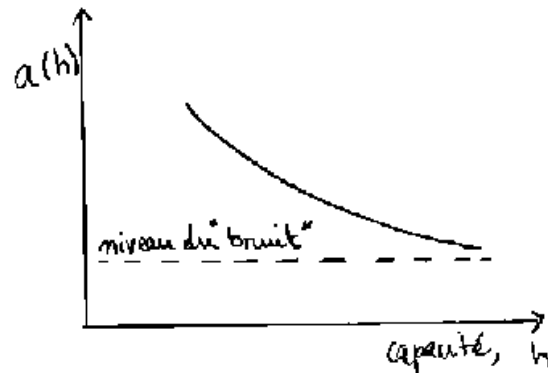
- Pour  $h$  fixe, quand le nombre d'exemples  $l$  augmente, l'erreur d'apprentissage  $\hat{R}(f^*(D_l), D_l)$  augmente, et l'erreur de généralisation  $R(f^*(D_l))$  diminue. Si l'algorithme d'apprentissage se comporte bien (on dit qu'il est *consistent*, qu'il converge vers la meilleure solution dans  $\mathcal{F}$ ), alors les deux courbes se rejoignent asymptotiquement à une valeur  $a(h)$  qui dépend de la capacité  $h$ . Une **capacité finie** est une condition nécessaire et suffisante pour obtenir cette convergence (dans la classe des algorithmes qui minimisent l'erreur empirique).



- Pour  $l$  fixe, quand la capacité  $h(\mathcal{F})$  de  $\mathcal{F}$  augmente,  $\hat{R}(f^*(D_l), D_l)$  diminue.
- Quand la capacité  $h(\mathcal{F})$  de  $\mathcal{F}$  augmente,  $R(f^*(D_l))$  commence par diminuer, puis atteint un minimum (à la **capacité optimale**  $h_{\text{opt}}$ ) et finalement remonte.  
À gauche, pour  $h < h_l^*$ , on dit que l'algorithme “*underfit*” (il ne colle pas assez aux données), alors qu'à droite, pour  $h > h_l^*$ , on dit qu'il “*overfit*” (il colle trop aux données).



- La capacité optimale dépend de  $l$ : au fur et à mesure que  $l$  augmente, la capacité optimale augmente. Plus on a d'exemples, plus on peut se permettre un modèle complexe.



- Au fur et à mesure que la capacité  $h$  augmente, l'erreur asymptotique  $a(h)$  diminue, jusqu'à atteindre le niveau de bruit intrinsèque au problème.

## References

- [Vapnik, 1982] Vapnik, V. (1982). *Estimation of Dependences Based on Empirical Data*. Springer-Verlag, Berlin.
- [Vapnik, 1995] Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Springer, New-York.
- [Vapnik, 1998] Vapnik, V. (1998). *Statistical Learning Theory*. Wiley, Lecture Notes in Economics and Mathematical Systems, volume 454.