Prérequis :

- Proba, calcul d'espérance
- loi gaussienne
- Algèbre linéaire, calcul matriciel.

Surapprendre : est-ce une si mauvaise idée ?

Cadre d'étude de l'apprentissage supervisé :

(X, Y) variables aléatoires.

X : entrée / descripteur / caractéristiques $\in \mathbb{R}^d$

Y : sortie / cible / étiquette / label

Modèle : on suppose que l'entrée et la sortie sont liées de la manière suivante.

$$Y = f^*(X) + \varepsilon$$

↳ ε est un terme de bruit **ll** de $\mathbb{E}$ tq $\mathbb{E}(\varepsilon) = 0$.

↳ f^* caractérise le lien entre X et Y : en particulier si on connaissait f^* on pourrait prédire Y à partir de la simple connaissance de X.

- Qd Y est une variable à valeurs réelles, on parlera de régression

  (ex) Y = nb de clics sur un lien
  X = description de l'utilisateur, âge, dernier type de page consulté, cb de fs, ...

  (ex) Y = prix d'une action en bourse

  (ex) Y = rendement d'un plant de céréales

- Qd Y est une variable à valeurs discrètes et finies, on parlera de classification

  (ex) Y = spam / non-spam
  X = contenu de l'email

  Y = 1 si le tableau a été peint par Picasso.
  X = image du tableau.

  Y = 1 si une transaction est frauduleuse.

  } classif° binaire

(ex) Y = chiffre manuscrit

Y = expression du visage (joie, tristesse, colère, ...)
X = image faciale

} classif° multi-classe

But : faire la prédiction de Y sachant X

ou encore estimer f par une fct $\hat{f}$ tq $\hat{f} \simeq f^*$

ou encore apprendre f.

Comment construire $\hat{f}$?

En apprentissage supervisé, on apprend à { faire des prédictions / construire $\hat{f}$ }

à partir d'exemples étiquetés, càd accompagnés de la valeur que l'on voudrait prédire. On a donc accès à un échantillon / des observations

$(X_1, Y_1), \dots, (X_n, Y_n)$ iid

parmi lesquelles les étiquettes $(Y_i)_{i \in n}$ vont jouer le rôle de "professeur" et vont venir "superviser" l'apprentissage de f^* par $\hat{f}$.

On pourra qualifier les ds° $(X_1, Y_1), \dots, (X_n, Y_n)$ { d'exemples d'entraînement. / de jeu d'apprentissage. }

Nous allons donc construire $\hat{f}$ à partir du jeu d'apprentissage tq

$Y_i \simeq \hat{f}(X_i)$ $\quad i = 1, \dots, n$.

càd tq l'erreur d'entraînement soit faible

Mais le véritable espoir est que $Y_{new} \simeq \hat{f}(X_{new})$ où (X_{new}, Y_{new}) des données que $\hat{f}$ n'a pas { "vu" / "utilisé" } lors de l'entraînement, càd que $\hat{f}$ généralise bien la prédiction à des nouvelles données.

En régression $(Y \in \mathbb{R})$, la "vraie" fonction f peut être interprétée comme

$$f^* \in \underset{f:\mathbb{R}^d \to \mathbb{R}}{\operatorname{argmin}} \underbrace{\mathbb{E}_{(X,Y)}\left[(Y - f(X))^2\right]}_{=:R(f)}$$

On appelle cette quantité le risque du prédicteur f, càd l'erreur moyenne quadratique commise par f sur des observations (Y, X).

Or la loi sous-jacente de X et Y est inconnue.

Donc le risque d'un prédicteur n'est pas calculable. On peut encore ⊖ le minimiser pour trouver f^*!

A la place, on va chercher à minimiser le risque empirique calculable à partir du jeu d'entraînement : étant donnés $(X_1, Y_1) \dots (X_n, Y_n)$

$$\hat{R}(f) := \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(X_i))^2.$$

⚠ On ne va pas chercher n'importe quelle fct qui minimise le risque empirique, on va se donner une classe de prédicteurs possibles. Pourquoi ? On pourrait construire la fct $f : \mathbb{R}^d \to \mathbb{R}$

$$x \mapsto \begin{cases} Y_i & \text{qd } x = X_i \\ 10^9 & \text{sinon.} \end{cases}$$

↳ elle a un risque empirique nul !

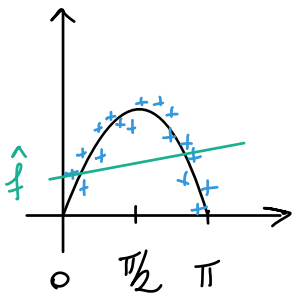
↳ elle fera certainement de très mauvaises prédictions pour toute nouvelle donnée.

On parle de sur-apprentissage qd le prédicteur ne sait pas séparer l'information du bruit : l'erreur en test explose !

Un modèle qui sur-apprend est généralement un modèle trop complexe qui "colle" trop aux données et capture le bruit.

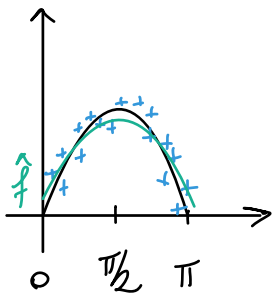
A l'inverse on parlera de **sous-apprentissage** qd le prédicteur est trop simple par rapport aux données.

Ex $X_i \in [0;\pi]$ $Y_i \in \mathbb{R}$ $Y_i = f^*(X_i) + \varepsilon_i$ $f^* = \sin$



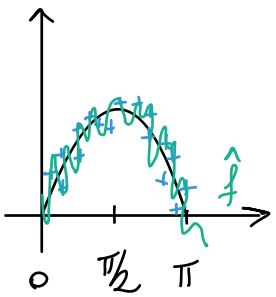
Sous-apprentissage

On contraint $\hat{f}$ à être une fct affine. On trouve la "meilleure fct affine"



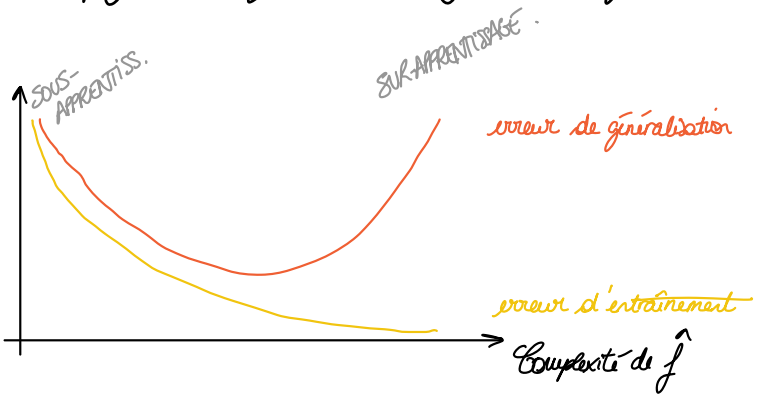
Bon compromis

polynôme de deg 4.



Sur-apprentissage

polynôme de deg 15.



C'est un **schéma classique** d'un 1er cours de ML. Pourtant, aujourd'hui des prédicteurs de + en + complexes et entraînés (réseaux de neurones) et leur performance s'améliore tjs. On va creuser plus en détail ce phénomène pour un modèle en particulier : le **modèle linéaire**, càd f est une fct affine des entrées X et ce qui jouera le rôle de la "**complexité**" du prédicteur, ce sera la **dimension** d de l'espace des entrées. Nous verrons finalement que le 1er graphique n'est qu'un bout de l'histoire.

Hyp : $Y_i = X_i^T \theta^* + \varepsilon_i$ pour $\theta^* \in \mathbb{R}^d$

HELP ON GAUSSIAN ?

$X_i \in \mathbb{R}^d$ chaque composante est $\mathcal{N}(0,1)$ iid.

$\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$

↳ Pouvez-vous me dire $\mathbb{E}[XX^T] \overset{?}{=} Id_d$.

On sait que $\theta^* = \underset{\theta}{\operatorname{argmin}}\ \mathbb{E}\left[(Y - X^T\theta)^2\right]$ pour (X,Y) copie de (X_i, Y_i).

$\forall \theta \quad \mathbb{E}\left[(Y - X^T\theta)^2\right] \geq \mathbb{E}\left[(Y - X^T\theta^*)^2\right]$

en effet $\mathbb{E}\left[(Y - X^T\theta)^2\right] = \mathbb{E}\left[(Y - X^T\theta^* + X^T\theta^* - X^T\theta)^2\right]$

$= \mathbb{E}\left[(Y - X^T\theta^*)^2\right] + \underbrace{\mathbb{E}\left[(X^T\theta^* - X^T\theta)^2\right]}_{\geq 0} + 2\,\mathbb{E}\left[\ \right]$ $= 0$

Pour estimer θ^*, on va chercher à minimiser le risque empirique

$$\hat{\theta} \in \underset{\theta}{\operatorname{argmin}}\ \frac{1}{n}\sum_{i=1}^{n} (Y_i - X_i^T\theta)^2 = \frac{1}{n}\|Y - X\theta\|_2^2$$

Fct cvx en θ : si min local $\Rightarrow$ min global.
Fct diff en θ : pt critique $\Rightarrow$ min.

$$\nabla \hat{R}(\theta) = \begin{pmatrix} \frac{\partial \hat{R}(\theta)}{\partial \theta_1} \\ \vdots \\ \frac{\partial \hat{R}(\theta)}{\partial \theta_d} \end{pmatrix} = \frac{1}{n}\, 2\, X^T (X\theta - Y)$$

$= 0 \iff$ **$X^T X \theta = X^T Y$**

$d < n$ **Régime sous-paramétré** : plus d'équations que d'inconnues.

$X^T X$ est inversible.

$\hat{\theta} = (X^T X)^{-1} X^T Y$.

$$\mathcal{R}(\hat{\theta}) = \mathbb{E}\left[(X^T\hat{\theta} - Y)^2\right]$$
$$= \mathbb{E}\left[(X^T\hat{\theta} - X^T\theta^*)^2\right] + \mathbb{E}\left[(X^T\theta^* - Y)^2\right].$$

$$\Rightarrow \mathcal{R}(\hat{\theta}) - \mathcal{R}(\theta^*) = \mathbb{E}\left[(X^T\hat{\theta} - X^T\theta^*)^2\right].$$

$$= \mathbb{E}\left[\langle X, \hat{\theta} - \theta^*\rangle^2\right]$$
$$= \mathbb{E}\left[(\hat{\theta} - \theta^*)^T X X^T (\hat{\theta} - \theta^*)\right]$$
$$= (\hat{\theta} - \theta^*)^T \underbrace{\mathbb{E}\left[XX^T\right]}_{Id.} (\hat{\theta} - \theta^*)$$
$$= \|\hat{\theta} - \theta^*\|_2^2$$

$$\mathbb{E}_{\mathcal{D}_n}\left[\mathcal{R}(\hat{\theta}) - \mathcal{R}(\theta^*)\right] = \mathbb{E}_{\mathcal{D}_n}\left[\|(X^TX)^{-1}X^TY - \theta^*\|_2^2\right]$$
$$= \mathbb{E}_{\mathcal{D}_n}\left[\|(X^TX)^{-1}X^TX\theta^* + (X^TX)^{-1}X^T\varepsilon - \theta^*\|_2^2\right]$$
$$= \mathbb{E}_{\mathcal{D}_n}\left[\|(X^TX)^{-1}X^T\varepsilon\|_2^2\right]$$
$$= \mathbb{E}_{\mathcal{D}_n}\left[Tr\left(\varepsilon^T X (X^TX)^{-1}(X^TX)^{-1}X^T\varepsilon\right)\right]$$
$$= \sigma^2\,\mathbb{E}\left[Tr\left((X^TX)^{-1}\right)\right]$$

$X \in \mathbb{R}^{n\times d}$ est une matrice gaussienne

On sait que $X^TX \in \mathbb{R}^{d\times d}$ a une distribut° de Wishart ac n ddl :

- presque sûrement inv.

Donc

$$\mathbb{E}_{\mathcal{D}_n}\left(\mathcal{R}(\hat{\theta}) - \mathcal{R}(\theta^*)\right) = \begin{cases} \dfrac{\sigma^2 d}{n-d-1} & \text{si } n \geq d+2 \\ \infty & \text{si } n = d \text{ ou } n = d+1. \end{cases}$$

$n \leq d$ "Régime surparamétré"

Il existe une infinité d'ERM, ds ce cas le risque empirique est nul dès lors que $Y = X\theta$,
i.e. $\forall i,\ Y_i = X_i^T \theta$.
On va se concentrer sur une solution en particulier, celle qui est de norme ℓ^2 min :

$$\hat{\theta} \in \operatorname{argmin} \|\theta\|_2^2 \text{ s.t } Y = X\theta \qquad (P)$$

i.e. $\hat{\theta} = X^T (XX^T)^{-1} Y$

On peut tjs réécrire (P) ĉ $\operatorname{argmin} \|\hat{\theta} + h\|_2^2$ s.t $h \in \operatorname{Ker} X$
Or $\hat{\theta} \in \operatorname{Im} X^T = (\operatorname{Ker} X)^\perp$ de $\operatorname{argmin} \|\hat{\theta}\|_2^2 + \|h\|_2^2$ s.t $h \in \operatorname{Ker} X$.
$\Rightarrow h = 0$!!

$$\mathbb{E}_{\mathcal{D}_n}\left[R(\hat{\theta}) - R(\theta^*)\right] = \mathbb{E}_{\mathcal{D}_n}\left[\|\hat{\theta} - \theta^*\|_2^2\right]$$

$$= \mathbb{E}_{\mathcal{D}_n}\left[\|X^T(XX^T)^{-1}Y - \theta^*\|_2^2\right]$$

$$= \mathbb{E}_{\mathcal{D}_n}\left[\|X^T(XX^T)^{-1}X\theta^* + X^T(XX^T)^{-1}\varepsilon - \theta^*\|_2^2\right]$$

$$= \mathbb{E}_{\mathcal{D}_n}\left[\|\underbrace{(X^T(XX^T)^{-1}X - I)}_{\text{Matrice de projection sur } \operatorname{Ker} X \text{ ou signe près.}}\theta^*\|_2^2\right] + \underbrace{\mathbb{E}_{\mathcal{D}_n}\left[\|X^T(XX^T)^{-1}\varepsilon\|_2^2\right]}$$

$$\mathbb{E}\left[\|X^T(XX^T)^{-1}\varepsilon\|_2^2\right] = \mathbb{E}\left[\varepsilon^T(XX^T)^{-1}XX^T(XX^T)^{-1}\varepsilon\right]$$

$$= \mathbb{E}\left[\mathrm{Tr}\left(\text{——}\right)\right]$$

$$= \sigma^2\,\mathbb{E}\left[\mathrm{Tr}\left((XX^T)^{-1}\right)\right]$$

$$= \frac{\sigma^2 n}{d-n+1} \quad \text{si } d \geq n+2.$$

$$\mathbb{E}\left[(\theta^*)^T\left(I - X^T(XX^T)^{-1}X\right)\theta^*\right] = \mathbb{E}\left[\|\mathrm{Proj}_{(\mathrm{Im}X^T)^\perp}(\theta^*)\|_2^2\right]$$

$$= \mathbb{E}\left[\|\mathrm{Proj}_{\ker X}(\theta^*)\|_2^2\right]$$

↗ matrice de projection donc idempotente !

Focus sur $\mathbb{E}\left[(\theta^*)^T X^T(XX^T)^{-1}X\theta^*\right]$:

On pose la rotation $R^{(\ell)}$ tq $\theta^* = \|\theta^*\|_2 R^{(\ell)} e_\ell$, alors

$$= \|\theta^*\|_2^2\,\mathbb{E}\left[e_\ell^T R^{(\ell)T} X^T(XX^T)^{-1}X R^{(\ell)} e_\ell\right]$$

$$= \|\theta^*\|_2^2\, e_\ell^T\,\mathbb{E}\left[(XR^{(\ell)})^T\left(XR^{(\ell)}R^{(\ell)T}X^T\right)^{-1}(XR^{(\ell)})\right]e_\ell.$$

par orthogonalité de $R^{(\ell)}$

OR la matrice aléatoire $XR^{(\ell)}$ est de même loi que X, on peut s'en convaincre :

$$(X_1, \dots, X_d)\,R^{(\ell)} = (U_1, \dots, U_d)$$

avec $U_j = (X_1, \dots, X_d)\,R^{(\ell)}_{\cdot j} = \sum_{k=1}^{d} X_k R^{(\ell)}_{kj}$

C'est une combinaison linéaire de gaussiennes de U_j est de loi gaussienne $\mathbb{E}[U_j] = \mathbb{E}\left[\sum_k X_j R_{kj}^{(\ell)}\right] = 0$

$$\mathrm{Var}[U_j] = \sum_{k=1}^{d} R_{kj}^{2(\ell)} \underbrace{\mathrm{Var}(X_j)}_{1} = \|R_{:j}^{(\ell)}\|_2^2 = 1 .$$

Donc finalement

$$= \|\theta^*\|_2^2 \, e_\ell^T \, \mathbb{E}\left[X^T (XX^T)^{-1} X\right] e_\ell . \quad \forall \ell$$

$$\mathbb{E}\left[(\theta^*)^T X^T (XX^T)^{-1} X \theta^*\right] = \frac{\|\theta^*\|_2^2}{d} \sum_{\ell=1}^{d} e_\ell^T \mathbb{E}\left[X^T (XX^T)^{-1} X\right] e_\ell$$

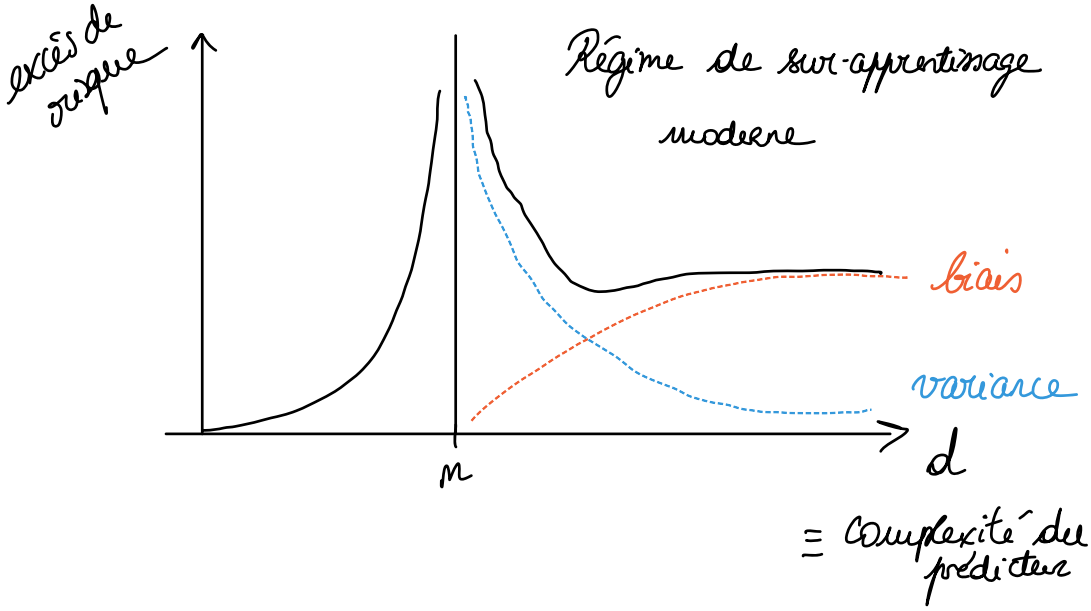
$$= \frac{\|\theta^*\|_2^2}{d} \mathbb{E}\left[\mathrm{Tr}\left(X^T (XX^T)^{-1} X\right)\right]$$

$$= \frac{\|\theta^*\|_2^2}{d} \mathbb{E}\left[\mathrm{Tr}(Id_n)\right]$$

$$= \frac{\|\theta^*\|_2^2}{d} n .$$

Finalement, $\mathbb{E}\left[(\theta^*)^T \left(I - X^T (XX^T)^{-1} X\right) \theta^*\right] = \|\theta^*\|_2^2 \, \frac{d-n}{d}$

$$\mathbb{E}_{\mathcal{D}_n}\left[R(\hat{\theta}) - R(\theta^*)\right] = \begin{cases} \sigma^2 \dfrac{n}{d-n+1} + \|\theta^*\|_2^2 \dfrac{d-n}{d} & \text{si } d \geq n+2 \\ +\infty & \text{si } \begin{cases} d = n+1 \\ d = n \end{cases} \end{cases}$$



Ds le régime de sur-apprentissage moderne, on voit que le risque n'explose pas qd $d \nearrow$, finalement on ne fait pas n'importe quoi qd on est à droite du graphique.
Evidemment ceci est vrai pour un minimiseur du risque empirique particulier, et pas n'importe lequel -

↳ Le biais augmente en d/n : qd $d > n$, $\hat{\theta}$ vit dans le "row space" de X, espace de dim n.

↳ La variance décroît en d/n ds le cas surparamétré

Pour des prédicteurs plus complexes tq les réseaux de neurones, l'entraînement revient à minimiser une fct non-convexe, donc il faut encore comprendre vers quel minimiseur du risque empirique l'entraînement nous amène et si celui-ci a de bonnes propriétés de généralisation.