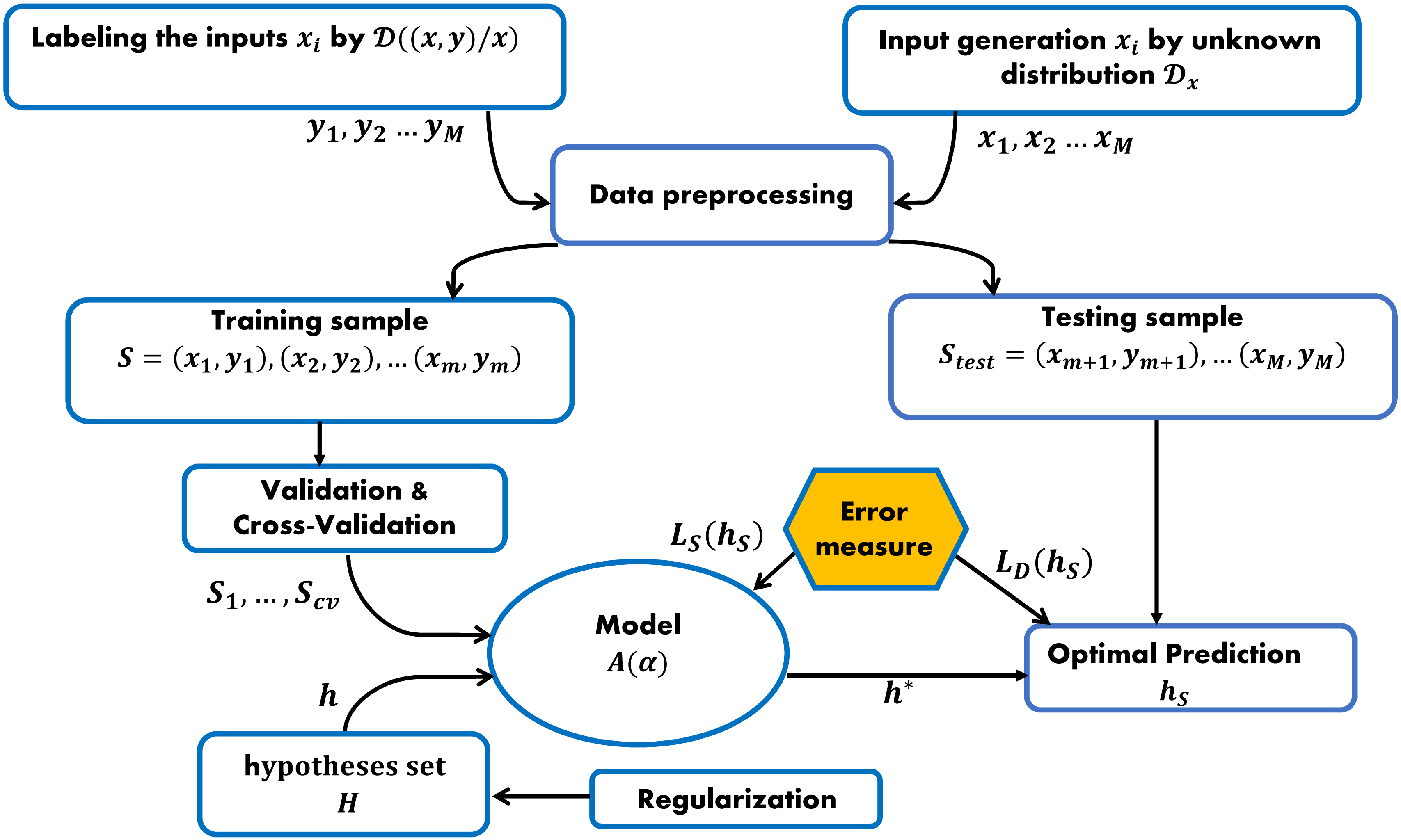
Part 1: Machine learning theory

1. Learning framework
2. Uniform convergence
3. Learnability of infinite size hypotheses set
4. Tradeoff Bias/Variance $E_S(L_D(h_S)) = \mathbf{Bais} + \mathbf{Variance} + \mathbf{bruit}$
    1. General Error Decomposition: Regression case.
    2. Bias-Variance Tradeoff.
        - Complexity of H.
        - Complexity of S.
    3. Bias-Variance Estimation: Bootstrap Replicate.
5. Validation/Cross-Validation
6. Regularization

Machine learning algorithm



Recall

The general form of error:

Let l be a cost function, such that: $l: H \times Z \longrightarrow \mathbb{R}^+$ and $Z = X \times Y$

The general error of h:

$$L_{\mathcal{D}}(h) = \underset{z \sim \mathcal{D}}{\mathrm{E}}[l(h,z)]$$

The empirical error of h:

$$L_S(h) = \frac{1}{m}\sum_{i=1}^{m} l(h, z_i)$$

Classification	Regression
$l(h,z) = \begin{cases} 1 \; si \; h(x) \neq y \\ 0 \; si \; h(x) = y \end{cases}$ with: $z = (x,y) \in Z = X \times \{0,1\}$ This function is also valid for the multinomial classification.	$l(h,z) = (h(x) - y)^2$ with: $z = (x,y) \in Z = X \times \mathbb{R}^+$

Motivation

Objective:

How can we estimate the general error?

Tool:

Bias-Variance decomposition.

- $\boldsymbol{L_D(h_S) \ll L_S(h_S)}$ ***overfitting***

- $\boldsymbol{E_S(L_D(h_S)) = \mathbf{Bais} + \mathbf{Variance} + Noise}$

ε_t is white noise if

$$\mathbf{E}(\boldsymbol{\varepsilon_t}) = \mathbf{0}\ \boldsymbol{Var(\varepsilon_t) = \sigma^2 = Cte},\ \boldsymbol{cov(\varepsilon_t, \varepsilon_{t+1}) = 0}$$

4.1. General Error Decomposition: Regression Case

Let h_S be the hypothesis selected by ERM_H.

We assume in what follow:

$$y = f(x) + \varepsilon(x)$$

Such that ε is a centered white noise.

In the regression case, to gauge the distance between $h_S(x)$ and y, we use the quadratic error:

$$l(h_S, y) = (h_S(x) - y)^2$$

So:

$$L_S(h_S) = \frac{1}{m}\sum_{i=1}^{m} l(h_S, y_i) = \frac{1}{m}\sum_{i=1}^{m} (h_S(x_i) - y_i)^2$$

On the other hand, we have:

$$L_D(h_S) = \underset{(x,y)\sim D}{\mathrm{E}}[l(h_S, y)] = \underset{(x,y)\sim D}{\mathrm{E}}[(h_S(x) - y)^2\,]$$

4.1. General Error Decomposition: Regression Case

$$E_S[L_D(h_S)] = E_S\left[\underset{(x,y)\sim D}{\mathrm{E}}[(h_S(x) - y)^2\,]\right]$$

$$= E_S\left[\underset{(x,y)\sim D}{\mathrm{E}}\left[\left(h_S(x) - f(x) - \varepsilon(x)\right)^2\,\right]\right]$$

$$= \underset{(x,y)\sim D}{\mathrm{E}}\left[E_S\left[\left(h_S(x) - f(x) - \varepsilon(x)\right)^2\,\right]\right]$$

So:

$$\boldsymbol{E_S[L_D(h_S)] = \underset{(x,y)\sim D}{E}\left[E_S[(h_S(x) - \bar{h}(x))^2]\,\right] + \underset{(x,y)\sim D}{E}\left[\left(\bar{h}(x) - f(x)\right)^2\right] + \underset{(x,y)\sim D}{E}\left[\left(\varepsilon(x)\right)^2\right]}$$

4.1. General Error Decomposition: Regression Case

$E_S[h_S(x)]$ is the estimation of a discrete random variable h_S that takes the following values $\{h_{S_1}, \dots, h_{S_k}\}$. So:

$$\boldsymbol{\bar{h}(x) = E_S[h_S(x)] = \sum_{j=1}^{k} h_{S_j} P(h_S = h_{S_j})}$$

Such that, $h_{S_1}, \dots, h_{S_k}$ are respectively generated by the samples $S_1, S_2, \dots, S_k$ trained by the same algorithm $\boldsymbol{A_\alpha}$, such that S_j have the same probability :

$$\boldsymbol{\bar{h}(x) = E_S[h_S(x)] = \frac{1}{k}\sum_{j=1}^{k} h_{S_j}(x)}$$

4.1. General Error Decomposition: Regression Case

Finally:

$$
\begin{aligned}
\boldsymbol{Bias(x)} &= \underset{(x,y)\sim D}{\boldsymbol{E}}\left[\left(\bar{\boldsymbol{h}}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{x})\right)^{2}\right] \\
\boldsymbol{Variance(x)} &= \underset{(x,y)\sim D}{\boldsymbol{E}}\left[\boldsymbol{E_S}[(\boldsymbol{h_S}(\boldsymbol{x}) - \bar{\boldsymbol{h}}(\boldsymbol{x}))^2]\right] \\
(\boldsymbol{x}) &= \underset{(x,y)\sim D}{\boldsymbol{E}}[\left(\boldsymbol{\varepsilon}(\boldsymbol{x})\right)^{2}]
\end{aligned}
$$

4.1. General Error Decomposition: Regression Case

$$\boldsymbol{Bias}(\boldsymbol{x}) = \underset{(\boldsymbol{x},\boldsymbol{y})\sim \boldsymbol{D}}{\boldsymbol{E}}\left[\left(\overline{\boldsymbol{h}}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{x})\right)^{\boldsymbol{2}}\right]$$

Definition: Bias

The bias measures the deviation between the hypothesis that we expect learn $\bar{h}$ throughout S and the target function f.

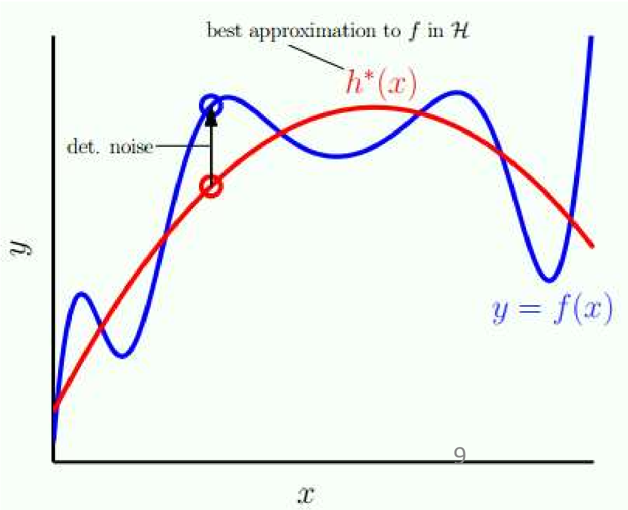
It is also named deterministic noise/error.

$$\boldsymbol{Bias}(\boldsymbol{x}) = \frac{\boldsymbol{1}}{\boldsymbol{m}}\sum_{\boldsymbol{i}=\boldsymbol{1}}^{\boldsymbol{m}}\left(\left(\overline{\boldsymbol{h}}(\boldsymbol{x}_{\boldsymbol{i}}) - \boldsymbol{f}(\boldsymbol{x}_{\boldsymbol{i}})\right)\right)^2$$

It describes the best model's error.

We want to learn: $\boldsymbol{y} = \boldsymbol{h}(\boldsymbol{x})$

But, we learn: $\boldsymbol{y} = \boldsymbol{h}(\boldsymbol{x}) + \boldsymbol{bruit\ d\acute{e}terministe}$



4.1. General Error Decomposition: Regression Case

$$\boldsymbol{Variance}(\boldsymbol{x}) = \underset{(\boldsymbol{x},\boldsymbol{y})\sim\boldsymbol{D}}{\boldsymbol{E}}\left[\boldsymbol{E}_{\boldsymbol{S}}[(\boldsymbol{h}_{\boldsymbol{S}}(\boldsymbol{x}) - \overline{\boldsymbol{h}}(\boldsymbol{x}))^2]\right]$$

Definition: Variance

The variance measures the deviation between the final hypothesis h_S and the hypothesis that we expect to learn $\bar{h}$.

It describes how much h_S variates from one training set S to another.

$$\boldsymbol{Variance}(\boldsymbol{x}) = \frac{\mathbf{1}}{\boldsymbol{m}}\sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{m}}\frac{\mathbf{1}}{\boldsymbol{k}}\sum_{\boldsymbol{j}=\mathbf{1}}^{\boldsymbol{k}}(\boldsymbol{h}_{\boldsymbol{S}_{\boldsymbol{j}}}(\boldsymbol{x}_{\boldsymbol{i}}) - \overline{\boldsymbol{h}}(\boldsymbol{x}_{\boldsymbol{i}}))^2$$

It measures the model’s instability.

4.1. General Error Decomposition: Regression Case

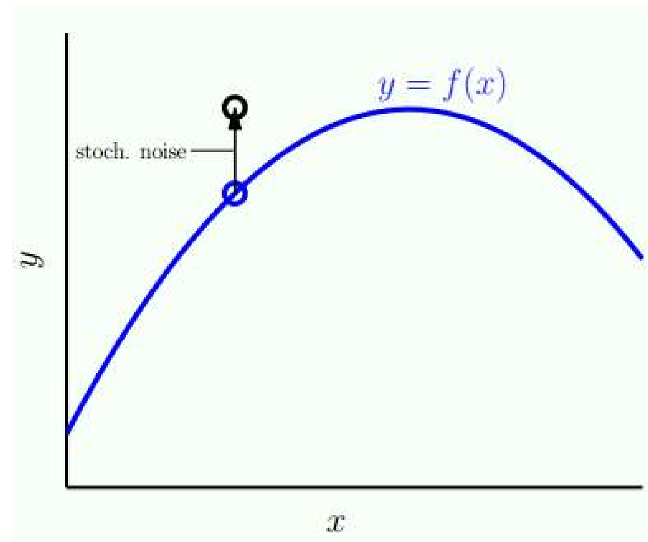
$$\boldsymbol{Noise(x)} = \underset{(x,y)\sim D}{\boldsymbol{E}}\left[\left(\boldsymbol{\varepsilon(x)}\right)^2\right]$$

Definition: Noise

The noise measures the deviation between the unknown target function f and the measured value y.

It describes the variance between y and f.

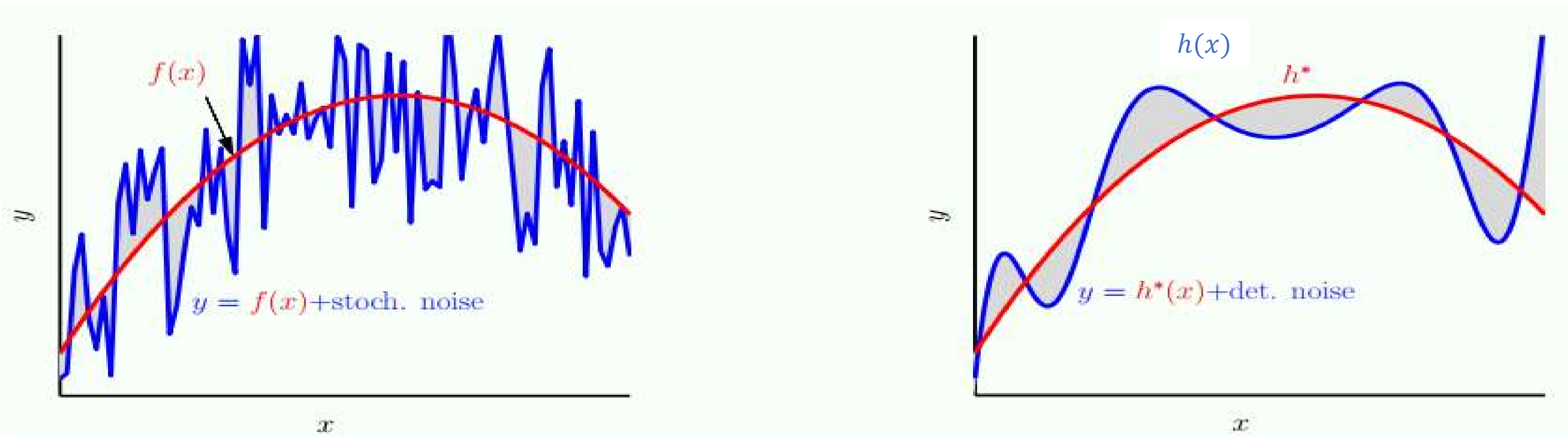
It is also named stochastic noise/error.



We wants to learn: $\boldsymbol{y = f(x)}$

But, we observe: $\boldsymbol{y = f(x) + bruit\ stochastique}$

4.1. General Error Decomposition: Regression Case



- Source: random measures.

If we measure y another time:

➢ Stochastic error changes.

If we change H:

➢ Stochastic error remains the same.

- Source: H cannot model f.

If we measure y another time:

➢ Deterministic error remains the same.

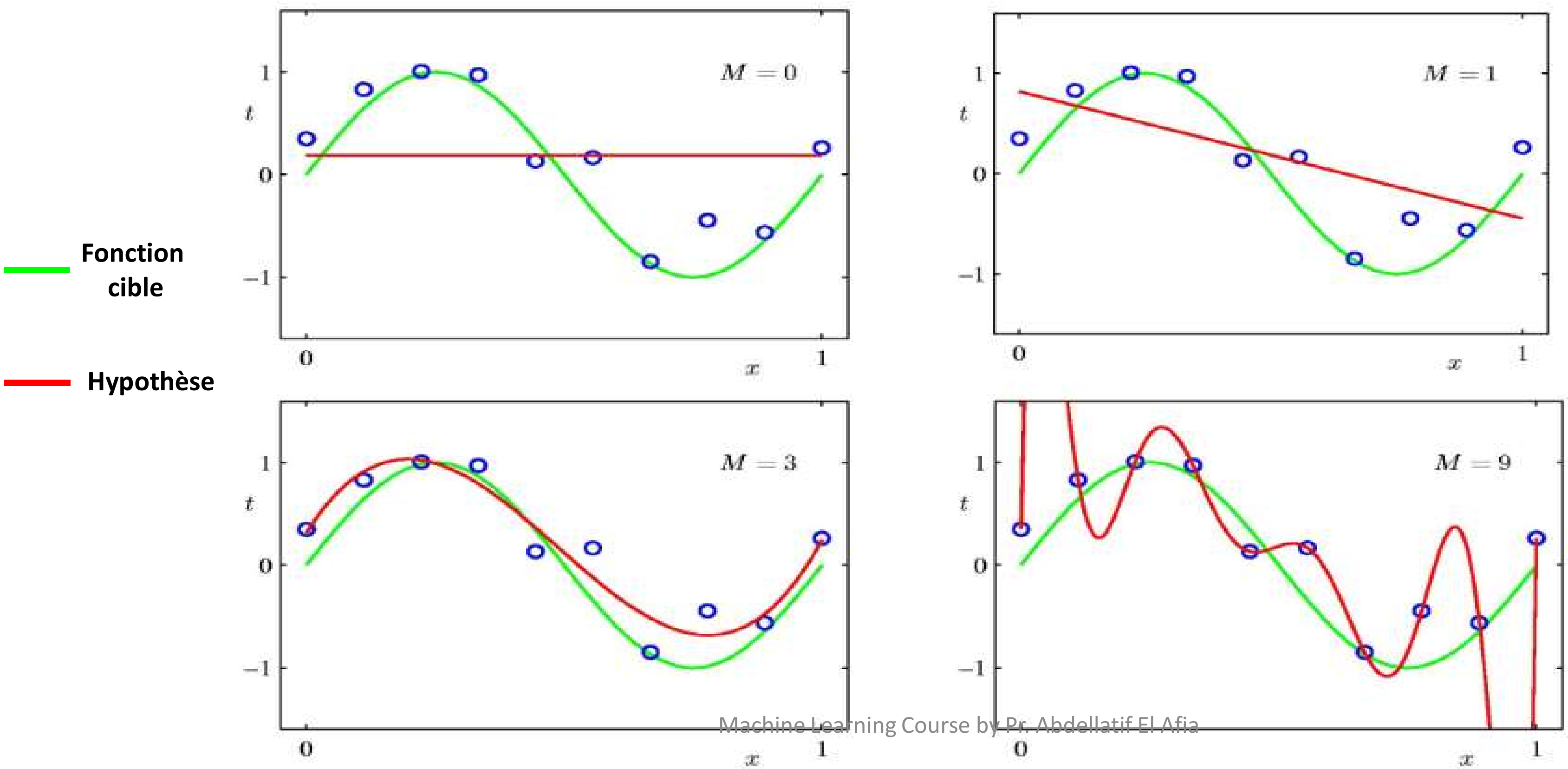
If we change H:

➢ Deterministic error changes.

4.2. Tradeoff Bias-Variance: Complexity of H

Complexity of H – Interpretation

- What is the best hypothesis for these data?



4.2. Tradeoff Bias-Variance: Complexity of H

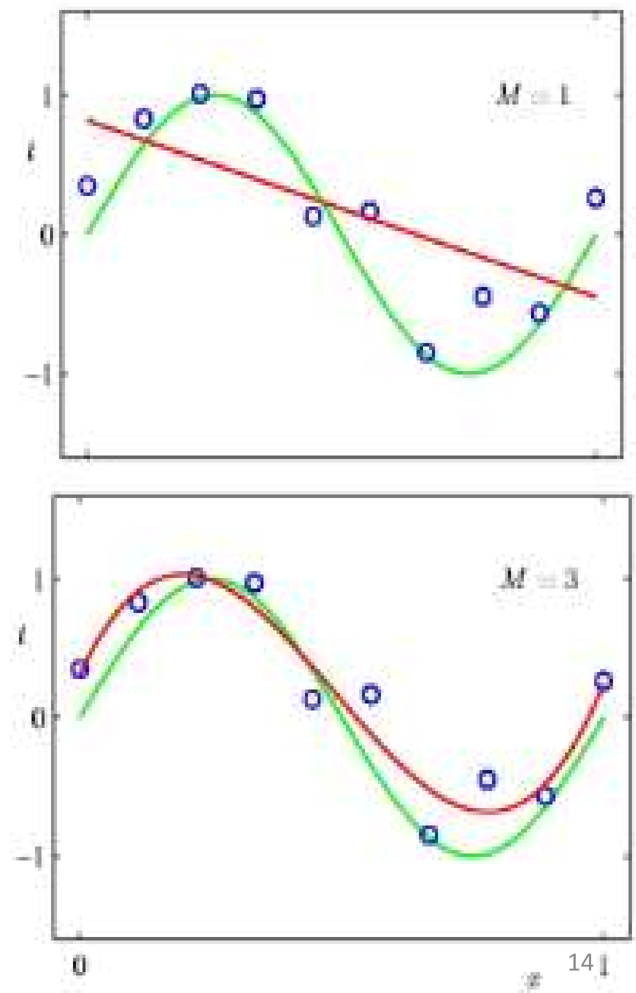
Complexity of H – Bias Interpretation

For a fixed size of data points:

- Simple hypothesis:
  - ➢ Small Polynomial degree.
  - ➢ High bias.

- Complex hypothesis:
  - ➢ strong Polynomial degree.
  - ➢ Low bias.

The Bias disappears when we select the perfect model.

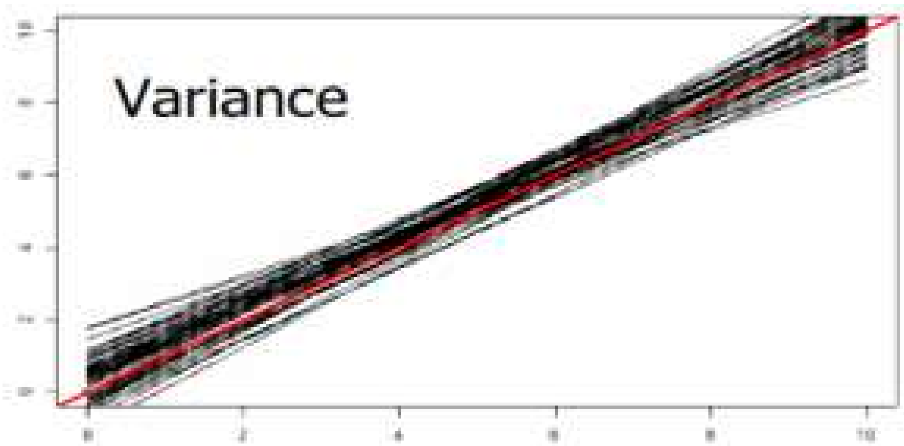


4.2. Tradeoff Bias-Variance: Complexity of H

Complexity of H – Variance Interpretation

We notice that the variance doesn't have a direct dependence on the real model.
Pour un nombre fixe de données:

- Simple hypothesis:
  - ➢ Small Polynomial degree.
  - ➢ low variance (high model's stability).



- Complex hypothesis:
  - ➢ strong Polynomial degree.
  - ➢ High variance (low model's stability).

The variance disappears when $|S| \rightarrow \infty$.

4.2. Tradeoff Bias-Variance: Complexity of H

Complexité de H –Types of H

There exist two characterizations of H:

Characterization 1:

- **Flexible H:** has a low bias and strong variance.
- **Rigid H:** has a strong bias and low variance.
- **Optimal H:** has a balance between bias/variance.

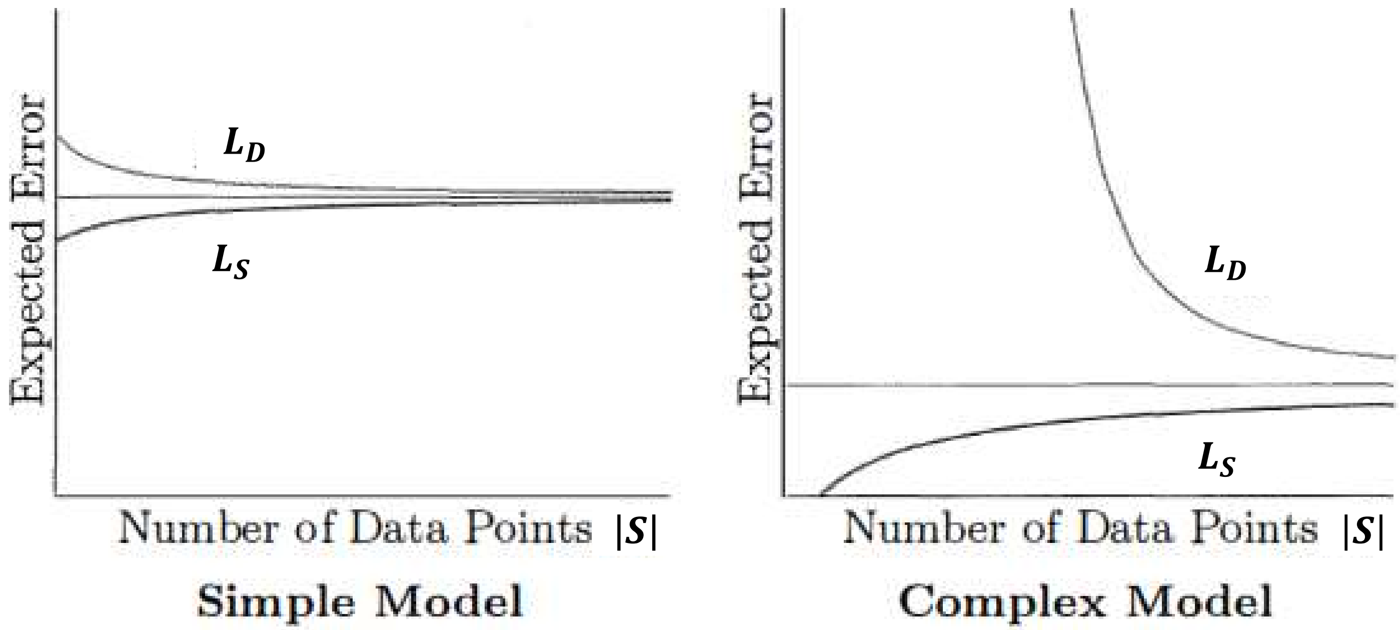
Characterization 2:

- **Simple H:** contains few parameters.
- C**omplexe H:** contains many parameters.

4.2. Tradeoff Bias-Variance: Complexity of S

Complexity of S – Learning curve

- It is assumed that the complexity of H is fixe.

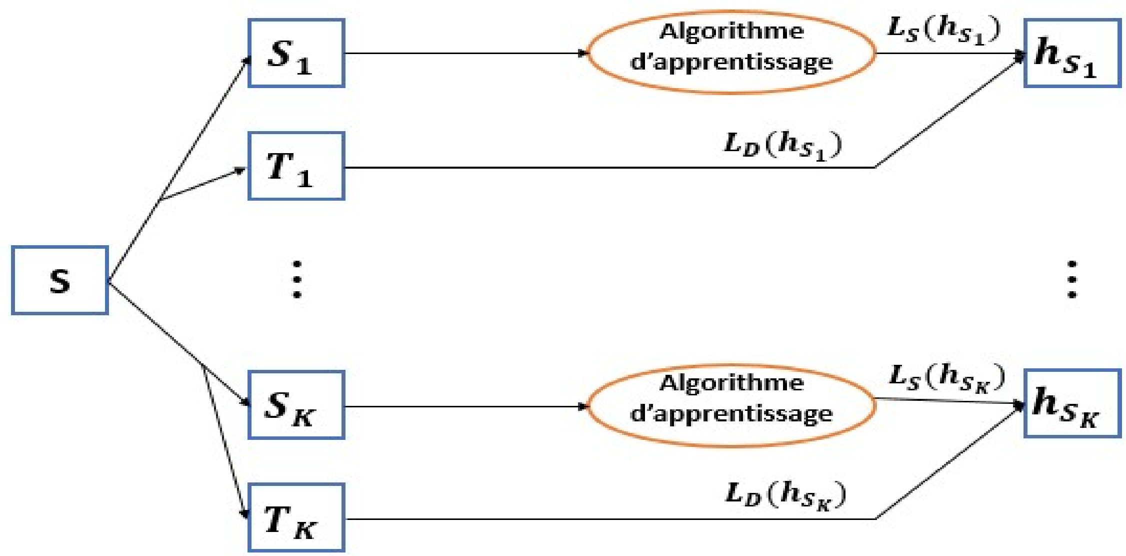


In the simple model, the learning curves converge more quickly but to worse performance than for the complex model.

4.3. Bias-Variance Estimation: Bootstrap Replicate

Definition:

Bootstrap Replicate is a technique for estimating the bias and the variance of learning.



The main hypothesis is:

$$\bar{h} = the\ most\ frequent\ hypothesis\ in\ H$$

We consider that the noise is null: $y = f(x)$