

# Aprenentatge Automàtic 1

**GCED**

Lluís A. Belanche

belanche@cs.upc.edu



Soft Computing Research Group  
Dept. de Ciències de la Computació (Computer Science)  
Universitat Politècnica de Catalunya

2019-2020

**LECTURE 3: Regression theory and linear regression models (I)**

# Theoretical issues for regression

## Outline

1. The regression framework
2. Bias-Variance analysis
3. Empirical and Structural risk minimization

# Theoretical issues for regression

## The regression framework

Given data  $\mathcal{D} = \{(\mathbf{x}_n, t_n)\}_{n=1, \dots, N}$ , where  $\mathbf{x}_n \in \mathbb{R}^d, t_n \in \mathbb{R}$ ,

**Statistics:** estimation of a continuous random variable (r.v.)  $T$  conditioned on a random vector  $\mathbf{X}$

**Mathematics:** estimation of a real function  $f$  based on a finite number of “noisy” examples  $(\mathbf{x}_n, f(\mathbf{x}_n))$

The departing **statistical setting** is  $t_n = f(\mathbf{x}_n) + \varepsilon_n$ ; a **model** is any approximation of  $f$

---

$\varepsilon_n$  are i.i.d. continuous r.v. such that  $\mathbb{E}[\varepsilon_n] = 0$  and  $\text{Var}[\varepsilon_n] = \sigma^2 < \infty$

# Theoretical issues for regression

## The regression framework

The **risk** of a model  $y$  is

$$R(y) := \int_{\mathbb{R}} \int_{\mathbb{R}^d} L(t, y(\mathbf{x})) p(t, \mathbf{x}) d\mathbf{x} dt$$

where  $L$  is a suitable **loss** function:

- $L(t, y(\mathbf{x})) \geq 0$
- $L(t, y(\mathbf{x})) = 0$  if  $t = y(\mathbf{x})$
- $L(t, y(\mathbf{x}))$  does not increase when  $|t - y(\mathbf{x})|$  decreases

related to the distribution of the  $\varepsilon_n$  (the “noise model”)

# Theoretical issues for regression

## The regression framework

Let us step firm ground and assume that  $\varepsilon_n \sim N(0, \sigma^2)$  (implications?)

Using a **Maximum Likelihood** argument, it can be shown that the “right” loss is the **square error**:

$$L_{SE}(t, y(\mathbf{x})) := (t - y(\mathbf{x}))^2$$

The **risk** is therefore

$$R(y) = \int_{\mathbb{R}} \int_{\mathbb{R}^d} (t - y(\mathbf{x}))^2 p(t|\mathbf{x}) p(\mathbf{x}) d\mathbf{x} dt$$

# Theoretical issues for regression

## The regression framework

If we enjoy complete freedom to choose  $y$ , the solution is:

$$y^*(\mathbf{x}) = \int_{\mathbb{R}} t p(t|\mathbf{x}) dt = f(\mathbf{x})$$

known as the **regression function**.

Since  $\mathbb{E}[\varepsilon_n] = 0$ , we can alternatively express the regression setting by stating that  $t$  is a continuous r.v. such that  $f(\mathbf{x}) = \mathbb{E}[t|\mathbf{X} = \mathbf{x}]$ .

$$\implies f = y^*$$

# Theoretical issues for regression

## The regression framework

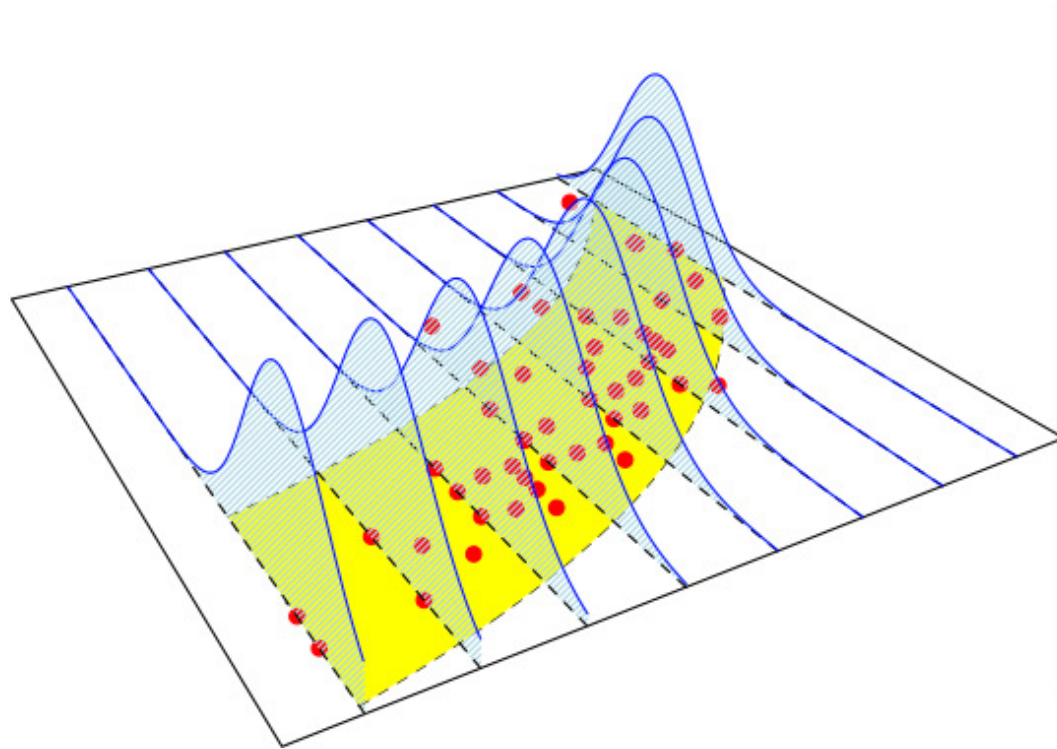


Illustration of the standard assumptions  
(normality, homoscedasticity)

# Theoretical issues for regression

## The regression framework

In a practical setting, we do not know  $p(t|\mathbf{x})$  ...

- Instead, we have a finite i.i.d. **data sample** of  $N$  labelled observations  $\mathcal{D} = \{(\mathbf{x}_n, t_n)\}_{n=1, \dots, N}$ , where  $\mathbf{x}_n \in \mathbb{R}^d, t_n \in \mathbb{R}$
- Intuition (?) is telling us to solve for  $y$  in:

$$\int_{\mathbb{R}^d} \left( f(\mathbf{x}) - y(\mathbf{x}) \right)^2 p(\mathbf{x}) d\mathbf{x}$$

- We must impose restrictions on the possible solutions  $y$  (a specific **class of functions**)



# Theoretical issues for regression

## The regression framework

We can compute an approximation to the true risk, called the **empirical risk**, by averaging the loss function on the available data  $\mathcal{D}$ :

$$R_{\text{emp}}(y) := \frac{1}{N} \sum_{n=1}^N (t_n - y(x_n))^2$$

(this quantity is also known as the **training**, resubstitution or apparent **error**)

The **Empirical Risk Minimization** (ERM) principle states that a learning algorithm should choose a hypothesis (model)  $\hat{y}$  which minimizes the empirical risk among a predefined class of functions  $\mathcal{Y}$ :

$$\hat{y} := \arg \min_{y \in \mathcal{Y}} R_{\text{emp}}(y)$$

# Theoretical issues for regression

## The regression framework

The quantity  $R_{\text{emp}}(\hat{y})$  is known as the **training error**.

In **theoretical ML**, we are very much interested in:

1. how this error fluctuates as a function of  $\mathcal{D}$
2. how far this error is from the true error, *i.e.*, to bound  $|R_{\text{emp}}(\hat{y}) - R(y)|$ ; at the very least, to bound  $|\mathbb{E}[R_{\text{emp}}(\hat{y})] - R(y)|$
3. how far this error is from the best possible error, *i.e.*, to bound  $|R_{\text{emp}}(\hat{y}) - R(y^*)|$ ; at the very least, to bound  $|\mathbb{E}[R_{\text{emp}}(\hat{y})] - R(y^*)|$

# Theoretical issues for regression

## Bias-Variance analysis

Recall the assumption that  $\varepsilon_n \sim N(0, \sigma^2)$  ...

In this case (using the square error), the risk can be decomposed as:

$$\begin{aligned} R(y) &= \int_{\mathbb{R}} \int_{\mathbb{R}^d} (t - y(\mathbf{x}))^2 p(t, \mathbf{x}) d\mathbf{x} dt \\ &= \int_{\mathbb{R}} \int_{\mathbb{R}^d} (t - f(\mathbf{x}))^2 p(t, \mathbf{x}) d\mathbf{x} dt \\ &\quad + \int_{\mathbb{R}^d} (f(\mathbf{x}) - y(\mathbf{x}))^2 p(\mathbf{x}) d\mathbf{x} \\ &= \sigma^2 + \int_{\mathbb{R}^d} (f(\mathbf{x}) - y(\mathbf{x}))^2 p(\mathbf{x}) d\mathbf{x} =: \sigma^2 + \text{MSE}(y) \end{aligned}$$

where  $f$  is the **regression function**.

# Theoretical issues for regression

## Bias-Variance analysis

Therefore we arrive at  $R(y) = \sigma^2 + \text{MSE}(y)$

We can now “forget” about  $\sigma^2$  and the risk and minimize instead the MSE “to the last bullet”:

$$\text{MSE}(y) = \int_{\mathbb{R}^d} \left( f(\mathbf{x}) - y(\mathbf{x}) \right)^2 p(\mathbf{x}) d\mathbf{x}$$

---

A **learning algorithm** for **regression** is a procedure that, given  $\mathcal{D}$  and  $\mathcal{Y}$ , outputs a model  $y_{\mathcal{D}} \in \mathcal{Y}$  that aims to minimize  $\text{MSE}(y)$ .

# Theoretical issues for regression

## Bias-Variance analysis

- Consider now one particular  $x_0$ : different  $\mathcal{D}$  will produce different  $y_{\mathcal{D}}$  and therefore different predictions  $y_{\mathcal{D}}(x_0)$  ...
- Let us concentrate on the quantity  $\left(f(x_0) - y_{\mathcal{D}}(x_0)\right)^2$
- We wish to eliminate the dependence on  $\mathcal{D}$ ; therefore we investigate its expected value:

$$\mathbb{E}_{\mathcal{D}}\left[\left(f(x_0) - y_{\mathcal{D}}(x_0)\right)^2\right], \quad \text{taken over all possible } \mathcal{D} \text{ of size } N$$

# Theoretical issues for regression

## Bias-Variance analysis

$$\begin{aligned}\mathbb{E}_{\mathcal{D}}\left[\left(f(\mathbf{x}_0) - y_{\mathcal{D}}(\mathbf{x}_0)\right)^2\right] &= \\ &\quad \left(f(\mathbf{x}_0) - \mathbb{E}_{\mathcal{D}}\left[y_{\mathcal{D}}(\mathbf{x}_0)\right]\right)^2 \\ &\quad + \\ &\quad \mathbb{E}_{\mathcal{D}}\left[\left(y_{\mathcal{D}}(\mathbf{x}_0) - \mathbb{E}_{\mathcal{D}}\left[y_{\mathcal{D}}(\mathbf{x}_0)\right]\right)^2\right]\end{aligned}$$

$$\Rightarrow \text{MSE}(y_{\mathcal{D}}(\mathbf{x}_0)) = \left(\text{Bias}(y_{\mathcal{D}}(\mathbf{x}_0))\right)^2 + \text{Var}(y_{\mathcal{D}}(\mathbf{x}_0))$$

---

$$R(y_{\mathcal{D}}(\mathbf{x}_0)) = \sigma^2 + \left(\text{Bias}(y_{\mathcal{D}}(\mathbf{x}_0))\right)^2 + \text{Var}(y_{\mathcal{D}}(\mathbf{x}_0))$$

# Theoretical issues for regression

## Bias-Variance analysis

The prediction risk at any given point  $x_0$  is the sum of three components:

**The noise variance:** variability of the target value around its conditional mean

**The (squared) bias:** average (square) deviation of our prediction at  $x_0$  and the best possible prediction

**The variance:** variability of our prediction as a function of the used data sample (regardless of the underlying function!)

# Theoretical issues for regression

## Bias-Variance analysis

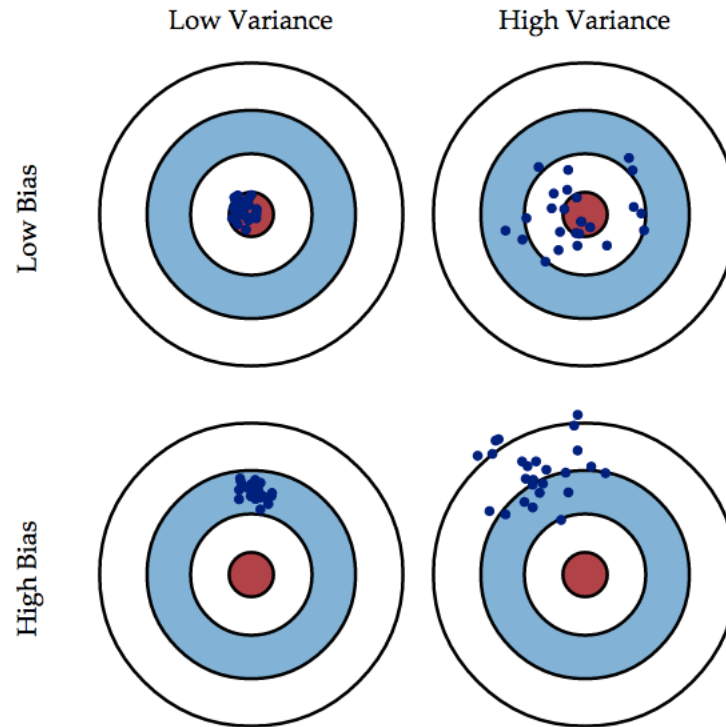


Illustration of the **Bias-Variance decomposition** using a dartboard



# Theoretical issues for regression

## Bias-Variance analysis

The derivation above depends on a particular point  $x_0$  ... let us put it back in place (*i.e.*, within their integrals):

$$\left(Bias(y_{\mathcal{D}})\right)^2 = \int_{\mathbb{R}^d} \left(Bias(y_{\mathcal{D}}(x))\right)^2 p(x) dx$$

$$Var(y_{\mathcal{D}}) = \int_{\mathbb{R}^d} Var(y_{\mathcal{D}}(x)) p(x) dx$$

---

$$R(y_{\mathcal{D}}) = \sigma^2 + \left(Bias(y_{\mathcal{D}})\right)^2 + Var(y_{\mathcal{D}})$$

# Theoretical issues for regression

## Bias-Variance analysis

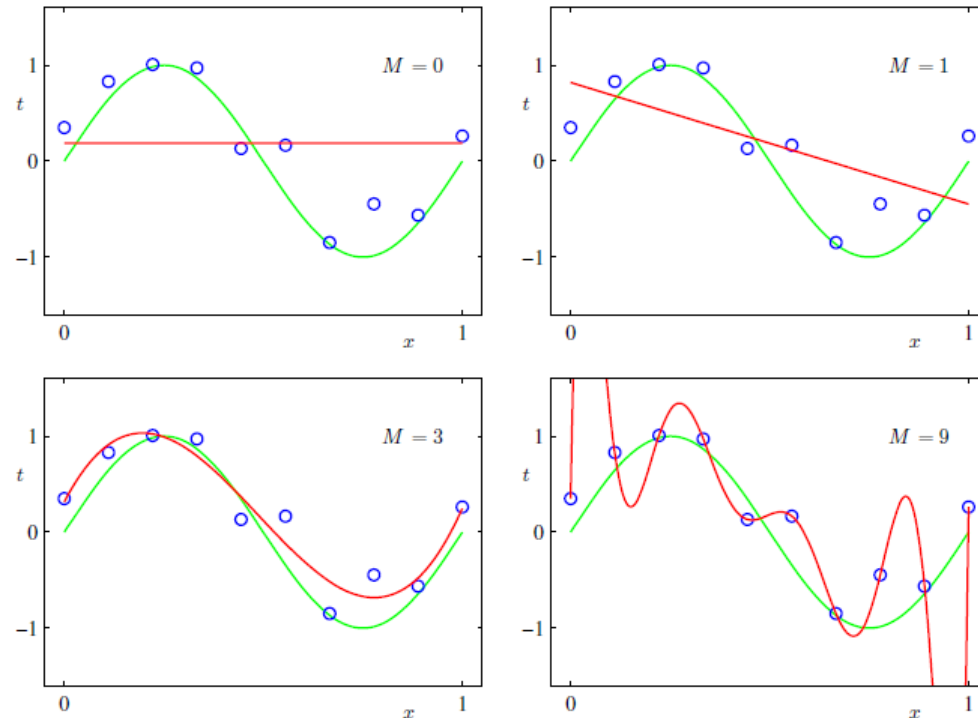


Illustration of the **Bias-Variance tradeoff** (a.k.a. **dilemma**)

# Theoretical issues for regression

## Bias-Variance analysis

In general,

- an **underfit** model will have a high bias
- an **overfit** model will have a high variance

The “ability to fit” has a name: **complexity** of the function class

- Models that are “more complex than needed” will tend to have a large prediction error, **which will be dominated by the variance term**
- Models that are “less complex than needed” will tend to have a large prediction error, **which will be dominated by the (square) bias term**

# Theoretical issues for regression

## Bias-Variance analysis

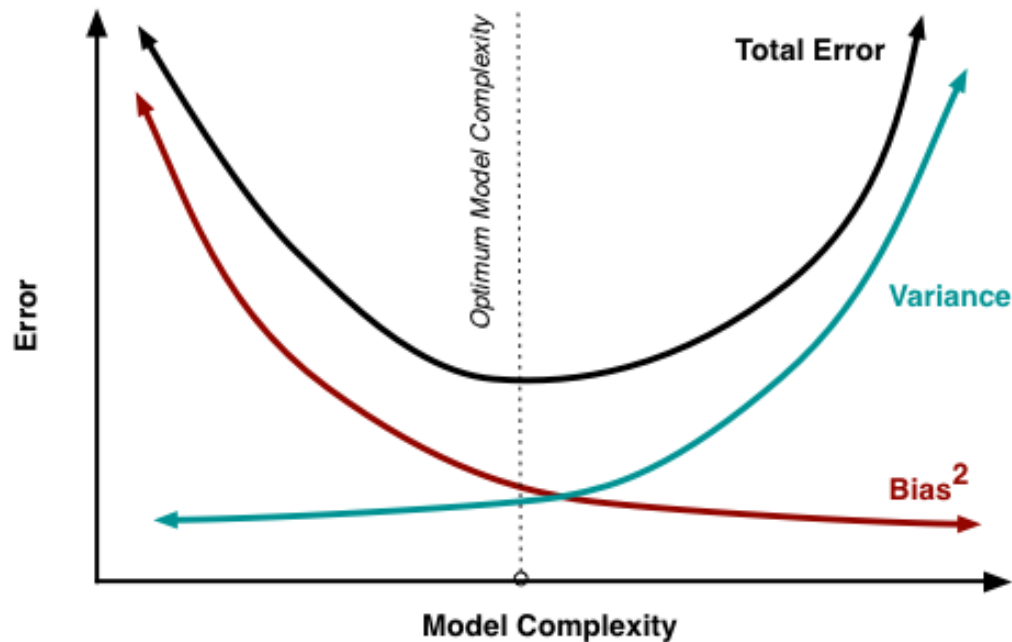
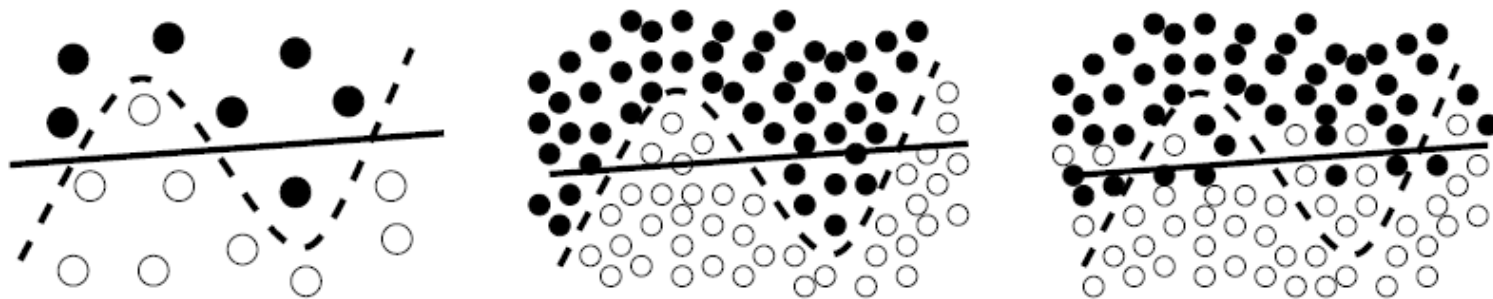


Illustration of **Bias<sup>2</sup>**, **Var**, MSE (Total Error) and Model Complexity

# Theoretical issues for regression

## Bias-Variance analysis



**Figure 2.1:** Illustration of the over-fitting dilemma: Given only a small sample (left) either, the solid or the dashed hypothesis might be true, the dashed one being more complex, but also having a smaller training error. Only with a large sample we are able to see which decision reflects the true distribution more closely. If the dashed hypothesis is correct the solid would under-fit (middle); if the solid were correct the dashed hypothesis would over-fit (right).

## Interpretation of the **Overfitting vs. underfitting** dilemma

---

(last two figures from S. Mika's PhD dissertation, Technische Universität Berlin, 2002)