The Bias-Variance Tradeoff

The compromise *bias-variance* express the effect of various possible factors on the final error between the hypothesis chosen by the LM and that which it would have had to choose, the ideal *target function*.

According to the general model of learning from examples, the LM receive from the environment a sample of data $\{\mathbf{x}_1,...,\mathbf{x}_m\}$ where $\mathbf{x}_i \in \mathcal{X}$. In the absence of additional information on their source, and for reasons of simplicity of modeling and mathematical analysis, one will suppose that these objects are drawn randomly and independently the ones of the others according to a probability distribution $\mathcal{D}_{\mathcal{X}}$ (it is what one calls *the assumption of independently and identically distributed).* Attached with each one of these data $\mathbf{x}_i$ the LM receives in addition one *label* or *supervision* u_i produced according to a functional dependence between $\mathbf{x}$ and u.

We note $\mathcal{S} = \{\mathbf{z}_1 = (\mathbf{x}_1, u_1),..., \mathbf{z}_m = (\mathbf{x}_m, u_m)\}$ the sample of learning made up here of supervised *examples*. To simplify, we will suppose that the functional dependence between an entry $\mathbf{x}$ and

its label u takes the form of a function f belonging to a family of functions $\mathcal{F}$. Without loosing the generality we also suppose that there can be erroneous labeling, in the form of a *noise*, i.e. a measurable bias between the proposed label and the true label according to *f*. The LM seeks to find a hypothesis function h, in the space of the functions $\mathcal{H}$ as near as possible to *f*, the target function. We will specify later the concept of proximity used to evaluate the distance between *f* and *h*.

Figure Error! **No text of specified style in document.**-1 illustrates the various sources of error between the target functions *f* and the hypothesis function *h*. We call *total error* the error resulting from the conjunction of these various errors between *f* and *h*. Let us detail them.

- The first source of error comes from the following fact: *nothing* does not allow *a priori* to postulate the equality between the target functions space $\mathcal{F}$ of the Nature and the hypotheses functions space $\mathcal{H}$ realizable by the LM. Of this fact, even if the LM provides an *optimal* assumption h^* (in the sense of the proximity measurement mentioned above), h^* is inevitably

taken in $\mathcal{H}$ and can thus be different from the target function *f*. It is the *approximation error* often called *inductive bias* (or simply bias) due to the difference between $\mathcal{F}$ and $\mathcal{H}$.

- Then, the LM does not provide in general h^* the optimal hypothesis in $\mathcal{H}$ but a hypothesis $\hat{h}$ based on the learning sample $\mathcal{S}$. Depending of this sample, the learned hypothesis $\hat{h}$ will be able to vary inside a set of functions that we denote by $\{\hat{h}\}_S$ to underline the dependence of each one of its elements on the random sample $\mathcal{S}$. The distance between h^* and the estimated hypothesis $\hat{h}$ who depends on the particularities of $\mathcal{S}$ is the *estimating error*. One can show formally that it is the *variance* related on the sensitivity of the calculation of the hypothesis $\hat{h}$ as function of the sample $\mathcal{S}$. More the hypotheses space $\mathcal{H}$ is rich, more, in general, this variance is important.

- Finally, it occurs the *noise* on labeling: because of transmission errors, the label *u* associated to $\mathbf{x}$ can to be inaccurate with respect to *f*. Hence the LM receives a sample of data

relative to *disturbed function* $f_b = f + noise$. It is the *intrinsic error* who generally complicates the research of the optimal hypothesis h^*.

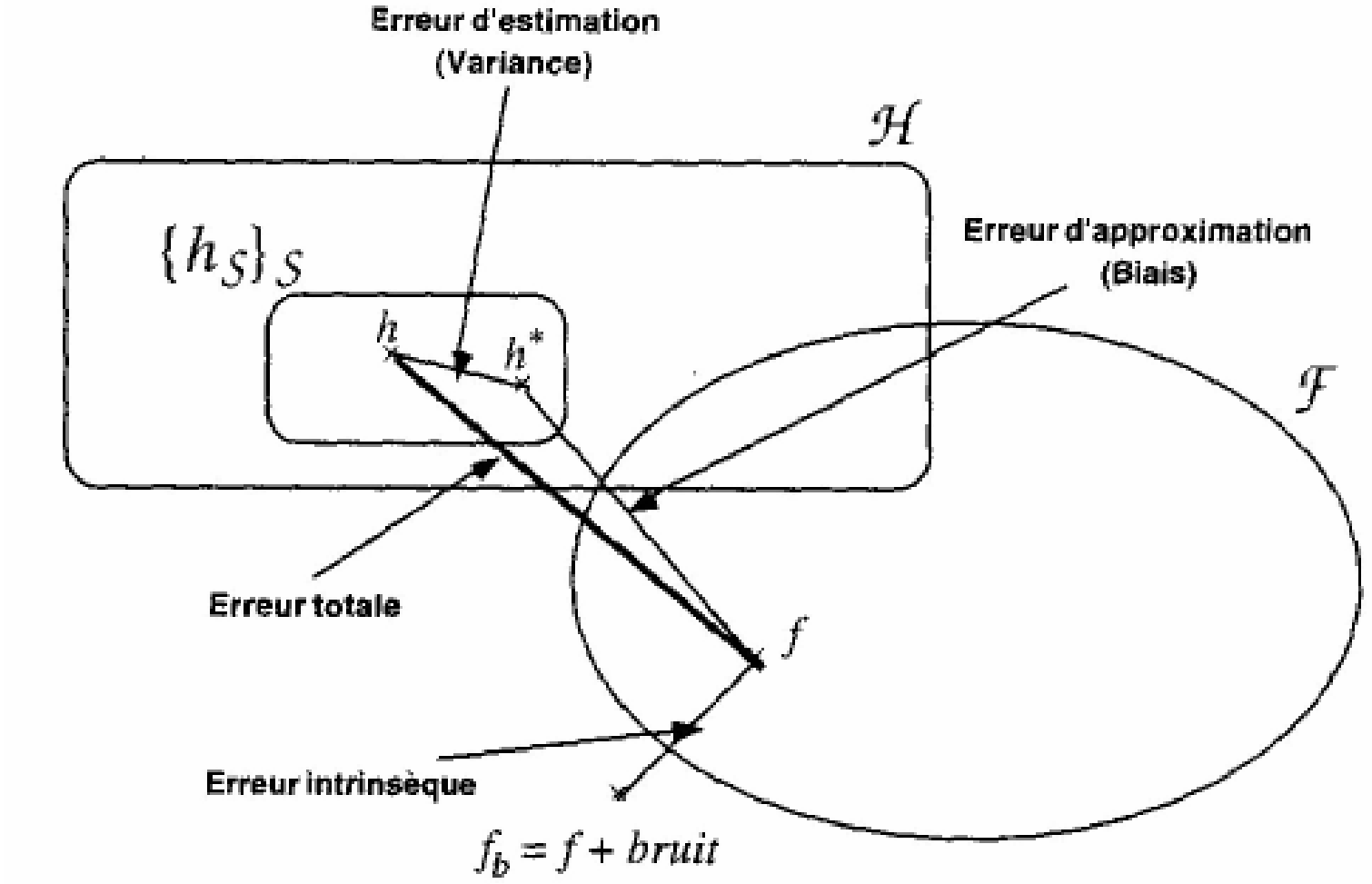

Erreur d'estimation
(Variance)
$\mathcal{H}$
$\{h_S\}_S$
Erreur d'approximation
(Biais)
h
h^*
$\mathcal{F}$
Erreur totale
f
Erreur intrinsèque
$f_b = f + bruit$

Figure Error! No text of specified style in document.**-1** *The various types of errors arising in the estimate of a targets function starting from a learning sample. With a more restricted space of hypotheses, one can reduce the variance, but generally at the price of a greater error of approximation.*

Being given these circumstances, the *bias-variance* trade off can be defined in the following way: to reduce bias due to the bad adequacy of $\mathcal{H}$ to $\mathcal{F}$ it is necessary to increase the richness of $\mathcal{H}$. Unfortunately, this enrichment will be paid, generally, with an increase in the variance. Of this fact, *the total error,* which is the sum of the approximation error and the estimation error, cannot significantly be decreased.

The bias-variance tradeoff should thus be rather called the compromise of the *approximation error/estimation error*. However, the important thing is that it is well a question of making a compromise, since one exploits a sum of terms that vary together in contrary direction. On the other hand the noise, or the *intrinsic error,* can only worsen the things while increasing. The ideal would be to have a null noise and a restricted $\mathcal{H}$ hypotheses space to reduce the variance, but at the same time *well informed,* i.e. containing only functions close to the target function, which would obviously be equivalent to have an *a priori* knowledge on Nature.

Regularization Methods

The examination of the compromise bias-variance and the analysis of the ERM principle by Vapnik have clearly shown that the mean of risk (the real risk) depends at the same time on the empirical risk measured on the learning sample and on the "capacity" of the space of the hypotheses functions. The larger this one is, the more there is a greater chance to find a hypothesis close to the target function (small approximation error), but also the hypothesis minimizing the empirical risk depends on the provided particular learning sample (big estimation error), which prohibits to exploit with certainty the performance measured by the empirical risk to the real risk.

In other words, supervised induction must always face the risk of *over-fitting*. If the space of the assumptions $\mathcal{H}$ is too rich, there are strong chances that the selected hypothesis, whose empirical risk is small, presents a high real risk. That is because several hypotheses can have a small empirical risk on a learning sample, while having very different real risks. It is thus not possible, only based on measured empirical risk, to distinguish the good hypothesis from the bad

one. It is thus necessary to restrict as much as possible the richness of the hypotheses space, while seeking to preserve a sufficient approximation capacity.

Tuning the Hypotheses Class

Since one can measure only the empirical risk, the idea is thus to try to evaluate the real risk by correcting the empirical risk, necessarily optimistic, by a *penalization term* corresponding to a measurement of the capacity of $\mathcal{H}$ the used hypotheses space. It is there the essence of all induction approaches, which revise the ERM principle (the adaptation to data) by a *regularization* term (depending on the hypotheses class). This fundamental idea is found in the heart of a whole set of methods like the *regularization theory, Minimum Description Length Principle: (MDLP), the Akaike information criterion* (AIC), and other methods based on complexity measures.

The problem thus defined is known, at least empirically, for a long time, and many techniques were developed to solve there. One can arrange them in three principal categories: methods of models selection, regularization techniques, and average methods.

- In the *methods of models selection*, the approach consists in considering a hypotheses space $\mathcal{H}$ and to decompose it into a discrete collection of nested subspaces $\mathcal{H}_1 \subseteq \mathcal{H}_2 \subseteq ... \subseteq \mathcal{H}_d \subseteq ...$ then, being given a learning sample, to try to identify the optimal subspace in which to choose the final hypothesis. Several methods were proposed within this framework, that one can gather in two types:

– *complexity penalization methods,* among which appear the *structural risk minimization principle (SRM)* of Vapnik, the *Minimum Description Length principle* of Rissanen (1978) and various methods or statistical criteria of selection,

– *methods of validation by multiple learning:* among which appears the *cross validation* and *bootstrapping*.

- The *regularization methods* act in the same spirit as the methods selection models, put aside that they do not impose a discrete decomposition on the hypotheses class. A penalization criterion of is associated to each hypothesis, which, either measure the complexity of their parametric structure, or the global properties of "regularity", related, for example, to the

derivability of the hypothesis functions or their dynamics (for example the high frequency functions, i.e. changing value quickly, will be more penalized comparatively to the low frequency functions).

- The *average methods* do not select a single hypothesis in the space $\mathcal{H}$, but choose a weighed combination of hypothesis to form one prediction function. Such a weighed combination can have like effect "to smooth" the erratic hypothesis (as in the methods of *Bayesian average* and in the *bagging methods)*,or to increase the capacity of representation of the hypothesis class if this one is not convex (as in the *boosting* methods*).*

All these methods generally led to notable improvements of the performances compared to the "naive" methods. However, they ask to be used carefully. On the one side, indeed, they correspond sometimes to an increase in the richness of the hypotheses space, and to an increased risk of over-fitting. On the other side, they require frequently an expertise to be applied, in particular because additional parameters should be regulated. Some recent work tries, for these

reasons, to determine automatically the suitable complexity of the candidate's hypotheses to adapt to the learning data.

Selection of the Models

We will define more formally the problem of the models selection, which is the objective of all these methods.

Let $\mathcal{H}_1 \subseteq \mathcal{H}_2 \subseteq ...\mathcal{H}_d \subseteq$ be a nested sequence of spaces or classes of hypotheses (or *models)* where the spaces $\mathcal{H}_d$ are of increasing capacity. The target function *f* can or cannot be included in one of these classes. Let h_d^* be the optimal hypothesis in the class of hypotheses $\mathcal{H}_d$ and $R(d) = R_{real}\left(h_d^*\right)$ the associate real risk. We note that the sequence $\left\{R(d)\right\}_{1 \le d \le \infty}$ is decreasing since the hypotheses classes $\mathcal{H}_d$ are nested, and thus their approximation capacity of the targets function *f* can only increase.

Using these notations, the problem of the models selection can be defined as follows.

Definition 1.2 The model selection problem *consist to choose, on the basis of a learning sample* $\mathcal{S}$ *of length m, a class of hypotheses* $\mathcal{H}_{d^*}$ *and a hypothesis* $h_d \in \mathcal{H}_{d^*}$ *such that the associate real risk* $R_{real}\left(h_d\right)$ *is minimal.*

The underlying conjecture is that the real risk associate with the selected hypothesis h_d for each class $\mathcal{H}_d$ present a global minimum for a nontrivial value of *d* (i.e. different from zero and *m*) corresponding to the "ideal" hypothesis space $\mathcal{H}_{d^*}$. (see Figure Error! **No text of specified style in document.**-2**)**.

It is thus a question of finding the ideal hypothesis space $\mathcal{H}_{d^*}$, and in addition to select the best hypothesis h_d in $\mathcal{H}_{d^*}$.The definition say nothing about this last problem. It is generally solved by using the *ERM* principle dictating to seek the hypothesis that minimizes the empirical risk.

For the selection of $\mathcal{H}_{d^*}$, one uses an estimate of the optimal real risk in each $\mathcal{H}_d$ by choosing the best hypothesis according to the empirical risk (the *ERM* method*)* and by correcting the associated empirical risk with a penalization term related to the characteristics of space $\mathcal{H}_d$. The problem of model selection consists then in solving an equation of the type:

$$d^* = \underset{d}{ArgMin}\left\{h_d \in \mathcal{H}_d : R_{real}^{estimated}\left(h_d\right)\right\}$$

$$= \underset{d}{ArgMin}\left\{h_d \in \mathcal{H}_d : R_{emp}\left(h_d\right)\right\} + \textit{penalization term}$$

Let us note that the choice of the best hypotheses space depends on the size *m* of the data sample. The larger this one is, the more it is possible, if necessary, to choose without risk (i.e. with a little variance or confidence interval) a rich hypotheses space making possible to approach as much as possible the targets function *f*.

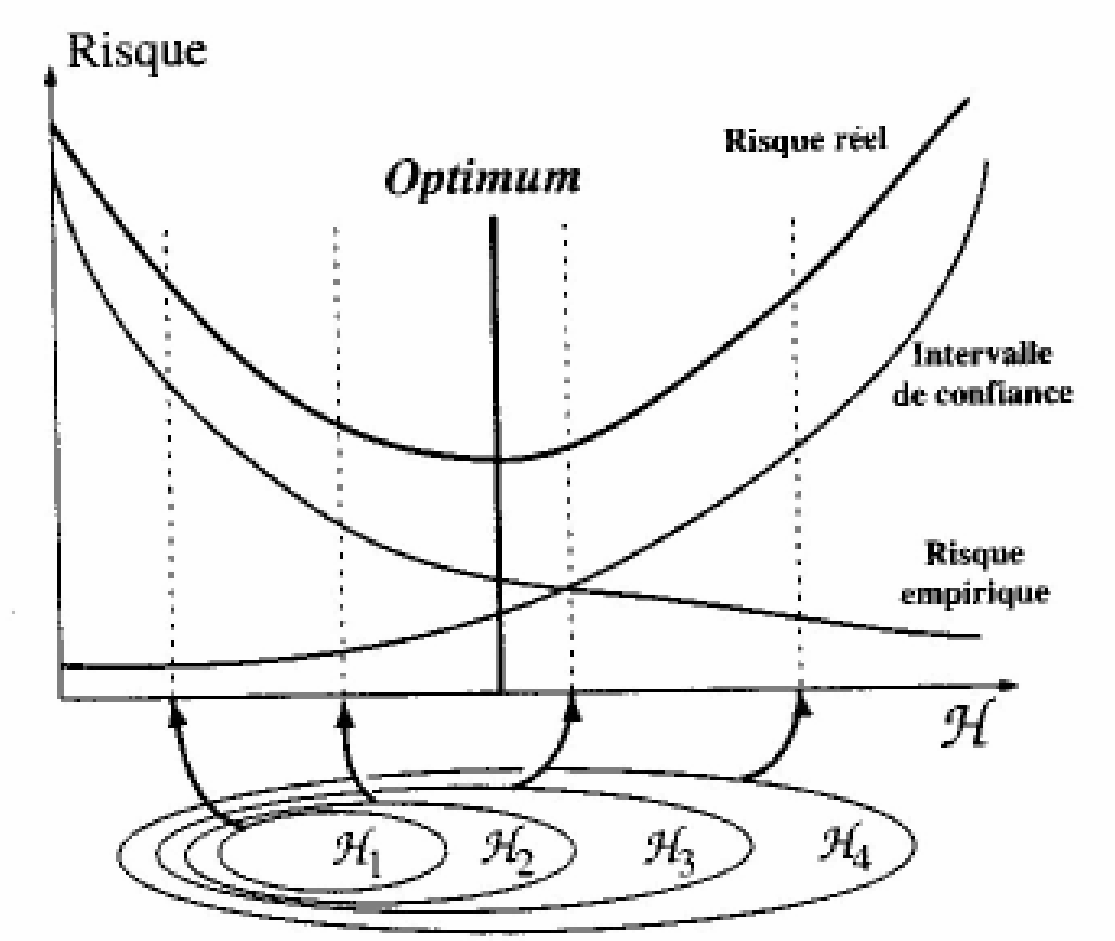
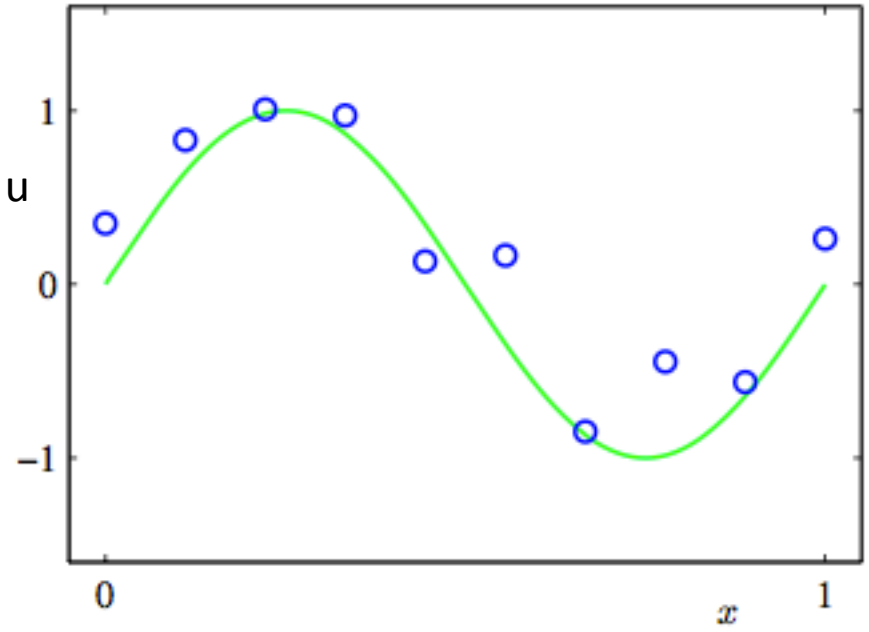


Figure Error! No text of specified style in document.**-2** *The bounds on the real risk results from the sum of the empirical risk and a confidence interval depending on the capacity of the associated hypotheses space. By supposing that one has a nested sequence of hypotheses spaces of increasing capacity and subscripted by* d*, the accessible optimal empirical risk decreases for increasing* d *(corresponding to the bias), while the confidence interval (corresponding to the variance) increases. The minimal bound on the real risk is reached for a suitable hypotheses space* $\mathcal{H}_d$.

Gasirea polinomului optim



$$\mathcal{S} = \{(x_1, u_1), \ldots, (x_m, u_m)\}$$

multime cu exemple de antrenare, m = 10

$$u_i = f(x_i) + \epsilon_i$$

functia tinta
(vrem sa o invatam)

zgomot aleator
(corupe datele)

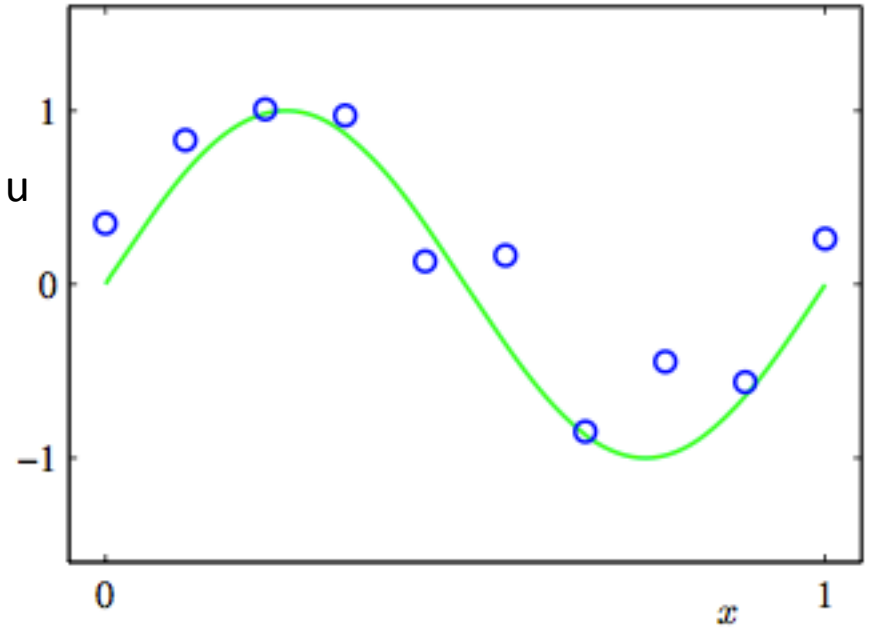
$$\epsilon_i \sim \mathcal{N}(\mu, \sigma)$$

Problema de regresie:
pe baza lui S gaseste (invata) functia h care stabileste corespondenta

$$u = h(x)$$

- foloseste h pentru predictie pentu noi valori ale lui x

Functia tinta



$$\mathcal{S} = \{(x_1, u_1), \ldots, (x_m, u_m)\}$$

multime cu exemple de antrenare, m = 10

$$u_i = f(x_i) + \epsilon_i$$

functia tinta
(vrem sa o invatam)

zgomot aleator
(corupe datele)

$$\epsilon_i \sim \mathcal{N}(\mu, \sigma)$$

$$f(x) = sin(2\pi x)$$

$$sin(x) = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)!} x^n = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \ldots$$

$$sin(2\pi x) \approx 6.28x - 41.34x^3 + 81.60x^5 - \ldots$$

Spatiul de functii

$$\mathcal{H}_i = \{h(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots w_i x^i\}$$

Spatiul functiilor polinomiale (curbe) de grad i
w e parametru, h e liniara in w, h e neliniara in x

$$\mathcal{H}_0 \subseteq \mathcal{H}_1 \subseteq \mathcal{H}_2 \subseteq \ldots$$

functie constanta — drepte — parabole

Riscul empiric al unei ipoteze h:

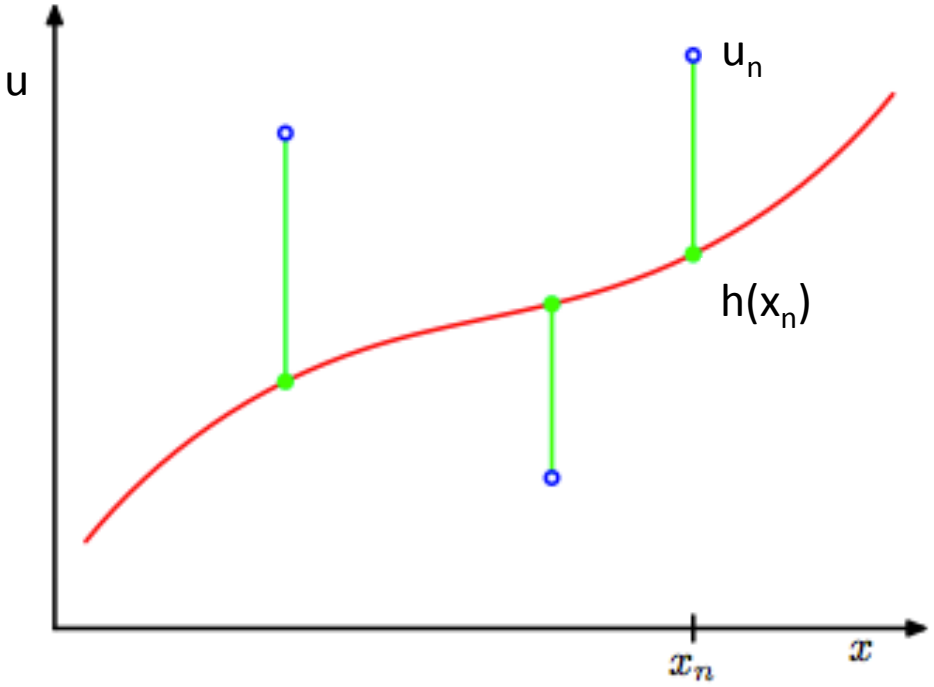
$$R_{emp}(h) = \frac{1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} l(u_i, h(x_i, \mathbf{w}))$$

functie cost (loss) – masoara costul pe care il implica luarea deciziei $h(x_i)$ in loc de u_i

Functia cost

$$R_{emp}(h) = \frac{1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} l(u_i, h(x_i, \mathbf{w}))$$

functie cost (loss) – masoara costul pe care il implica luarea deciziei $h(x_i)$ in loc de u_i



Functia cost

$$R_{emp}(h) = \frac{1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} l(u_i, h(x_i, \mathbf{w}))$$

functie cost (loss) – masoara costul pe care il implica luarea deciziei $h(x_i)$ in loc de u_i

Exemple de functii cost:

$$l(u_i, h(x_i, \mathbf{w})) = \sum_{i=1}^{|\mathcal{S}|} (u_i - h(x_i, \mathbf{w}))^2$$

$$l(u_i, h(x_i, \mathbf{w})) = \sum_{i=1}^{|\mathcal{S}|} |u_i - h(x_i, \mathbf{w})|$$

Principiul ERM

- gaseste ipoteza h* care minimizeaza riscul empiric (eroarea de antrenare)

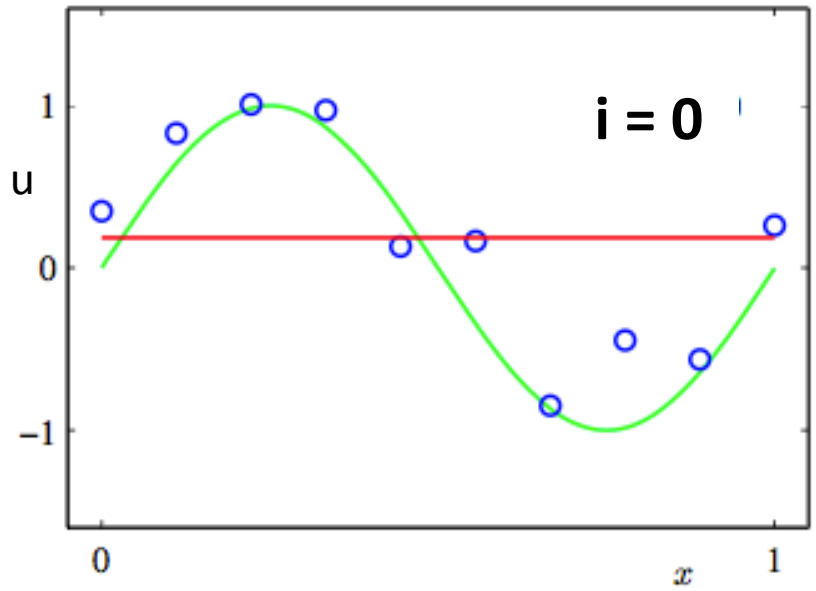
$$h^*_{\mathcal{S},\mathcal{H}} = arg \min_{h \in \mathcal{H}} R_{emp}(h)$$

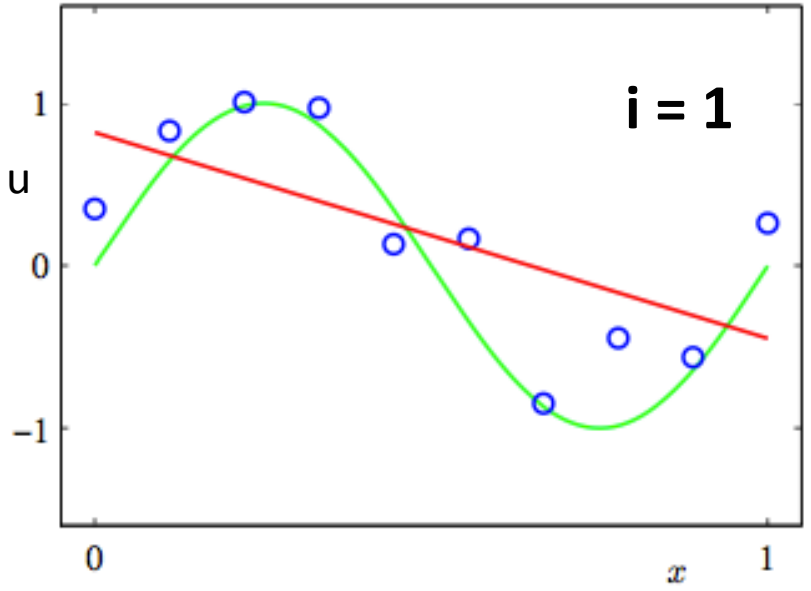
$$R_{emp}(h) = \frac{1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} l(u_i, h(x_i, \mathbf{w}))$$

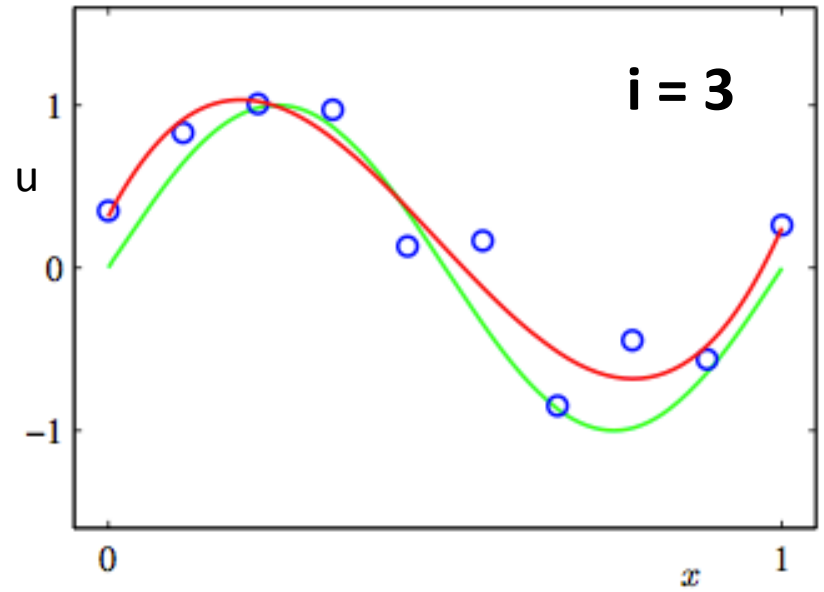
- afla parametri w care minimizeaza riscul empiric folosind functia de cost

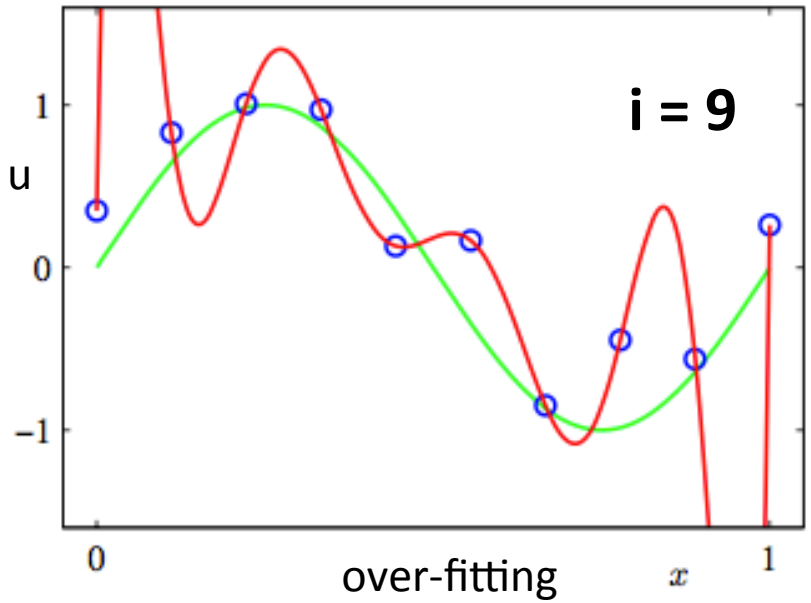
$$l(u_i, h(x_i, \mathbf{w})) = \sum_{i=1}^{|\mathcal{S}|} (u_i - h(x_i, \mathbf{w}))^2$$

Alegerea modelului $\mathcal{H}_i$ si aflarea lui $h^*_{\mathcal{S},\mathcal{H}_i}$



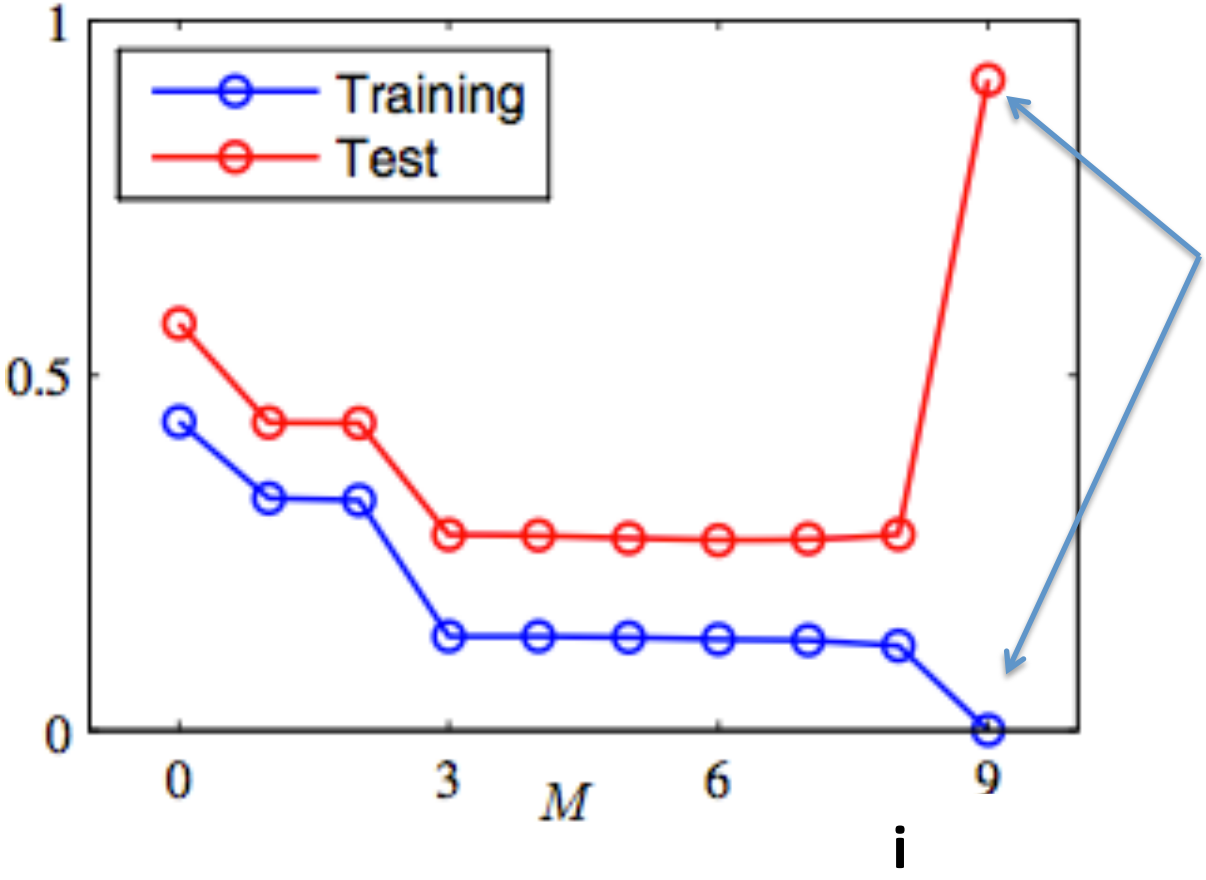






Evolutia riscului empiric

Evaluam ipotezele pe o multime de test de 100 de exemple



Over-fitting:

- eroare pe multimea de antrenare mica
- eroare pe multimea de testare mare

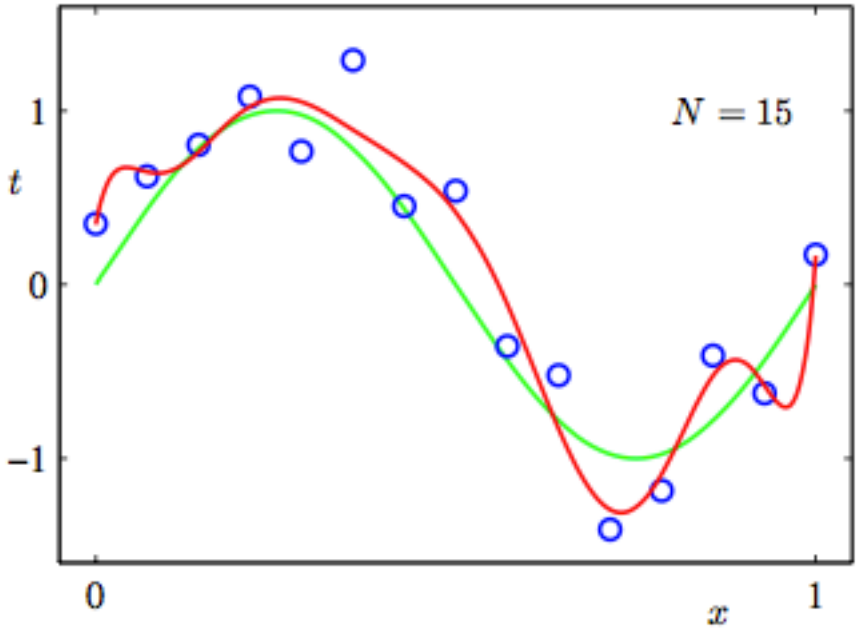
Coeficientii lui $h^*_{\mathcal{S},\mathcal{H}_i}$

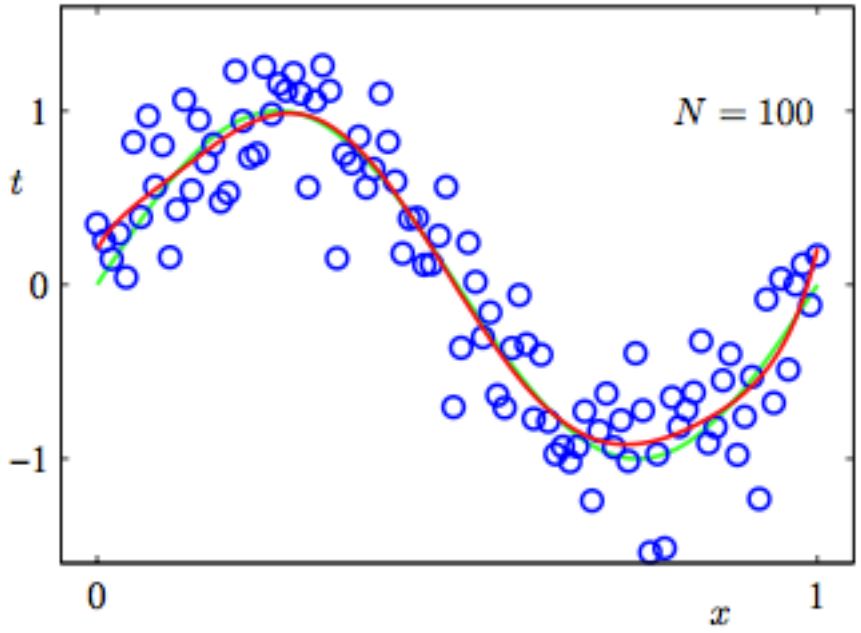
$$\mathcal{H}_i = \{h(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots w_i x^i\}$$

	i = 0	i = 1	i = 6	i = 9
$w_0^\star$	0.19	0.82	0.31	0.35
$w_1^\star$		-1.27	7.99	232.37
$w_2^\star$			-25.43	-5321.83
$w_3^\star$			17.37	48568.31
$w_4^\star$				-231639.30
$w_5^\star$				640042.26
$w_6^\star$				-1061800.52
$w_7^\star$				1042400.18
$w_8^\star$				-557682.99
$w_9^\star$				125201.43

$$sin(2\pi x) \approx 6.28x - 41.34x^3 + 81.60x^5 - \ldots$$

Comportamentul unui model in functie de marimea lui S





Problema de over-fitting se elimina treptat pe masura ce creste numarul de exemple de antrenare

Metode de regularizare

$$l(u_i, h(x_i, \mathbf{w})) = \sum_{i=1}^{|\mathcal{S}|} (u_i - h(x_i, \mathbf{w}))^2 + \underbrace{\lambda ||w||^2}_{\text{penalitate}}$$

$$||w||^2 = w_0^2 + w_1^2 + \ldots + w_i^2$$

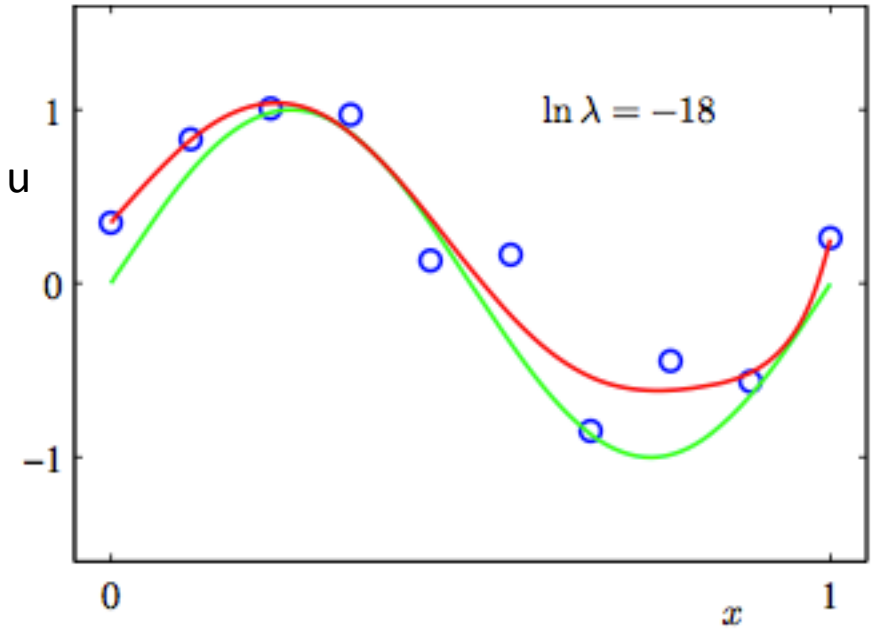
λ - controleaza importanta termenului de regularizare/penalitate

Impactul includerii unei termen de regularizare

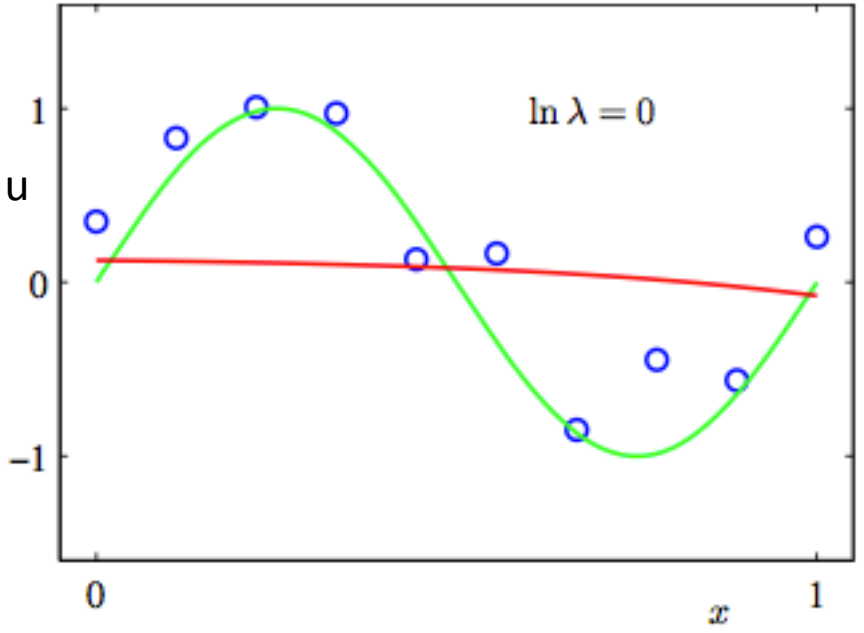
$$l(u_i, h(x_i, \mathbf{w})) = \sum_{i=1}^{|\mathcal{S}|} (u_i - h(x_i, \mathbf{w}))^2 + \lambda ||w||^2$$

penalitate

$h^*_{\mathcal{S},\mathcal{H}_9}$



$h^*_{\mathcal{S},\mathcal{H}_9}$



Impactul includerii unei termen de regularizare

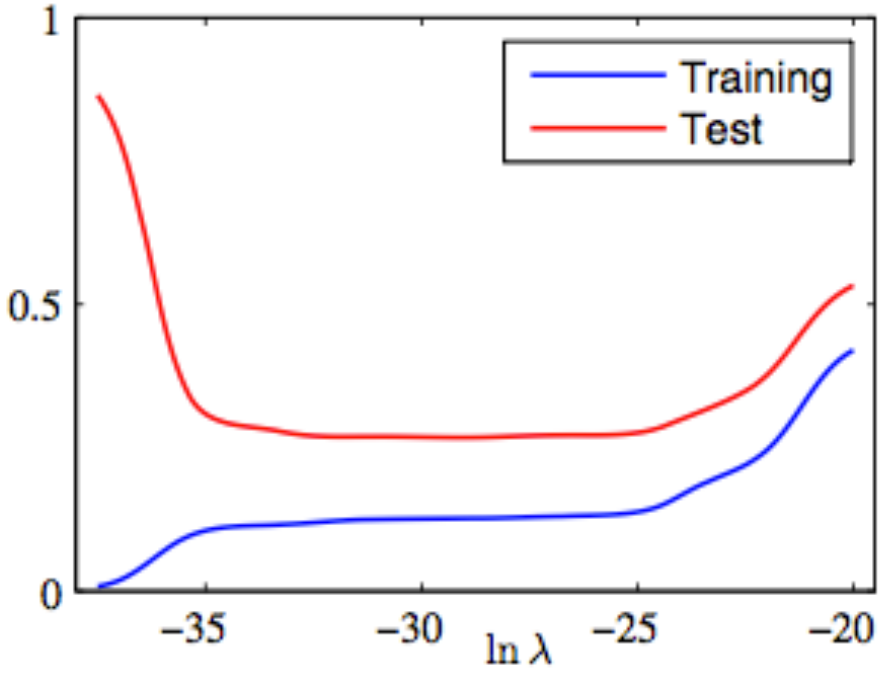
	i = 0	i = 1	i = 6	i = 9
$w_0^\star$	0.19	0.82	0.31	0.35
$w_1^\star$		-1.27	7.99	232.37
$w_2^\star$			-25.43	-5321.83
$w_3^\star$			17.37	48568.31
$w_4^\star$				-231639.30
$w_5^\star$				640042.26
$w_6^\star$				-1061800.52
$w_7^\star$				1042400.18
$w_8^\star$				-557682.99
$w_9^\star$				125201.43

i = 9

	$\ln \lambda = -\infty$	$\ln \lambda = -18$	$\ln \lambda = 0$
$w_0^\star$	0.35	0.35	0.13
$w_1^\star$	232.37	4.74	-0.05
$w_2^\star$	-5321.83	-0.77	-0.06
$w_3^\star$	48568.31	-31.97	-0.05
$w_4^\star$	-231639.30	-3.89	-0.03
$w_5^\star$	640042.26	55.28	-0.02
$w_6^\star$	-1061800.52	41.32	-0.01
$w_7^\star$	1042400.18	-45.95	-0.00
$w_8^\star$	-557682.99	-91.53	0.00
$w_9^\star$	125201.43	72.68	0.01

$$sin(2\pi x) \approx 6.28x - 41.34x^3 + 81.60x^5 - \ldots$$

Impactul includerii unei termen de regularizare



i = 9

	$\ln \lambda = -\infty$	$\ln \lambda = -18$	$\ln \lambda = 0$
$w_0^\star$	0.35	0.35	0.13
$w_1^\star$	232.37	4.74	-0.05
$w_2^\star$	-5321.83	-0.77	-0.06
$w_3^\star$	48568.31	-31.97	-0.05
$w_4^\star$	-231639.30	-3.89	-0.03
$w_5^\star$	640042.26	55.28	-0.02
$w_6^\star$	-1061800.52	41.32	-0.01
$w_7^\star$	1042400.18	-45.95	-0.00
$w_8^\star$	-557682.99	-91.53	0.00
$w_9^\star$	125201.43	72.68	0.01

Alegerea modelului

Impartim datele initiale in 2 multimi: multimea de antrenare si multimea de validare.

Alegem i sau λ pe baza erorii pe multimii de validare