Lecture 8: Learning theory, Bias-Variance

Andreas Wichert

Department of Computer Science and Engineering

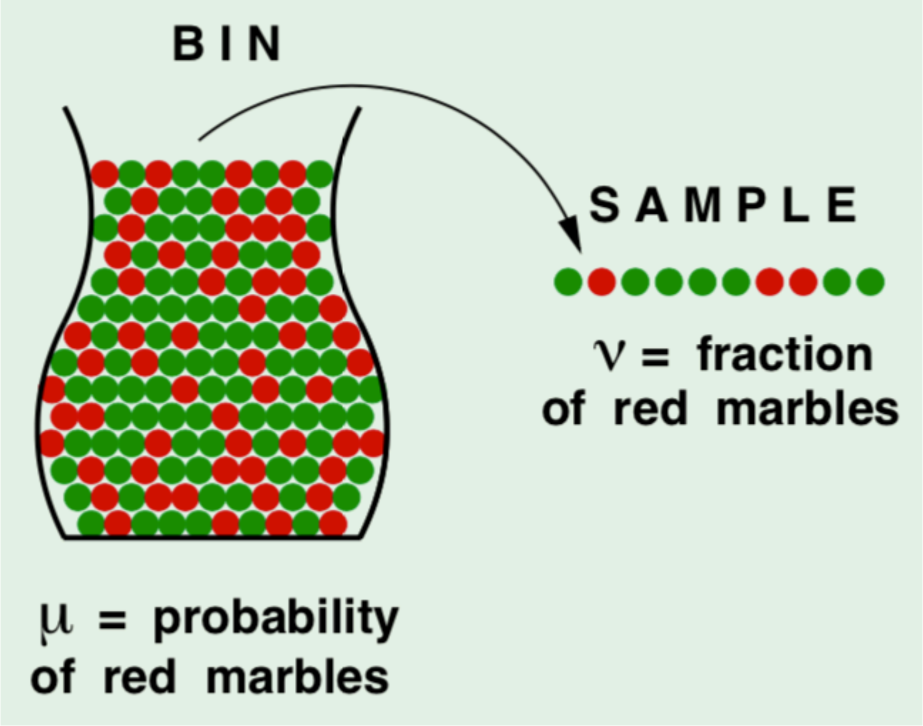
Técnico Lisboa

- Under what conditions is successful learning possible?
- Under what conditions is a particular learning algorithm assured of learning successfully?

Generalization

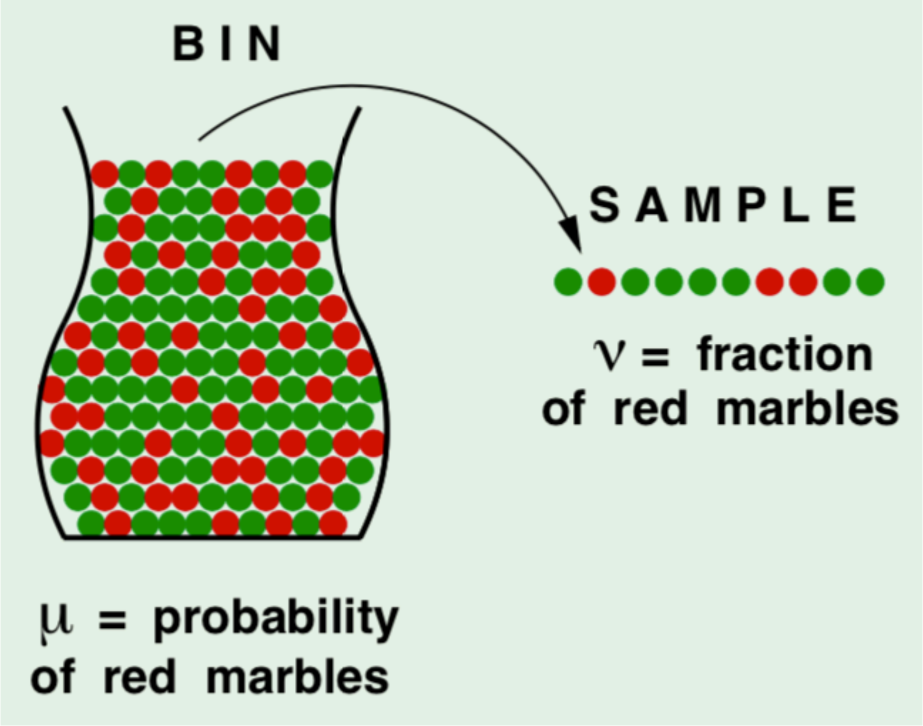
- We pick a random sample of *N* independent marbles (with replacement) from this bin, and observe the fraction *v* of red marbles

- What does the value of *v* tell us about the value of *μ*?

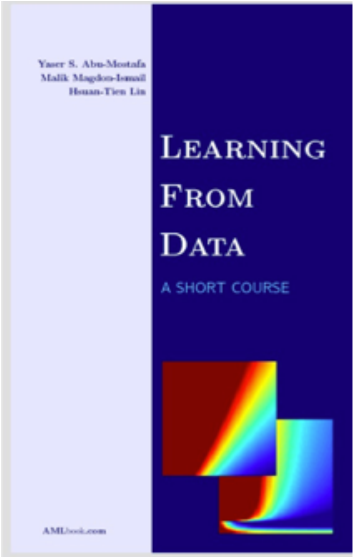


- As the sample size *N* grows *ν* approaches value of *μ*

$$p(|\nu - \mu| > \epsilon) \leq \frac{2}{e^{2\cdot\epsilon^2\cdot N}}$$



ν and μ depend on the hypothesis



Yaser S. Abu-Mostafa
Malik Magdon-Ismail
Hsuan-Tien Lin

Learning From Data

A Short Course

AMLbook.com

ν is in sample $E_{in}(h)$

μ is out of sample $E_{out}(h)$

$$p(|E_{in}(h) - E_{out}(h)| > \epsilon) \leq \frac{2}{e^{2\cdot\epsilon^2\cdot N}}$$

M experiments

$$p(|E_{in}(g) - E_{out}(g)| > \epsilon) \leq \frac{2 \cdot M}{e^{2\cdot\epsilon^2\cdot N}}$$

- Can we overcome the problem of large sample by regularisation?
- The use of least squares, can lead to severe over-fitting if complex models are trained using data sets of limited size.
- However, limiting the number of parameters in order to avoid over-fitting has the side effect of limiting the flexibility of the model.
- Although the introduction of regularization terms can control over-fitting for models with many parameters, this raises the question of how to determine a suitable value for the regularization coefficient $\boldsymbol{\lambda}$

Expectations

- One of the most important operations involving probabilities is that of finding weighted averages of functions.
- The average value of some function *f(x)* under a probability distribution *p(x)* is called the expectation of *f(x)*

$$\mathbb{E}(f) = \sum_x p(x) f(x)$$

$$\mathbb{E}(f) = \int p(x) f(x) dx$$

For a finite number N of points the expectation can be approximated (similar or equal) as a

$$\mathbb{E}(f) \simeq \frac{1}{N} \sum_{\eta=1}^{N} f(x_n)$$

A conditional expectation with respect to a conditional distribution is given by

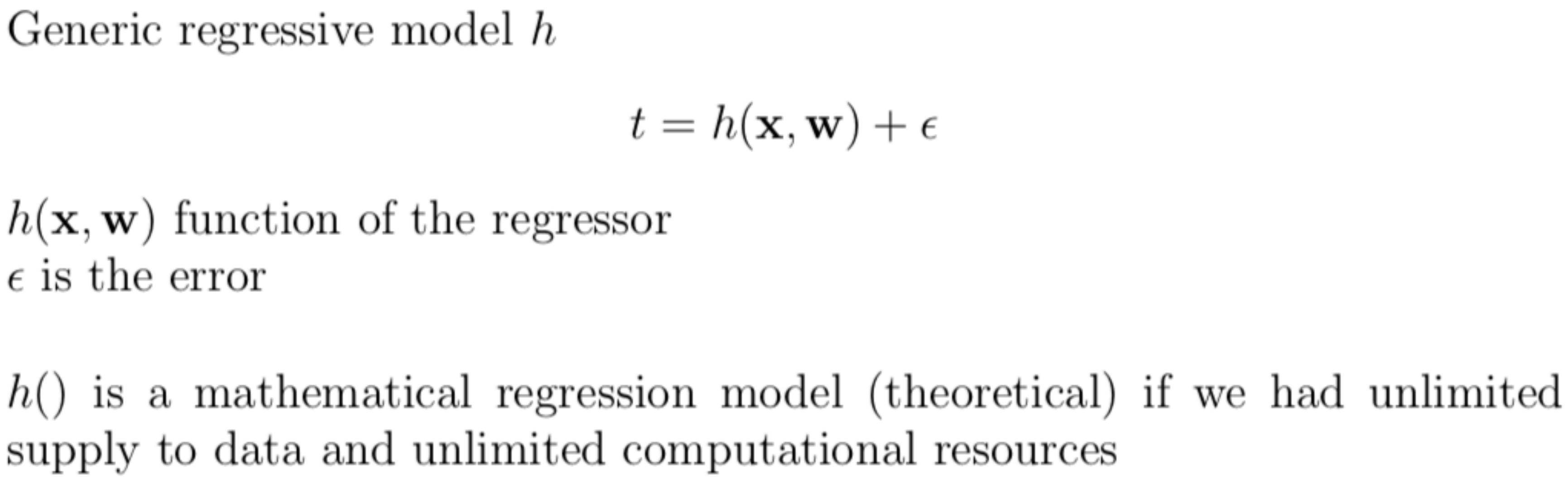
$$\mathbb{E}(f|y) = \sum_{x} p(x|y) f(x)$$

The variance of $f(x)$ is defined by

$$var[f] = \mathbb{E}\left[(f(x) - \mathbb{E}[f(x)])^2\right]$$

and provides a measure of how much variability there is in $f(x)$ around its mean value $\mathbb{E}[f(x)]$

Finite Sample-Size Consideration



Generic regressive model h

$$t = h(\mathbf{x}, \mathbf{w}) + \epsilon$$

$h(\mathbf{x}, \mathbf{w})$ function of the regressor
ϵ is the error

$h()$ is a mathematical regression model (theoretical) if we had unlimited supply to data and unlimited computational resources

Empirical knowledge represented by training sample D (real world)

$$D \to \hat{\mathbf{w}}$$

with

$$y(\mathbf{x}, \hat{\mathbf{w}})$$

is an approximation of the regression model.

Given the training sample D the estimator $\hat{\mathbf{w}}$ is the minimizer of the cost function

$$E(\hat{\mathbf{w}}) = \frac{1}{2} \cdot \sum_{\eta=1}^{N} (t_\eta - y(\mathbf{x}_\eta, \hat{\mathbf{w}}))^2$$

$\frac{1}{2}$ is used for consistency only and can be ignored

$$E(\hat{\mathbf{w}}) = \sum_{\eta=1}^{N} (t_\eta - y(\mathbf{x}_\eta, \hat{\mathbf{w}}))^2$$

$$E(\hat{\mathbf{w}}) = \sum_{\eta-1}^{N} \left(t_\eta - y(\mathbf{x}_\eta, \hat{\mathbf{w}})\right)^2$$

$\mathbb{E}_D$ denotes the average operator taken over the entire training sample D
The variables or their functions that come under the average operator $\mathbb{E}_D$ are denoted by $\mathbf{x}$ and t
The pair $(\mathbf{x}, t)$ represents an example in the training sample D

$\mathbb{E}$ acts on the whole ensemble of $\mathbf{x}$ and t (population) of which D is a subset.

Because of $D \rightarrow \hat{\mathbf{w}}$ we may write $y(\mathbf{x}, D)$ instead of $y(\mathbf{x}, \hat{\mathbf{w}})$

$$E(\hat{\mathbf{w}}) = \mathbb{E}_D\left[(t - y(\mathbf{x}, D)^2\right]$$

Next we write

$$t - y(\mathbf{x}, D) = (t - h(\mathbf{x}, \mathbf{w})) + (h(\mathbf{x}, \mathbf{w}) - y(\mathbf{x}, D))$$

because of

$$t = h(\mathbf{x}, \mathbf{w}) + \epsilon$$

$$t - y(\mathbf{x}, D) = \epsilon + (h(\mathbf{x}, \mathbf{w}) - y(\mathbf{x}, D))$$

and

$$E(\hat{\mathbf{w}}) = \mathbb{E}_D\left[(\epsilon + (h(\mathbf{x}, \mathbf{w}) - y(\mathbf{x}, D)))^2\right]$$

$$E(\hat{\mathbf{w}}) = \mathbb{E}_D\left[(\epsilon + (h(\mathbf{x}, \mathbf{w}) - y(\mathbf{x}, D)))^2\right]$$

we multiply out

$$E(\hat{\mathbf{w}}) = \mathbb{E}_D\left[\epsilon^2\right] + \mathbb{E}_D\left[(h(\mathbf{x}, \mathbf{w}) - y(\mathbf{x}, D))^2\right] + 2 \cdot \mathbb{E}_D\left[\epsilon \cdot h(\mathbf{x}, \mathbf{w}) - \epsilon \cdot y(\mathbf{x}, D))\right]$$

Now the last expectation term is zero

$$2 \cdot \mathbb{E}_D\left[\epsilon \cdot h(\mathbf{x}, \mathbf{w}) - \epsilon \cdot y(\mathbf{x}, D))\right] = 0$$

Now the last expectation term is zero

$$2 \cdot \mathbb{E}_D \left[\epsilon \cdot h(\mathbf{x}, \mathbf{w}) - \epsilon \cdot y(\mathbf{x}, D))\right] = 0$$

because

- The expectational error ϵ is uncorrelated with the regression function $h(\mathbf{x}, \mathbf{w})$

  $$\mathbb{E}_D \left[\epsilon \cdot h(\mathbf{x}, \mathbf{w})\right] = 0$$

  This is the **principle of orthogonality** which states that all the information about D available to us through input $\mathbf{x}$ has been encoded into the regression function $h(\mathbf{x}, \mathbf{w})$ the mean value of the exceptional error ϵ given any realisation $\mathbf{x}$ is zero

  $$\mathbb{E}_D \left[\epsilon | \mathbf{x}\right] = 0$$

  $$\mathbb{E}_D \left[\epsilon \cdot h(\mathbf{x}, \mathbf{w})\right] = \mathbb{E}_D \left[\mathbb{E}_D \left[\epsilon \cdot h(\mathbf{x}, \mathbf{w}) | \mathbf{x}\right]\right] = \mathbb{E}_D \left[h(\mathbf{x}, \mathbf{w}) \cdot \mathbb{E}_D \left[\epsilon | \mathbf{x}\right]\right]$$

- The expectational error ϵ pertains to the regression model $h(\mathbf{x}, \mathbf{w})$ whereas the approximation function pertains to the physical model $y(\mathbf{x}, D))$

$$E(\hat{\mathbf{w}}) = \mathbb{E}_D\left[\epsilon^2\right] + \mathbb{E}_D\left[(h(\mathbf{x},\mathbf{w}) - y(\mathbf{x},D))^2\right]$$

$\mathbb{E}_D\left[\epsilon^2\right]$ is the variance of the expectational error evaluated over the training sample D

It is the constant noise error because it is independent of the weight vector $\mathbf{w}$

The natural measure of effectiveness of $y(\mathbf{x},\hat{\mathbf{w}})$ as a predictor of the desired response t is defined as the average loss is defined as

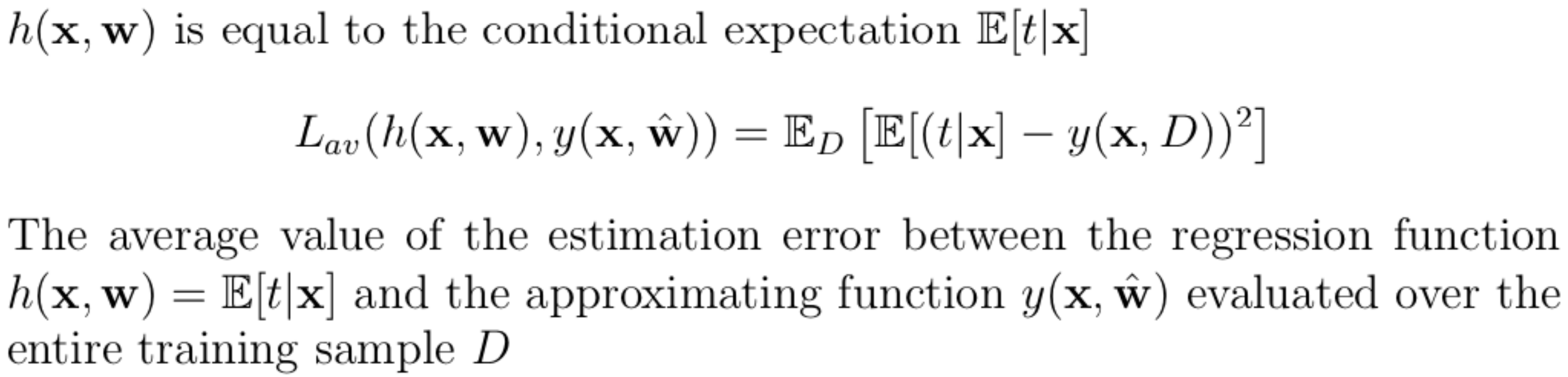
$$L_{av}(h(\mathbf{x},\mathbf{w}), y(\mathbf{x},\hat{\mathbf{w}})) = \mathbb{E}_D\left[(h(\mathbf{x},\mathbf{w}) - y(\mathbf{x},D))^2\right]$$

The natural measure of effectiveness of $y(\mathbf{x}, \hat{\mathbf{w}})$ as a predictor of the desired response t is defined as the average loss is defined as

$$L_{av}(h(\mathbf{x}, \mathbf{w}), y(\mathbf{x}, \hat{\mathbf{w}})) = \mathbb{E}_D \left[(h(\mathbf{x}, \mathbf{w}) - y(\mathbf{x}, D))^2 \right]$$

This natural measure is fundamentally important because it provides the mathematical basis for the tradeoff between the bias and variance that results from the use of $y(\mathbf{x}, \hat{\mathbf{w}})$ as the approximation to $h(\mathbf{x}, \mathbf{w})$

Bias-Variance



$h(\mathbf{x}, \mathbf{w})$ is equal to the conditional expectation $\mathbb{E}[t|\mathbf{x}]$

$$L_{av}(h(\mathbf{x}, \mathbf{w}), y(\mathbf{x}, \hat{\mathbf{w}})) = \mathbb{E}_D \left[\mathbb{E}[(t|\mathbf{x}] - y(\mathbf{x}, D))^2\right]$$

The average value of the estimation error between the regression function $h(\mathbf{x}, \mathbf{w}) = \mathbb{E}[t|\mathbf{x}]$ and the approximating function $y(\mathbf{x}, \hat{\mathbf{w}})$ evaluated over the entire training sample D

We rewrite

$$\mathbb{E}[t|\mathbf{x}] - y(\mathbf{x}, D)) = (\mathbb{E}[t|\mathbf{x}] - \mathbb{E}_D\left[y(\mathbf{x}, D)\right]) + (\mathbb{E}_D\left[y(\mathbf{x}, D)\right] - y(\mathbf{x}, D))$$

and

$$L_{av}(h(\mathbf{x}, \mathbf{w}), y(\mathbf{x}, \hat{\mathbf{w}})) = \mathbb{E}_D\left[(\mathbb{E}[t|\mathbf{x}] - \mathbb{E}_D\left[y(\mathbf{x}, D)\right] + \mathbb{E}_D\left[y(\mathbf{x}, D)\right] - y(\mathbf{x}, D))^2\right]$$

with

$$B(\hat{\mathbf{w}}) = \mathbb{E}_D\left[y(\mathbf{x}, D)\right] - \mathbb{E}[t|\mathbf{x}]$$

we simplify to

$$L_{av}(h(\mathbf{x}, \mathbf{w}), y(\mathbf{x}, \hat{\mathbf{w}})) = \mathbb{E}_D\left[(-B(\hat{\mathbf{w}}) - y(\mathbf{x}, D) + \mathbb{E}_D\left[y(\mathbf{x}, D)\right])^2\right]$$

- we rewrite with a minus sign

$$L_{av}(h(\mathbf{x},\mathbf{w}),y(\mathbf{x},\hat{\mathbf{w}})) = \mathbb{E}_D\left[(-B(\hat{\mathbf{w}}) - (y(\mathbf{x},D) - \mathbb{E}_D\left[y(\mathbf{x},D)\right]))^2\right]$$

$$L_{av}(h(\mathbf{x},\mathbf{w}),y(\mathbf{x},\hat{\mathbf{w}})) = \mathbb{E}_D\left[B(\hat{\mathbf{w}})^2\right] + 2\cdot\mathbb{E}_D\left[B(\hat{\mathbf{w}})\cdot(y(\mathbf{x},D) - \mathbb{E}_D\left[y(\mathbf{x},D)\right])\right]$$

$$+\mathbb{E}_D\left[(y(\mathbf{x},D) - \mathbb{E}_D\left[y(\mathbf{x},D)\right])^2\right]$$

The expectation is zero

$$\mathbb{E}_D\left[B(\hat{\mathbf{w}})\cdot(y(\mathbf{x},D) - \mathbb{E}_D\left[y(\mathbf{x},D)\right])\right] = 0$$

The expectation is zero

$$\mathbb{E}_D\left[B(\hat{\mathbf{w}})\cdot(y(\mathbf{x},D)-\mathbb{E}_D\left[y(\mathbf{x},D)\right])\right]=0$$

means

$$\mathbb{E}_D\left[(\mathbb{E}_D\left[y(\mathbf{x},D)\right]-\mathbb{E}[t|\mathbf{x}])\cdot(y(\mathbf{x},D)-\mathbb{E}_D\left[y(\mathbf{x},D)\right])\right]=$$

$$\mathbb{E}_D\left[y(\mathbf{x},D)\cdot\mathbb{E}_D\left[y(\mathbf{x},D)\right]\right]-\mathbb{E}_D\left[(\mathbb{E}_D\left[y(\mathbf{x},D)\right])^2\right]$$

$$-\mathbb{E}_D\left[\mathbb{E}[t|\mathbf{x}]\cdot(y(\mathbf{x},D)\right]+\mathbb{E}_D\left[\mathbb{E}[t|\mathbf{x}]\cdot\mathbb{E}_D\left[y(\mathbf{x},D)\right]\right]=0$$

Because

$$\mathbb{E}_D\left[y(\mathbf{x},D)\cdot\mathbb{E}_D\left[y(\mathbf{x},D)\right]\right]=y(\mathbf{x},D)\cdot\mathbb{E}_D\left[y(\mathbf{x},D)\right]$$

$$\mathbb{E}_D\left[(\mathbb{E}_D\left[y(\mathbf{x},D)\right])^2\right]=(\mathbb{E}_D\left[y(\mathbf{x},D)\right])^2$$

$$\mathbb{E}_D\left[\mathbb{E}[t|\mathbf{x}]\cdot(y(\mathbf{x},D)\right]=y(\mathbf{x},D)\cdot\mathbb{E}_D\left[y(\mathbf{x},D)\right]$$

$$\mathbb{E}_D\left[\mathbb{E}[t|\mathbf{x}]\cdot\mathbb{E}_D\left[y(\mathbf{x},D)\right]\right]=(\mathbb{E}_D\left[y(\mathbf{x},D)\right])^2$$

Bias-Variance Dilemma

$$L_{av}(h(\mathbf{x},\mathbf{w}), y(\mathbf{x},\hat{\mathbf{w}})) = \left(\mathbb{E}_D\left[y(\mathbf{x},D)\right] - \mathbb{E}[t|\mathbf{x}]\right)^2 + \mathbb{E}_D\left[\left(y(\mathbf{x},D) - \mathbb{E}_D\left[y(\mathbf{x},D)\right]\right)^2\right]$$

with

$$bias: \;\; B(\hat{\mathbf{w}}) = \mathbb{E}_D\left[y(\mathbf{x},D)\right] - \mathbb{E}[t|\mathbf{x}]$$

$$variance: \;\; V(\hat{\mathbf{w}}) = \mathbb{E}_D\left[\left(y(\mathbf{x},D) - \mathbb{E}_D\left[y(\mathbf{x},D)\right]\right)^2\right]$$

$$L_{av}(h(\mathbf{x},\mathbf{w}), y(\mathbf{x},\hat{\mathbf{w}})) = (B(\hat{\mathbf{w}}))^2 + V(\hat{\mathbf{w}}) = (bias)^2 + variance$$

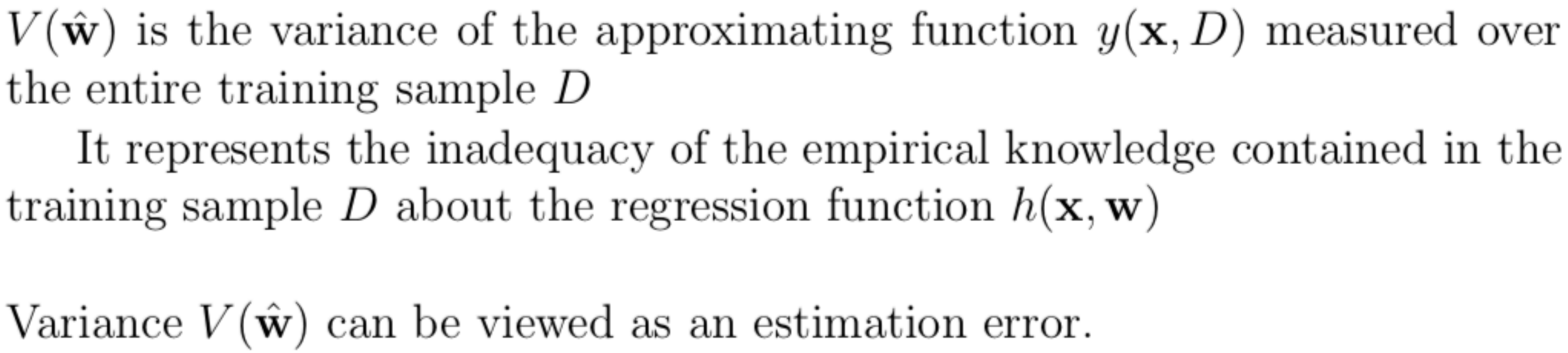
Bias

The first term, $B(\hat{\mathbf{w}})$ is the bias of the average value of the approximation function $y(\mathbf{x}, D)$ measured with respect to the regression function $h(\mathbf{x}, \mathbf{w}) = \mathbb{E}[t|\mathbf{x}]$

$B(\hat{\mathbf{w}})$ represents the inability of the physical model defined by the function $y(\mathbf{x}, D)$ to accurately approximate the regression function $h(\mathbf{x}, \mathbf{w}) = \mathbb{E}[t|\mathbf{x}]$

The bias $B(\hat{\mathbf{w}})$ can be viewed as an approximation error.

Variance



$V(\hat{\mathbf{w}})$ is the variance of the approximating function $y(\mathbf{x}, D)$ measured over the entire training sample D

It represents the inadequacy of the empirical knowledge contained in the training sample D about the regression function $h(\mathbf{x}, \mathbf{w})$

Variance $V(\hat{\mathbf{w}})$ can be viewed as an estimation error.

Interpretation

- In a complex physical model that learns by example and does so with a training sample of limited size, the price for achieving a small bias is a large variance
- For any physical model, it is only when the size of the training sample becomes infinitely large that we can hope to eliminate both bias and variance at the same time
- Our goal is to minimize the expected loss, which we have decomposed into the sum of a (squared) bias, a variance, and a constant noise term

- There is a trade-off between bias and variance, with very flexible models having low bias and high variance
- Rigid models having high bias and low variance
- The model with the optimal predictive capability is the one that leads to the best balance between bias and variance

Example

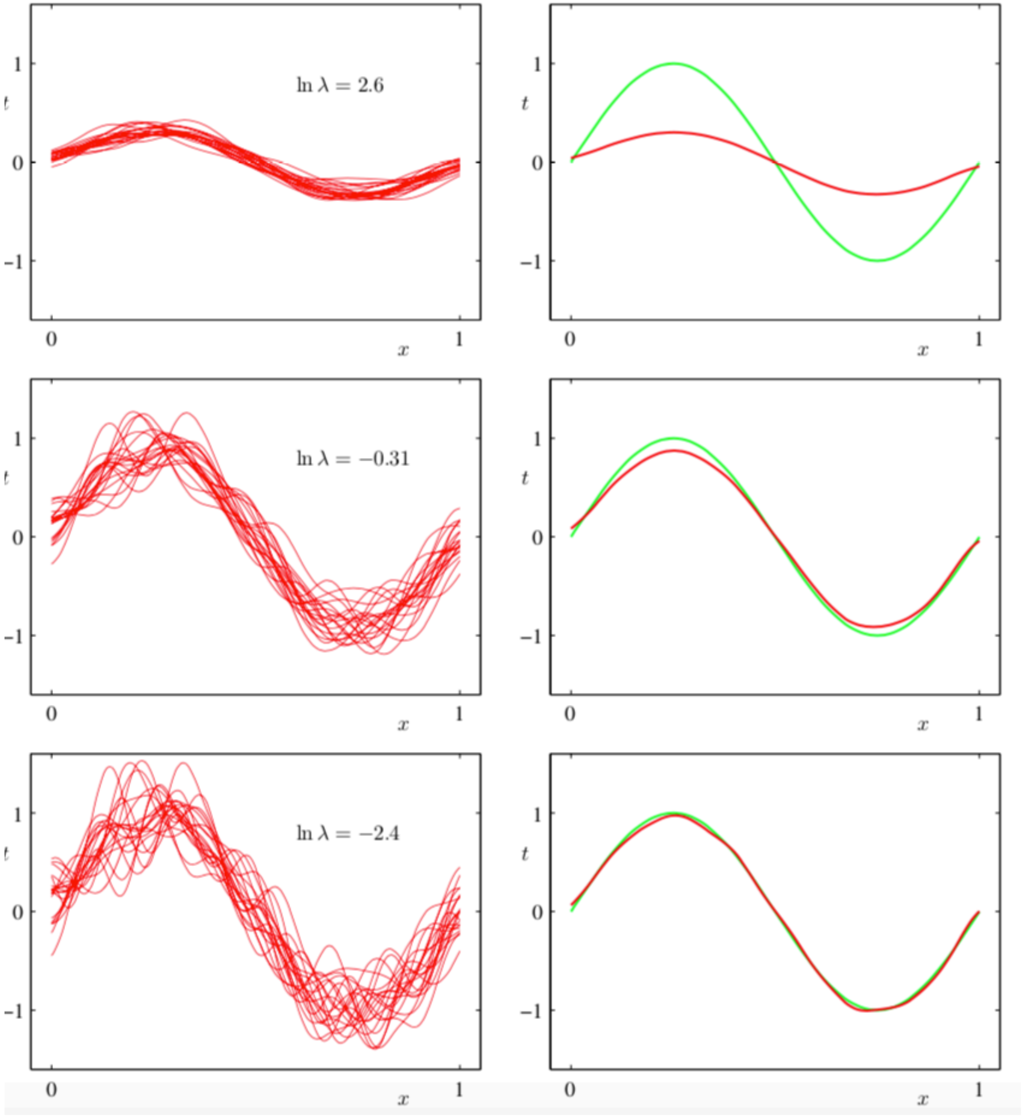
- There are *L = 100* data sets, each having *N = 25* data points, and there are *24* Gaussian basis functions in the model so that the total number of parameters is *M = 25* including the *“bias” (not statistical, remember?)* parameter

$$y(x, \mathbf{w}) = w_0 + \sum_{j=1}^{24} w_j \cdot x^j = \sum_{j=0}^{25} \phi_j(x)$$

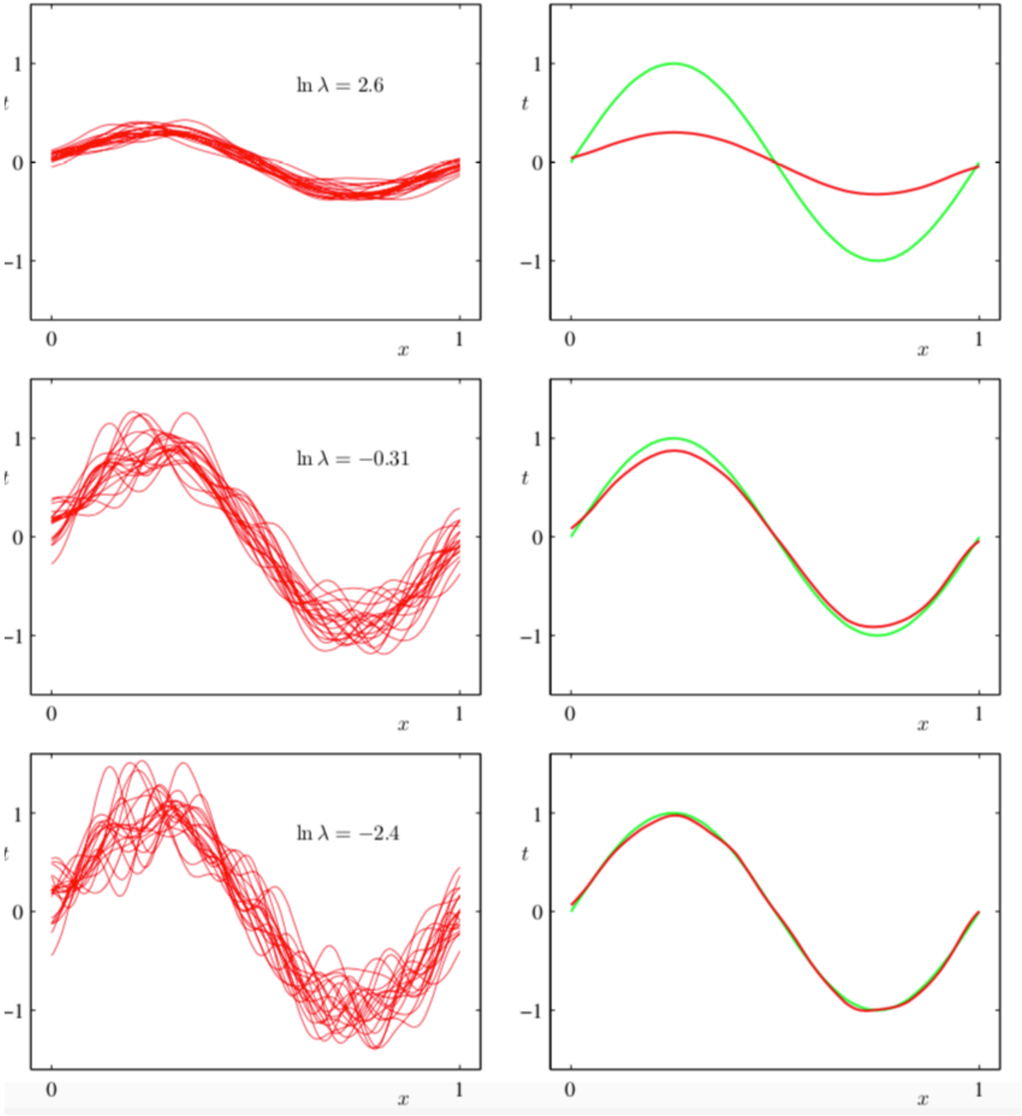
We will minimizing the regularized error function to give a prediction function

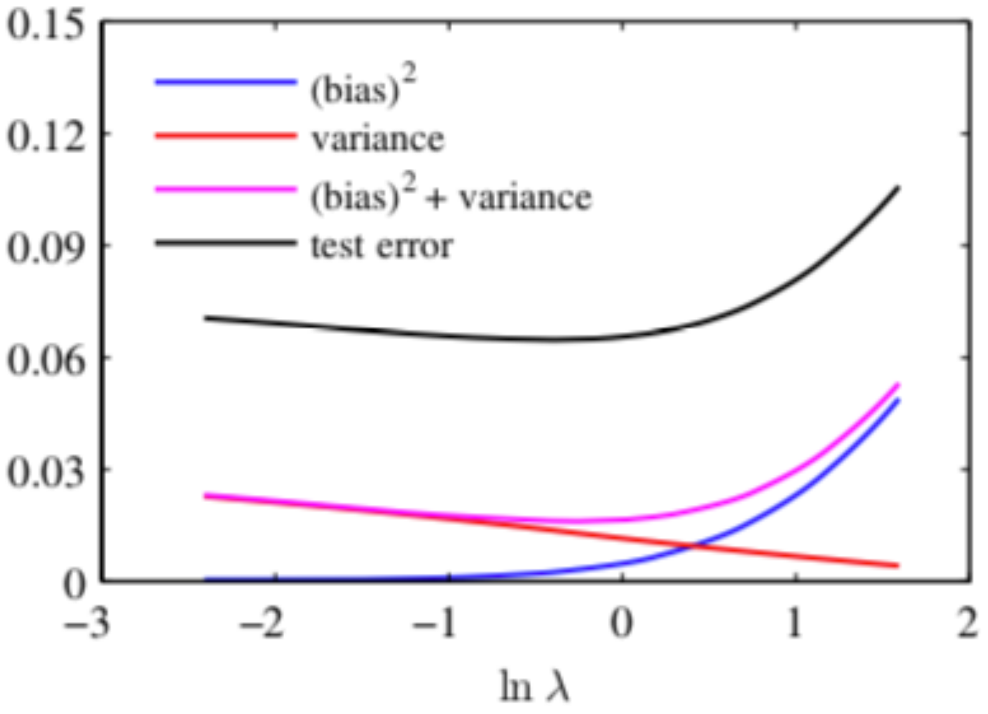
$$E(\mathbf{w}) = \frac{1}{2} \cdot \sum_{\eta=1}^{N} (t_\eta - \mathbf{w}^T \cdot \phi(\mathbf{x}_\eta))^2 + \frac{\lambda}{2} \|\mathbf{w}\|_2^2$$

- A large value of the regularisation coefficient λ that gives low variance (because the red curves in the left plot look similar) but high bias (because the two curves in the right plot are very different)

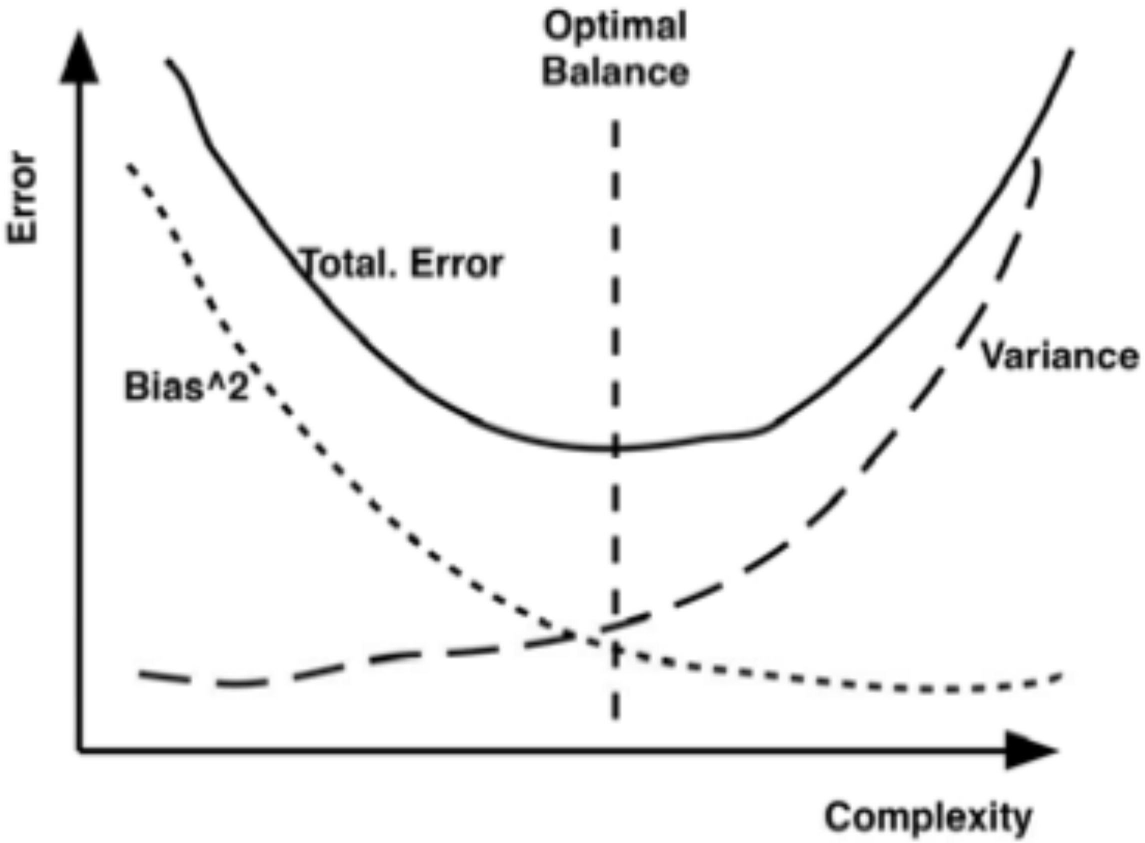


- A small value of the regularisation coefficient λ that gives large variance (shown by the high variability between the red curves in the left plot) but low bias (shown by the good fit between the average model fit and the original sinusoidal function)





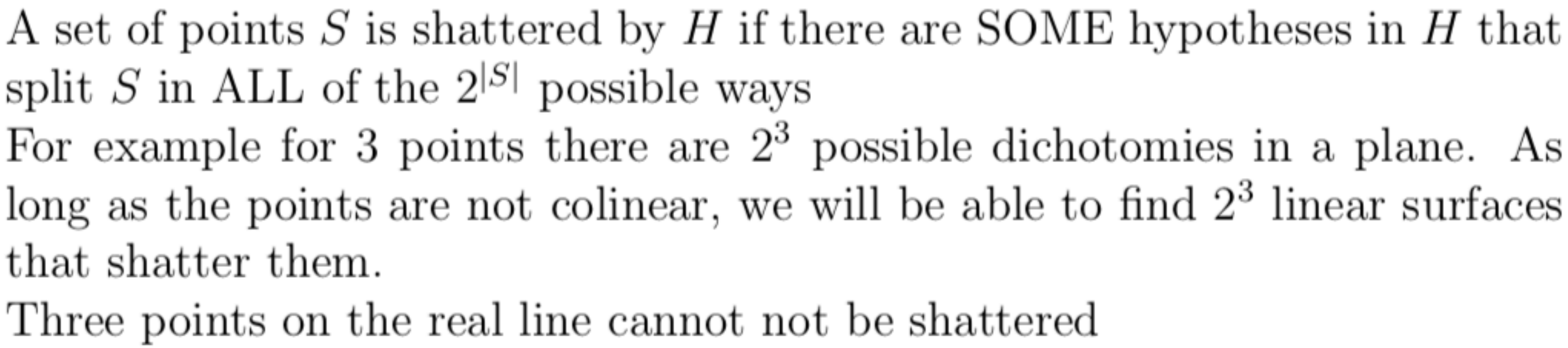
- The model with the optimal predictive capability is the one that leads to the best balance between bias and variance


Optimal
Balance
Error
Total. Error
Variance
Bias^2
Complexity

The VC Dimension

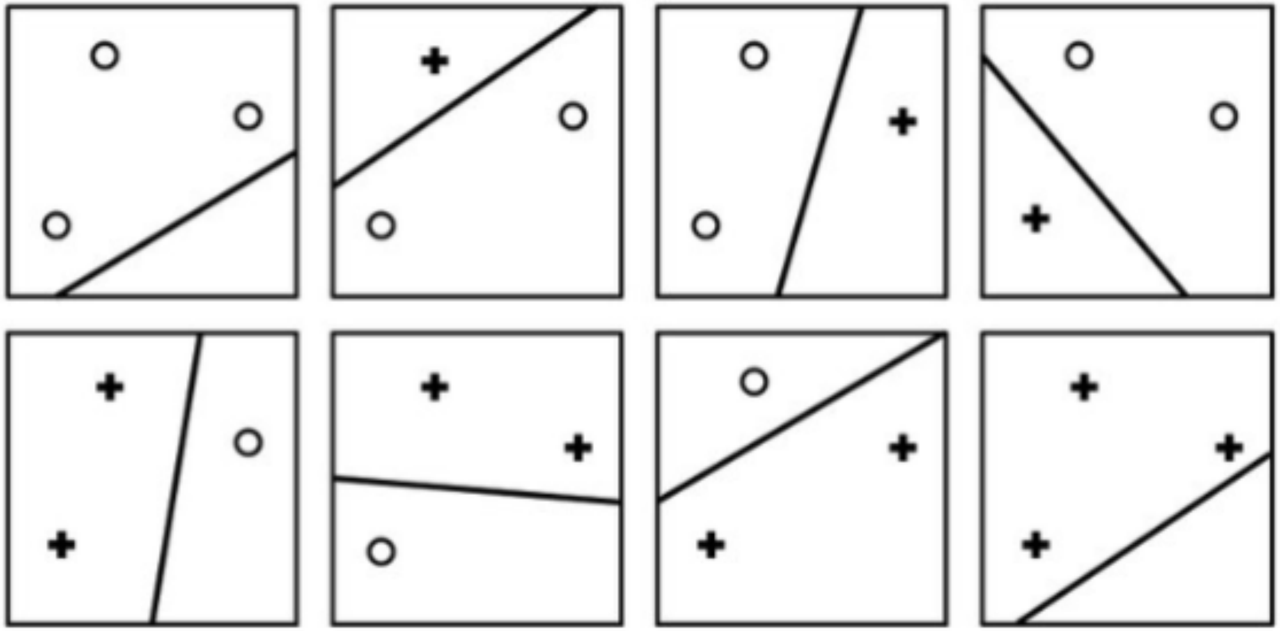
- The Vapnik-Chervonenkis dimension, or VC dimension.
  The VC dimension measures the capacity of a binary classifier.
- A dichotomy is a partition of a whole (or a set) into two parts (subsets). *From Ancient Greek: equally divided, cut in half*
- A set of instances *S* is shattered by hypothesis space *H* if and only if for **EVERY** dichotomy of *S* there exists *SOME* hypothesis in *H* consistent with this dichotomy

- A set of instances *S* is shattered by hypothesis space *H* if and only if for **EVERY** dichotomy of *S* there exists *SOME* hypothesis in *H* consistent with this dichotomy



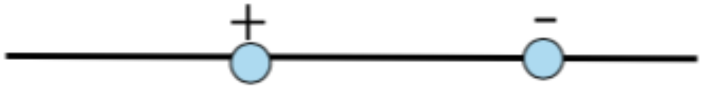
A set of points S is shattered by H if there are SOME hypotheses in H that split S in ALL of the $2^{|S|}$ possible ways
For example for 3 points there are 2^3 possible dichotomies in a plane. As long as the points are not colinear, we will be able to find 2^3 linear surfaces that shatter them.
Three points on the real line cannot not be shattered

Shattering of three Points

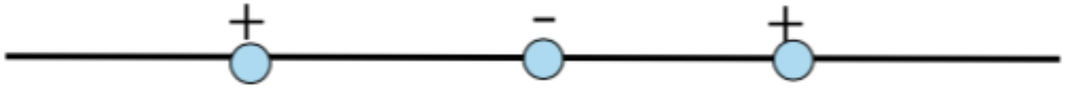


Shattering on a Line

- Yes

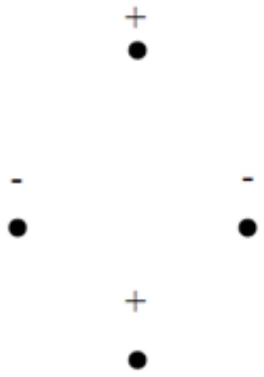


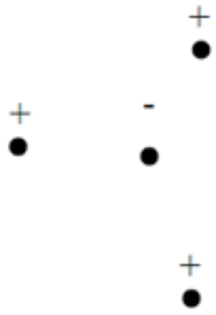
- No



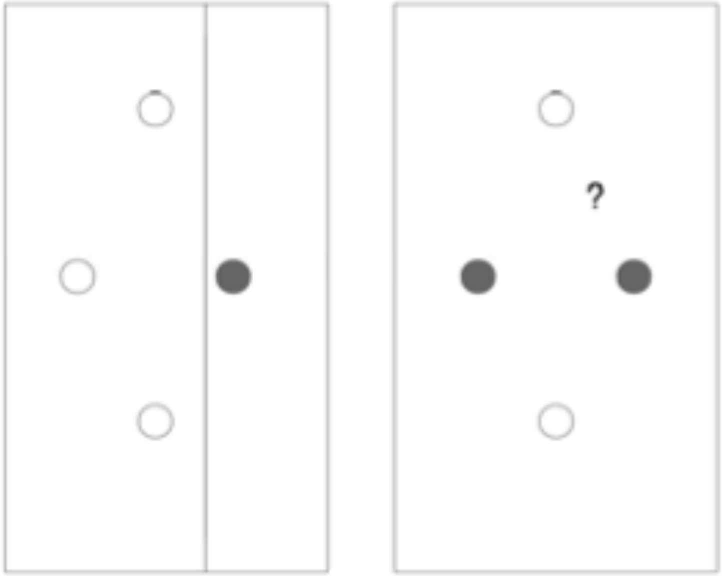
Cannot be Shattered

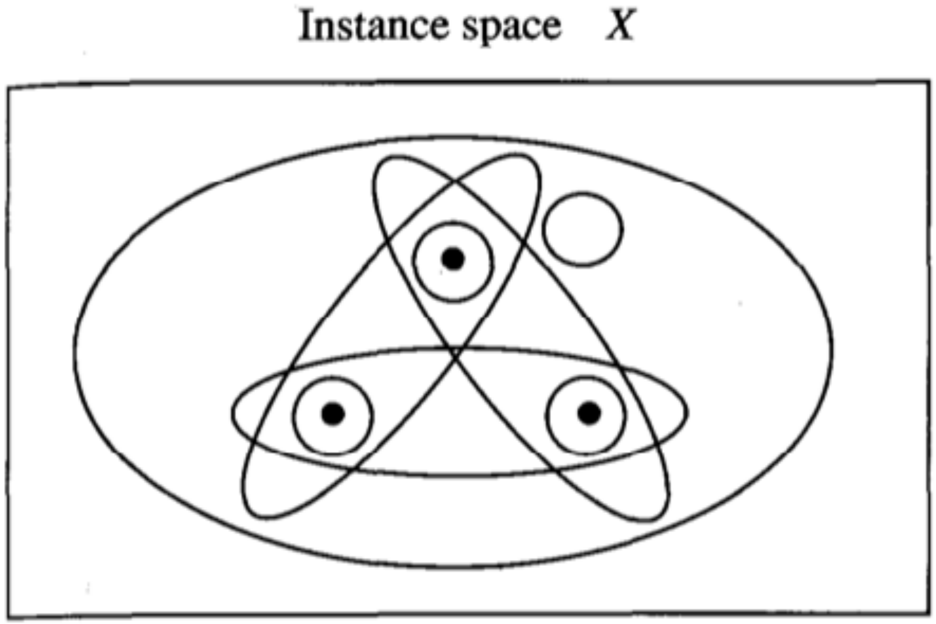
- Four points on a plane, two examples
- i) ii)





On a plane four points cannot be shattered



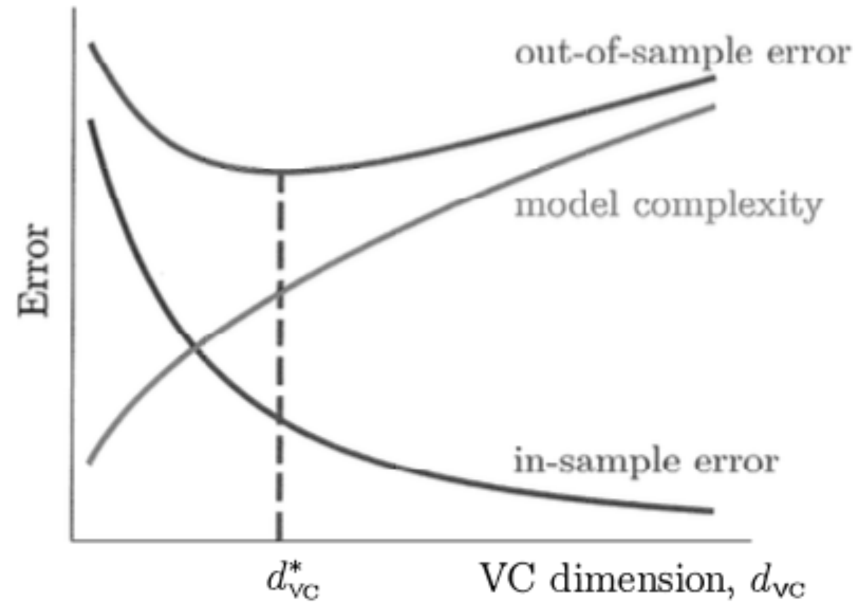

Instance space X

VC-dimension

- The VC-dimension of a hypothesis space *H* is the cardinality of the largest set S that can be shattered by *H*

- It can be shown that the VC dimension of linear decision surfaces in an *d* dimensional space (i.e., the VC dimension of a perceptron with *d* inputs) is *d + 1*

- Perceptron in *d* dimensions has *d+1* parameter (bias) Through *d+1* linear independent chosen points we can learn all dichotomies
- For *d+2* points in a perceptron in *d* dimension some vectors (at least two) are represented as a linear combination, we cannot learn all dichotomies



When we use a more complex learning model, one that has higher VC dimension d_{vc}, we are likely to fit the training data better resulting in a lower in sample error, but we pay a higher penalty for model complexity. A combination of the two, which estimates the out of sample error, thus attains a minimum at some intermediate d^*_{vc}

$$m \geq \frac{1}{\epsilon}(4 \cdot \log_2(2/\delta) + VC(H) \cdot log_2(12/\epsilon))$$

This number m of training examples is sufficient to assure that any consistent hypothesis will be probably (with probability $(1-\delta)$) approximately (within error ϵ) correct.

Blumer, A., Ehrenfeucht, A., Haussler, D., and Warmuth, M. (1989). Learnability and the Vapnik- Chemonenkis dimension. Journal of the ACM, 36(4) (October), 929-965.

`https://scholar.google.pt/citations?user=4esyQS4AAAAJ&hl=pt-PT`

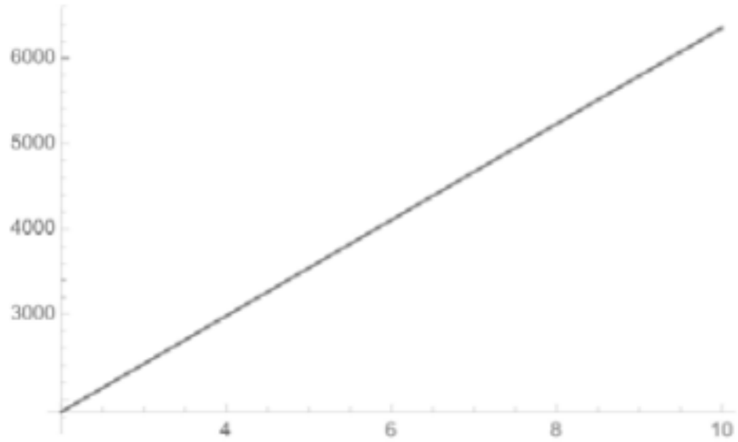


Figure 7.15: Estimation of the number N of training examples for $d = 2$ to $d = 10$ for a perceptron with $\epsilon = 0.1$ and $\delta = 0.1$.

Perceptron: For a perceptron [Mitchell, 1997]

$$N \geq \frac{1}{\epsilon}\left(4 \cdot \log_2(2/\delta) + 8 \cdot (d+1) \cdot log_2(13/\epsilon)\right)$$

since

$$VC(H^{perceptron}) = d + 1$$

- Neural Network: For acyclic layered network *G* containing *s* perceptrons [Kearns and Vazirani, 1994], [Mitchell, 1997] each with d inputs we have

$$VC(H_G^{perceptron}) \leq 2 \cdot (d+1) \cdot s \cdot \log(e \cdot s) = 2 \cdot (d+1) \cdot s \cdot (\log(s)+1)$$

$$N \geq \frac{1}{\epsilon}\left(4 \cdot \log_2(2/\delta) + 16 \cdot (d+1) \cdot s \cdot \log(e \cdot s) \cdot log_2(13/\epsilon)\right)$$

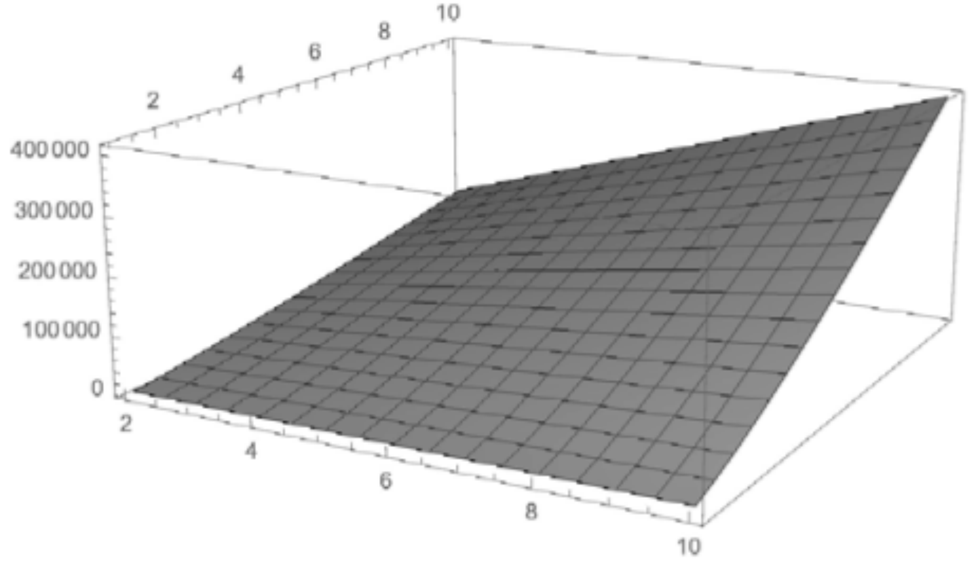
Example: For example for a network of seven units, each with an input of five, and $\epsilon = 0.01$ and $\delta = 0.01$

$$N \geq 2051800$$

which is a huge number. If we reduce the accuracy to $\epsilon = 0.1$ and $\delta = 0.1$

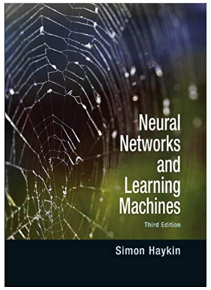
$$N \geq 139191$$

which is still a huge number. In backpropagation algorithm the VC dimension is usually lower, and due to regularization we can reduce the number considerably.

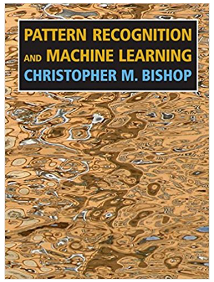


Estimation of the number N of training examples for *d = 2* to *d = 10* and *s = 1* (units) to *s = 10* for a neural network with *ε= 0.1* and *δ = 0.1*. In backpropagation algorithm the VC dimension is usually lower, and due to regularization we can reduce the number considerably.

Literature

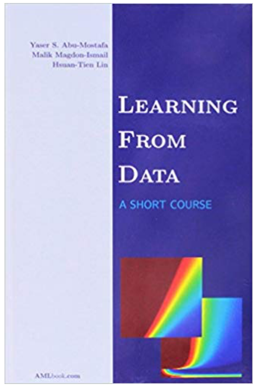


- Simon O. Haykin, Neural Networks and Learning Machine, (3rd Edition), Pearson 2008
    - Chapter 2, Section 2.7



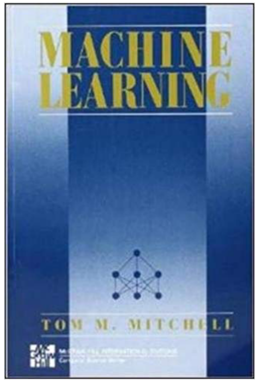
- Christopher M. Bishop, Pattern Recognition and Machine Learning (Information Science and Statistics), Springer 2006
    - Chapter 3, Section 3.2

Literature (Additional)



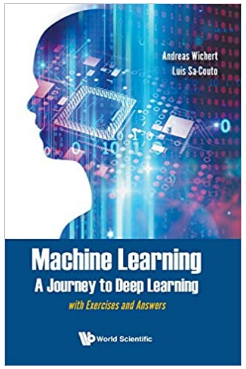
- Yaser S. Abu-Mostafa, Malik Magdon-Ismail, Hsuan-Tien Lin, Learning from Data, AMLBook (2012)

- https://work.caltech.edu/telecourse



- Tom M. Mitchell, Machine Learning, McGraw-Hill; 1st edition (October 1, 1997)

Literature



- Machine Learning - A Journey to Deep Learning, A. Wichert, Luis Sa-Couto, World Scientific, 2021
  - Chapter 7