Machine Intelligence 1

2.1 Elements of Statistical Learning Theory

Prof. Dr. Klaus Obermayer

Fachgebiet Neuronale Informationsverarbeitung (NI)

WS 2017/2018

Background

Central question

What must one know a priori about an unknown functional dependency in order to estimate it on the basis of observations? (Vapnik 1995)

- Regularization principles for solving ill-posed problems (e.g. Tikhonov)
- Nonparametric statistics (e.g. Parzen)
- Law of large numbers in functional space (Vapnik & Chervonenkis)
- Algorithmic complexity (Kolmogorov, Solomonoff, Chaitin)

Empirical Risk Minimization (ERM)

$$E^G_{[\underline{\mathbf{w}}]} \;=\; \int P_{(y_T,\underline{\mathbf{x}})}\, e_{(y_T,\underline{\mathbf{x}};\underline{\mathbf{w}})}\, d\underline{\mathbf{x}}\, dy_T \;\overset{!}{=}\; \min \qquad \text{(generalization error)}$$

mathematical expectation

$\Downarrow$ ERM $\Downarrow$ $\qquad\leftarrow$ When does this work?

empirical average

$$E^T_{[\underline{\mathbf{w}}]} \;=\; \frac{1}{p}\sum_{\alpha=1}^{p} e_{(y_T^{(\alpha)},\underline{\mathbf{x}}^{(\alpha)};\underline{\mathbf{w}})} \;\overset{!}{=}\; \min \qquad \text{(training error)}$$

The law of large numbers

> *"The sequence of means converges to the expectation of a random variable (if it exists) as the number p of samples increases."*

- can only be applied if set of predictors has a finite number of elements (use Hoeffding's inequality for bound on deviations)
- extension to infinite sets (e.g. MLPs) necessary

Statistical learning theory (SLT)

Goal: Estimate a desired function from

- a wide set of functions
- a limited number of examples

Statistical learning theory (SLT)

Goal: Estimate a desired function from

- a wide set of functions
- a limited number of examples

Achievements of statistical learning theory

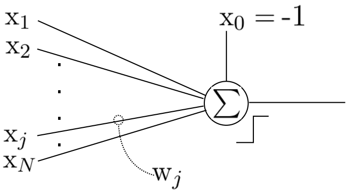
1. formulation of conditions under which ERM works
2. bounds describing **generalization ability** of ERM
3. inductive inference for **small sample size**s based on these bounds
4. methods for implementing this new type of inference ($\rightarrow$ **SVMs**)

2.1.1 Linear Classifiers: An Example

Linear classifiers

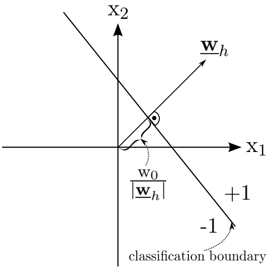
- **data representation:** samples $\underline{\mathbf{x}}^{(\alpha)} \in \mathbb{R}^N$ and classes $y_T^{(\alpha)} \in \{-1, 1\}$
- **model class:** binary connectionist neuron

binary connectionist neuron



$$y(\underline{\mathbf{x}}) = \operatorname{sign}(\underline{\mathbf{w}}^T \underline{\mathbf{x}})$$

classification boundary



x_2
$\underline{\mathbf{w}}_h$
x_1
$\frac{\mathrm{w}_0}{|\underline{\mathbf{w}}_h|}$
+1
-1
classification boundary

$$\underline{\mathbf{w}}^T \underline{\mathbf{x}} = 0$$

Optimization with ERM

- **performance measure:** classification error
- **optimization:** minimizing the training error (ERM)
  - which conditions yield small training errors?
- **validation:** evaluate test error on unseen data
  - which conditions yield small generalization errors?

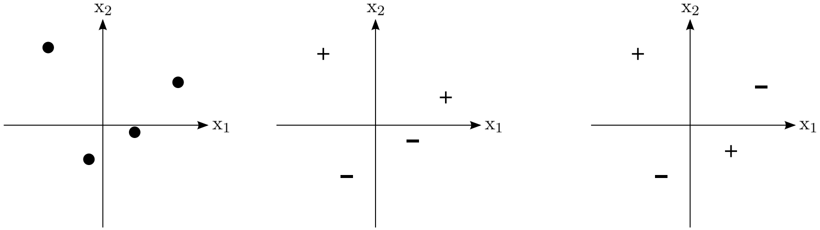
Linear separability

$\rightsquigarrow$ classes do not overlap

$\rightsquigarrow$ classes can be separated by a hyperplane

} all data points can be correctly classified by at least one classifier from the set

- observations $\underline{\mathbf{x}}^{(\alpha)} \in \mathbb{R}^N, \alpha = 1, \ldots, p$, in *general location* (→ each subset of N points should be linearly independent)



in general location   separable assignment   ass. not linearly separable

- 14 linearly separable configurations, 2 configurations not separable

Statistics of linear separability (Cover, 1965)

(1) **number** $C_{(p,N)}$ of linearly separable assignments $y_T : \underline{\mathbf{x}} \to y_T$:

$$C_{(p,N)} = 2 \sum_{k=0}^{N-1} \binom{p-1}{k}$$

Statistics of linear separability (Cover, 1965)

(1) **number** $C_{(p,N)}$ of linearly separable assignments $y_T : \underline{\mathbf{x}} \to y_T$:

$$C_{(p,N)} = 2 \sum_{k=0}^{N-1} \binom{p-1}{k}$$

(2) **fraction** $\Pi_{(p,N)}$ of linearly separable assignments (total: 2^p):

$$\Pi_{(p,N)} = \frac{C_{(p,N)}}{2^p} = \frac{1}{2^{p-1}} \sum_{k=0}^{N-1} \binom{p-1}{k}$$

Asymptotic distribution

$$\Pi_{(p,N)} = \sum_{k=0}^{N-1} \overbrace{\binom{p-1}{k} \left(\frac{1}{2}\right)^k \left(\frac{1}{2}\right)^{p-1-k}}^{f(k\,|\,p-1,\frac{1}{2})}$$

$$f(k|n,q) = \binom{n}{k} q^k (1-q)^{n-k} \qquad \text{(Binomial distribution)}$$

- $\Pi_{(p,N)}$ is the CDF of the Binomial distribution $f(k\,|\,p-1,\frac{1}{2})$

Asymptotic distribution

$$\Pi_{(p,N)} = \sum_{k=0}^{N-1} \overbrace{\binom{p-1}{k} \left(\frac{1}{2}\right)^k \left(\frac{1}{2}\right)^{p-1-k}}^{f(k\,|\,p-1,\frac{1}{2})}$$

$$f(k|n,q) = \binom{n}{k} q^k (1-q)^{n-k} \qquad \text{(Binomial distribution)}$$

- $\Pi_{(p,N)}$ is the CDF of the Binomial distribution $f(k\,|\,p-1, \frac{1}{2})$

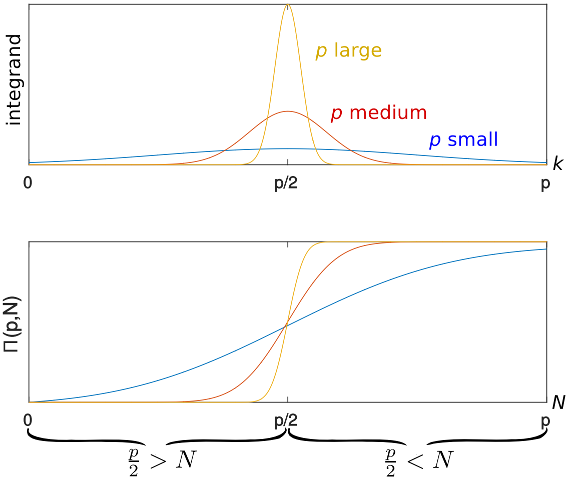
Binomial distribution converges to Gaussian distribution

$$f(k\,|\,n,q) \;\approx\; \mathcal{N}(k\,|\,nq,\, nq(1-q)) \qquad \text{(asymptotic distribution)}$$

In the limit $N, p \to \infty$, with $\frac{p}{N} = \text{const}$, one obtains:

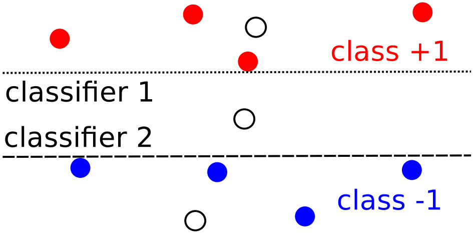
$$\Pi_{(p,N)} \approx \sum_{k=0}^{N-1} \mathcal{N}(k\,|\,\tfrac{p-1}{2}, \tfrac{p-1}{4}) \underset{p\to\infty}{\approx} \int_{-\infty}^{N} dk\, \exp\Big(-\tfrac{1}{p/2}(k-\tfrac{p}{2})^2\Big)$$

Capacity of the set of linear classifiers

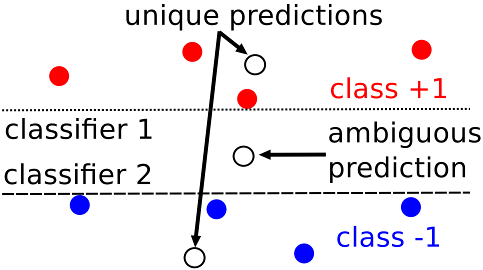


- all 2^p label configurations can be realized below $\beta := \frac{p}{N} = 2$
- β can be understood as a natural capacity measure

Generalization performance (1)



Generalization performance (1)



- $C_{(p,N-1)}$: number of linearly separable assignments on $p-1$ data points, where there are still two choices left for the remaining data point (Cover, 1965).

Generalization performance (2)

- number of ambiguous, linearly separable assignments: $C_{(p,N-1)}$
- fraction of ambiguous, linearly separable assignments

$$g_{(p,N)} := \frac{\text{ambiguous assignments}}{\text{all separable assignments}} = \frac{C_{(p,N-1)}}{C_{(p,N)}}$$

- if fraction of ambiguous assignments is large, prediction for new data points will not be informative

Generalization performance (2)

- number of ambiguous, linearly separable assignments: $C_{(p,N-1)}$

- fraction of ambiguous, linearly separable assignments

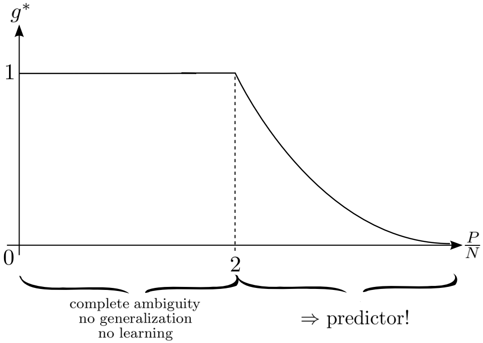
$$g_{(p,N)} := \frac{\text{ambiguous assignments}}{\text{all separable assignments}} = \frac{C_{(p,N-1)}}{C_{(p,N)}}$$

- if fraction of ambiguous assignments is large, prediction for new data points will not be informative

- asymptotic result (Cover, 1965):

$$g^*_{(\frac{p}{N})} = \lim_{\substack{p,N\to\infty, \\ \frac{p}{N}=\text{const}}} g_{(p,N)} = \begin{cases} 1, & \text{for } 0 \leq \frac{p}{N} \leq 2 \\ \frac{1}{\frac{p}{N}-1}, & \text{for } \frac{p}{N} > 2 \end{cases}$$

How much data do we need to predict?



- for $p/N \leq 2$ almost all predictions are ambiguous.
- for $p/N > 2$ classifier is able to generalize to new data points

"The probability of ambiguous generalization is large unless the number of training patterns exceeds the capacity of the set of separating surfaces."

The Vapnik-Chervonenkis (VC) dimension

key question of inductive reasoning

When will conclusions inferred from a sample of data be meaningful?
→ concept of non-falsifiability (Popper)

- a theory is useful only, if it is falsifiable (cmp. astrology),
- i.e. if it cannot predict all potentially observable patterns.

The Vapnik-Chervonenkis (VC) dimension

key question of inductive reasoning

When will conclusions inferred from a sample of data be meaningful?
→ concept of non-falsifiability (Popper)

- a theory is useful only, if it is falsifiable (cmp. astrology),
- i.e. if it cannot predict all potentially observable patterns.

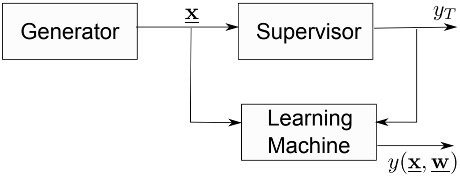
generalization of the capacity measure

VC-dimension d_{VC} describes the capacity of a family of functions $\mathcal{M}$

- d_{VC} is the **largest number** of samples in general locations...
  - that can be classified by some $f \in \mathcal{M}$ for any possible labeling
  - that can be classified by some $f \in \mathcal{M}$ without any empirical error

2.1.2 Formulation of the Inductive Learning Problem

Inductive learning for binary classification



- generator draws $\underline{\mathbf{x}} \in R^n$ from unknown $P(\underline{\mathbf{x}})$
- supervisor returns output $y_T \in \{-1, +1\}$ to every $\underline{\mathbf{x}}$ from unknown $P(y_T|\underline{\mathbf{x}})$
- learning machine estimates function $y(\underline{\mathbf{x}}, \underline{\mathbf{w}}) \approx y_T$ by selecting $\underline{\mathbf{w}} \in \Lambda$ from the set of possible parameters Λ
- performance is measured with 0-1 loss function

$$e_{(y_T, \underline{\mathbf{x}}; \underline{\mathbf{w}})} \quad = \quad \frac{1}{2}\big(1 - y_T \; y(\underline{\mathbf{x}}, \underline{\mathbf{w}})\big)$$

Empirical Risk Minimization (ERM)

$$e_{(y_T, \underline{\mathbf{x}}; \underline{\mathbf{w}})} \quad = \quad \frac{1}{2}\big(1 - y_T \; y(\underline{\mathbf{x}}, \underline{\mathbf{w}})\big)$$

$$\begin{aligned} E^G_{[\underline{\mathbf{w}}]} &= \text{probability of misclassification} \\ &= \int P_{(y_T, \underline{\mathbf{x}})} \, e_{(y_T, \underline{\mathbf{x}}; \underline{\mathbf{w}})} \, d\underline{\mathbf{x}} \, dy_T \;\overset{!}{=}\; \min \qquad \text{(generalization error)} \end{aligned}$$

mathematical expectation

$\Downarrow$ ERM $\Downarrow$ ← When does this work?

empirical average

$$\begin{aligned} E^T_{[\underline{\mathbf{w}}]} &= \text{fraction of misclassifications} \\ &= \frac{1}{p} \sum_{\alpha=1}^{p} e_{(y_T^{(\alpha)}, \underline{\mathbf{x}}^{(\alpha)}; \underline{\mathbf{w}})} \;\overset{!}{=}\; \min \qquad \text{(training error)} \end{aligned}$$

Change in nomenclature ($\rightarrow$ Vapnik, 1998)

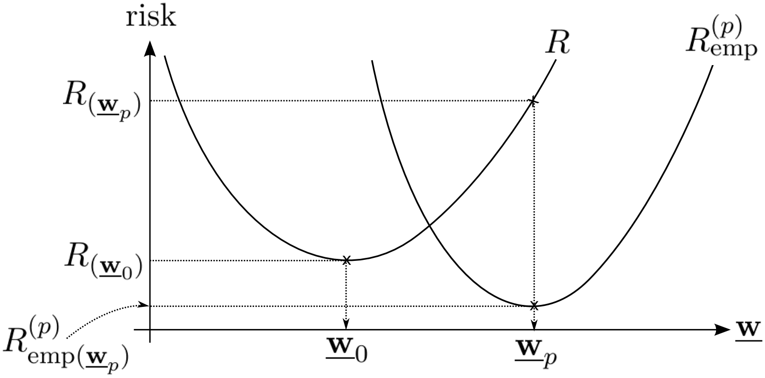
probability of misclassification: $E^G_{[\underline{\mathbf{w}}]} \rightarrow R_{[\underline{\mathbf{w}}]}$ (risk)

fraction of misclassification: $E^T_{[\underline{\mathbf{w}}]} \rightarrow R_{\text{emp}[\underline{\mathbf{w}}]}$ (empirical risk)

data representation

- p observations $\underline{\mathbf{z}}^{(\alpha)} := (\underline{\mathbf{x}}^{(\alpha)}, y_T^{(\alpha)}), \alpha \in \{1, \ldots, p\}$
- observations are drawn **i.i.d.** from an unknown distribution $P(\underline{\mathbf{x}}, y_T) = P(y_T|\underline{\mathbf{x}})\, P(\underline{\mathbf{x}})$.
- **stationarity:** training and test distributions are identical

The learning problem

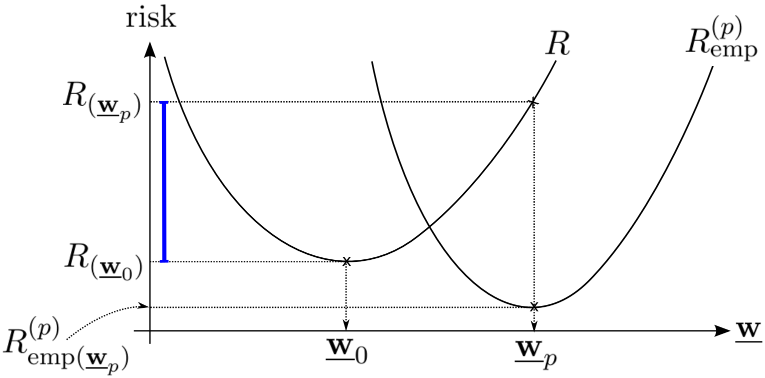

risk
R
$R_{\text{emp}}^{(p)}$
$R_{(\underline{\mathbf{w}}_p)}$
$R_{(\underline{\mathbf{w}}_0)}$
$R_{\text{emp}(\underline{\mathbf{w}}_p)}^{(p)}$
$\underline{\mathbf{w}}_0$
$\underline{\mathbf{w}}_p$
$\underline{\mathbf{w}}$

- samples is drawn *i.i.d.* from $P(\underline{\mathbf{x}}, y_T)$

  $\rightsquigarrow$ $R_{\text{emp}}^{(p)}$ is a random variable

  $\rightsquigarrow$ $\underline{\mathbf{w}}_p$ is a random variable

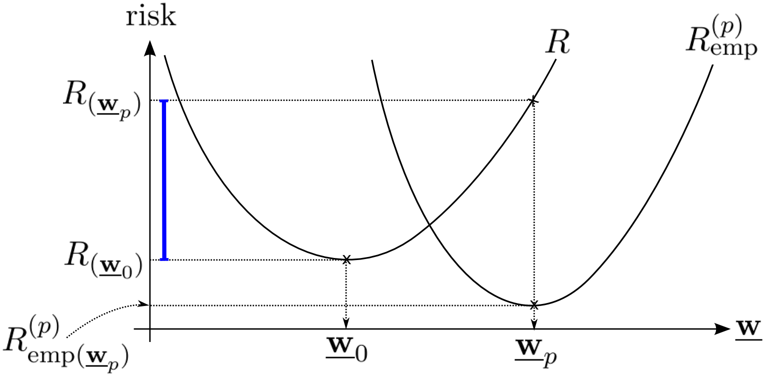
The key questions of Statistical Learning Theory



1. When does inductive learning through ERM work?
   - $R(\underline{\mathbf{w}}_p)$ should reflect the true optimal risk $R(\underline{\mathbf{w}}_0)$ in the limit $p \to \infty$:

$$\lim_{p\to\infty} P\Big\{ \big| R_{(\underline{\mathbf{w}}_p)} - R_{(\underline{\mathbf{w}}_0)} \big| \geq \eta \Big\} \;=\; 0\,, \quad \text{for all} \quad \eta > 0\,.$$

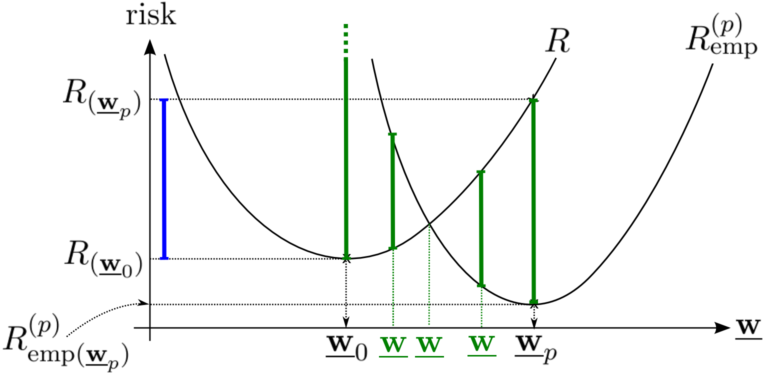
The key questions of Statistical Learning Theory



2. How strongly does $R_{(\underline{\mathbf{w}}_p)}$ differ from $R_{(\underline{\mathbf{w}}_0)}$ for **finite** samples?
   - For a given confidence ϵ, find η such that

$$P\left\{ \left| R_{(\underline{\mathbf{w}}_p)} - R_{(\underline{\mathbf{w}}_0)} \right| \geq \eta \right\} \;<\; \epsilon \,.$$

The key questions of Statistical Learning Theory


risk
R
$R_{\text{emp}}^{(p)}$
$R_{(\underline{\mathbf{w}}_p)}$
$R_{(\underline{\mathbf{w}}_0)}$
$R_{\text{emp}(\underline{\mathbf{w}}_p)}^{(p)}$
$\underline{\mathbf{w}}$
$\underline{\mathbf{w}}_0$
$\underline{\mathbf{w}}$
$\underline{\mathbf{w}}$
$\underline{\mathbf{w}}$
$\underline{\mathbf{w}}_p$

3. Do training errors $R_{\text{emp}(\underline{\mathbf{w}})}$ reflect generalization performance $R_{(\underline{\mathbf{w}})}$?
   - Avoid overfitting for finite samples, i.e. given confidence ϵ, find η such that:

$$P\left\{ \left| R_{(\underline{\mathbf{w}})} - R_{\text{emp}(\underline{\mathbf{w}})} \right| > \eta \right\} \;<\; \epsilon \,, \qquad \forall \underline{\mathbf{w}} \in \Lambda \,.$$

2.1.3 General Classifiers: Conditions for Inductive Learning

Conditions for successful learning with ERM

The success of ERM depends on...

- the model class (→ set of classifiers)
- the availability of data (→ growth function)

Conditions for successful learning with ERM

The success of ERM depends on...

- the model class ($\rightarrow$ set of classifiers)
- the availability of data ($\rightarrow$ growth function)

Key concepts: growth function & capacity measures (VC-dimension)

results (milestones)

1. asymptotic result: learning success for a potentially infinite number of training data
2. finite samples: deviation of the generalization error from the generalization error of the optimal model
3. finite samples: bound on the generalization error for every model

for technical details, conditions for non-trivial consistency etc. see Vapnik (1998)

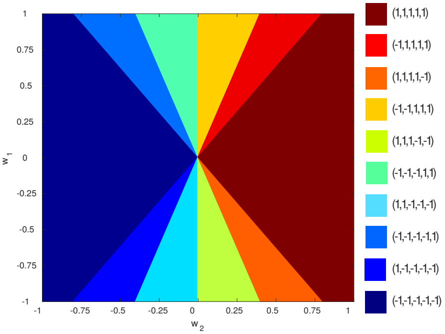
The growth function $G^{\Lambda}_{(p)}$

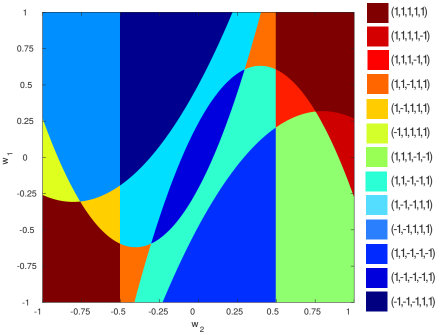
- **data representation:** $\underline{\mathbf{x}} \in \mathbb{R}^N, \quad y_T \in \{-1, +1\}$
- **model class:** set Λ of functions $y_{(\underline{\mathbf{x}};\underline{\mathbf{w}})} \in \{-1, +1\}$
- **binary label vector:** $\underline{\mathbf{y}}_{(\underline{\mathbf{w}})} = \Big(y_{(\underline{\mathbf{x}}^{(1)}, \underline{\mathbf{w}})}, y_{(\underline{\mathbf{x}}^{(2)}, \underline{\mathbf{w}})}, \ldots, y_{(\underline{\mathbf{x}}^{(p)}, \underline{\mathbf{w}})} \Big)$
  - different classifiers can induce the same label vector on the training set
- **number** of different vectors $\underline{\mathbf{y}}_{(\underline{\mathbf{w}})}$ induced by all $\hat{y} \in \Lambda$:

$$N^{\Lambda}_{(\underline{\mathbf{x}}^{(1)}, \ldots, \underline{\mathbf{x}}^{(p)})} \quad \leq 2^p \qquad \text{(depends on } \Lambda \text{ and the samples)}$$

The growth function $G^{\Lambda}_{(p)}$

5 samples $x^{(i)} \in \{-0.8, -0.4, 0, 0.4, 0.8\}$





$$f(x|\underline{\mathbf{w}}) = \text{sign}\Big(w_1\, x + w_2\Big) \qquad f(x|\underline{\mathbf{w}}) = \text{sign}\Big(w_1\, x + (x - w_2)^2 - \frac{1}{4}\Big)$$

$$N^{\Lambda}_{(x^{(1)},\ldots,x^{(5)})} \;=\; 10 \;\leq\; 2^5 \qquad N^{\Lambda}_{(x^{(1)},\ldots,x^{(5)})} \;=\; 13 \;\leq\; 2^5$$

The growth function $G^{\Lambda}_{(p)}$

$$G^{\Lambda}_{(p)} = \underbrace{\ln\left(\sup_{\underline{\mathbf{x}}^{(1)},\ldots\underline{\mathbf{x}}^{(p)}} N^{\Lambda}_{(\underline{\mathbf{x}}^{(1)},\ldots,\underline{\mathbf{x}}^{(p)})} \right)}_{\text{worst case}}$$

The growth function $G_{(p)}^{\Lambda}$

$$G_{(p)}^{\Lambda} = \underbrace{\ln \left(\sup_{\underline{\mathbf{x}}^{(1)}, \ldots \underline{\mathbf{x}}^{(p)}} N_{(\underline{\mathbf{x}}^{(1)}, \ldots, \underline{\mathbf{x}}^{(p)})}^{\Lambda} \right)}_{\text{worst case}}$$

- bound on growth function (Vapnik, 1998)

$$G_{(p)}^{\Lambda} \begin{cases} = p \ln 2 & \text{for } p \leq \mathrm{d}_{\mathrm{VC}} \\ \leq \mathrm{d}_{\mathrm{VC}} \Big(1 + \ln \frac{p}{\mathrm{d}_{\mathrm{VC}}} \Big) & \text{for } p > \mathrm{d}_{\mathrm{VC}} \end{cases}$$

The growth function $G_{(p)}^{\Lambda}$

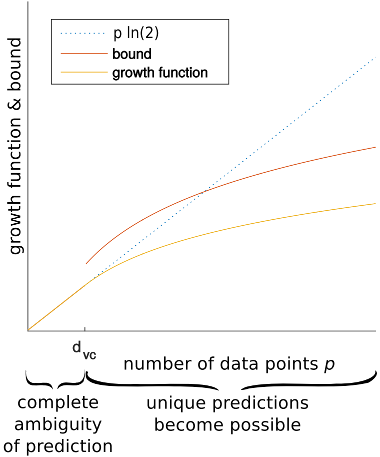
$$G_{(p)}^{\Lambda} = \underbrace{\ln\left(\sup_{\underline{\mathbf{x}}^{(1)},\ldots\underline{\mathbf{x}}^{(p)}} N_{(\underline{\mathbf{x}}^{(1)},\ldots,\underline{\mathbf{x}}^{(p)})}^{\Lambda}\right)}_{\text{worst case}}$$

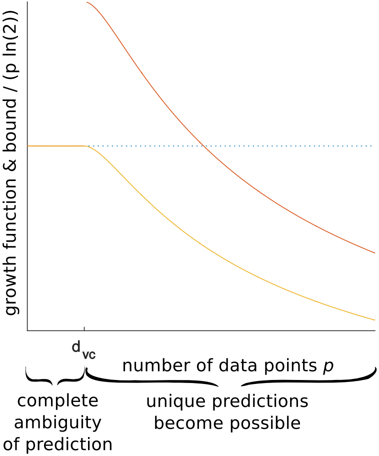
- bound on growth function (Vapnik, 1998)

$$G_{(p)}^{\Lambda} \begin{cases} = p \ln 2 & \text{for } p \leq \mathrm{d}_{\mathrm{VC}} \\ \leq \mathrm{d}_{\mathrm{VC}}\left(1 + \ln \frac{p}{\mathrm{d}_{\mathrm{VC}}}\right) & \text{for } p > \mathrm{d}_{\mathrm{VC}} \end{cases}$$

- Vapnik-Chernovenkis dimension d_{VC}: capacity measure of the model class

The growth function $G^{\Lambda}_{(p)}$





- d_{VC}: capacity measure of the model class
  - small d_{VC}: small sample size sufficient for learning
  - large d_{VC}: large sample size required

The Vapnik-Chervonenkis dimension

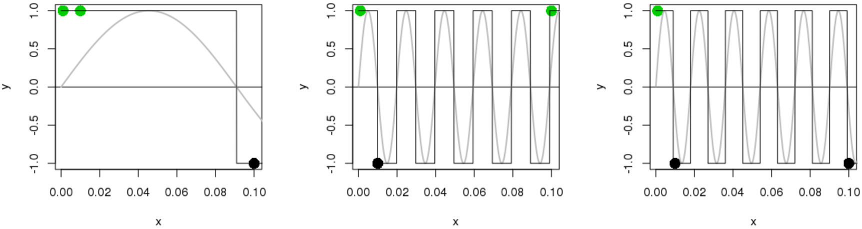
- d_{VC} is ...
  - largest number of data points on which (at least for one set) 2^p different loss vectors (i.e. label assignments) can be induced
  - capacity measure for a given set of models
  - purely combinatorial concept (worst case scenario)
  - provides a bound on the growth function ...
  - ... $\rightsquigarrow$ determines the number of samples needed to reliably learn

- d_{VC} is typically hard to compute
  - bounds have been derived for some architectures
  - e.g. MLPs with sigmoidal units: $\mathrm{d}_{\mathrm{VC}} = \mathcal{O}(W^2)$, where W is number of parameters, Koiran and Sontag (1996)

An unlearnable problem

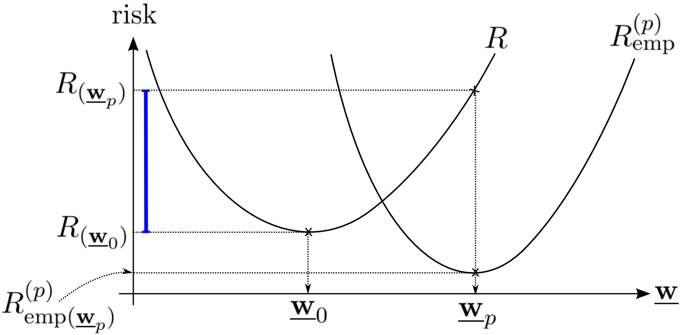
- set of classifiers: $\hat{y}(x;a) = \mathrm{sign}[\sin(ax)]$, single model parameter a

- selection of classifiers (cf. Vapnik, 1998)
  - take $x_i = 10^{-i},\ i = 1, ..., N$ and assign any $y_i \in \{-1, +1\}$
  - set $a = \pi(1 + \sum_{i=1}^{N} \frac{(1-y_i)10^i}{2})$



- a single model parameter, but an infinite d_{VC}

The learning problem and its solution (1)

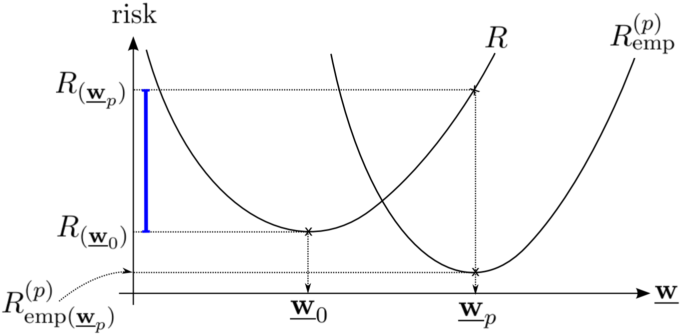

risk
R
$R_{\text{emp}}^{(p)}$
$R_{(\underline{\mathbf{w}}_p)}$
$R_{(\underline{\mathbf{w}}_0)}$
$R_{\text{emp}(\underline{\mathbf{w}}_p)}^{(p)}$
$\underline{\mathbf{w}}_0$
$\underline{\mathbf{w}}_p$
$\underline{\mathbf{w}}$

1. **learnability**: asymptotic result

$$\mathrm{d}_{\mathrm{VC}} \text{ finite} \quad \Rightarrow \quad \lim_{p \to \infty} P\left\{ \left| R_{(\underline{\mathbf{w}}_p)} - R_{(\underline{\mathbf{w}}_0)} \right| \geq \eta \right\} \;=\; 0\,, \quad \forall \eta > 0$$

- convergence of ERM is only guaranteed if d_{VC} is finite

The learning problem and its solution (2)

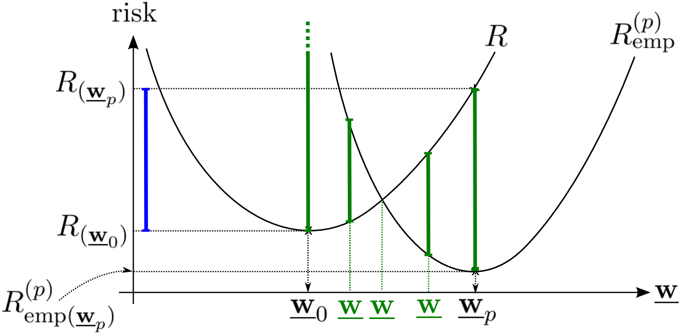


❷ **finite samples:** deviation from the optimal model

$$P\left\{ R_{(\underline{\mathbf{w}}_p)} - R_{(\underline{\mathbf{w}}_0)} > \left(\frac{G^{\Lambda}_{(2p)} - \ln \frac{\epsilon}{8}}{p} \right)^{\frac{1}{2}} + \left(- \frac{\ln \frac{\epsilon}{2}}{2p} \right)^{\frac{1}{2}} + \frac{1}{p} \right\} < \epsilon$$

- finite d_{VC}: increased sample size $p \Rightarrow$ reduction of the bound

The learning problem and its solution (3)


risk
R
$R_{\text{emp}}^{(p)}$
$R_{(\underline{\mathbf{w}}_p)}$
$R_{(\underline{\mathbf{w}}_0)}$
$R_{\text{emp}(\underline{\mathbf{w}}_p)}^{(p)}$
$\underline{\mathbf{w}}$
$\underline{\mathbf{w}}_0$
$\underline{\mathbf{w}}$
$\underline{\mathbf{w}}$
$\underline{\mathbf{w}}$
$\underline{\mathbf{w}}_p$

③ **finite samples:** bound on the generalization error

$$P\left\{ \sup_{\underline{\mathbf{w}} \in \Lambda} \left| R_{(\underline{\mathbf{w}})} - R_{\text{emp}(\underline{\mathbf{w}})}^{(p)} \right| > \eta \right\} < 4 \exp\left(G_{(2p)}^{\Lambda} - p\left(\eta - \frac{1}{p}\right)^2 \right)$$

- bound non-trivial only if $G_{(2p)}^{\Lambda}$ is sub-linear in p (see blackboard)

Regression problems

- regression problems can be treated in a similar manner

End of Section 2.1

the following slides contain

OPTIONAL MATERIAL

Mathematical preliminaries

binomial Coefficient

$$\binom{n}{k} = \begin{cases} \frac{n!}{k!(n-k)!}, & n \geq k \\ 0, & \text{else} \end{cases}$$

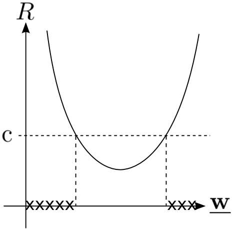
binomial distribution

$$f(k; n, p) = \binom{n}{k} p^k (1-p)^{n-k}$$

cumulative distribution

$$F(k; n, p) = Pr(X \leq k) = \sum_{x=0}^{k} f(x; n, p)$$

Strict Consistency

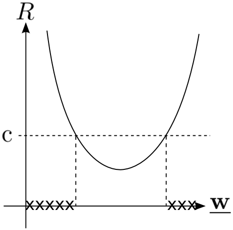


$\Lambda = \{\underline{\mathbf{w}}\}$: set of all predictors

$\Lambda_{(c)} \subset \Lambda$: set of all "bad" predictors

××× : $\Lambda_{(c)} = \{\underline{\mathbf{w}} : R_{(\underline{\mathbf{w}})} \geq c\}$

Strict Consistency



$\Lambda = \{\underline{\mathbf{w}}\}$: set of all predictors

$\Lambda_{(c)} \subset \Lambda$: set of all "bad" predictors

$\times\times\times$: $\Lambda_{(c)} = \{\underline{\mathbf{w}} : R_{(\underline{\mathbf{w}})} \geq c\}$

definition and implications of **strict consistency**

$$\lim_{p \to \infty} P\left\{\left| \inf_{\underline{\mathbf{w}} \in \Lambda_{(c)}} R^{(p)}_{\text{emp}(\underline{\mathbf{w}})} - \inf_{\underline{\mathbf{w}} \in \Lambda_{(c)}} R_{(\underline{\mathbf{w}})} \right| \geq \eta \right\} = 0$$

strict consistency $\overset{\rightarrow}{\not\leftarrow}$ inductive learning via ERM

proof: supplementary material

Strict Consistency $\rightarrow$ convergence in the limit

key theorem of learning theory

let $a \leq R_{(\underline{\mathbf{w}})} \leq A$ for all $\underline{\mathbf{w}} \in \Lambda$ and $P_{(\underline{\mathbf{z}})} \in \Pi$, then:

the ERM procedure is strictly consistent for Λ and Π

$$\Updownarrow$$

$$\lim_{p\rightarrow\infty} P\left\{ \sup_{\underline{\mathbf{w}}\in\Lambda} \left(R_{(\underline{\mathbf{w}})} - R^{(p)}_{\mathrm{emp}(\underline{\mathbf{w}})} \right) > \eta \right\} = 0\,, \qquad \text{for all} \qquad P_{(\underline{\mathbf{z}})} \in \Pi\,.$$

proof: supplementary material

- uniform (one-sided) convergence of $R^{(p)}_{\text{emp}(\underline{\mathbf{w}})}$ to $R_{(\underline{\mathbf{w}})}$ for $p \rightarrow \infty$