Statistical Learning Theory
An Introduction

Nothing is more practical than a good theory

V. Vapnik

By Seyed Hossein Ghafarian

Supervisor: Dr Sadoghi

Statistical Learning Theory

- Statistical Learning Theory (SLT)
  - Motivation: To provide a suitable learning algorithm to answer the following question:
    - **Obtain valid conclusions from experimental data.**

Statistical Learning Theory Background

- **Learning Objective: Generate a general rule that:**
  - **Be able to describe/explain the examples seen (so far).**
  - **Be able to generalize to unseen examples.**

Statistical Learning Theory

- Goal : **to learn a function** $f : X \rightarrow Y$
- In order to achieve this , we need a measure of *"**how good**" a function* , $f(x,\alpha)$, $\alpha \in \Lambda$ is.
- **Risk Minimization**
  - **Loss** between the response y of the supervisor and the response $f(x,\alpha)$ by learning machine : $L(y, f(x, \alpha))$
  - **Expected value of loss**

$$R(\alpha) = E[L(y, f(x,\alpha))] = \int L(y, f(x,\alpha)) dP(x, y)$$

  - **Find** $f(x,\alpha_0)$ ***which minimize*** **the risk functional** $R(\alpha)$
    - *Over the class of functions,* $f(x,\alpha)$, $\alpha \in \Lambda$
    - *Where the joint probability distribution* $P(x,y)$ *is unknown*
    - The only available information is the training set

$$(x_1, y_1), \ldots, (x_l, y_l)$$

Statistical Learning Theory

- **The Problem of Pattern Recognition**

  Let the supervisor's output *y* take on only two values *y={0,1}* and let *f(x,α) , α ϵ Λ* be a set of *indicator functions* .

$$L(y, f(x,\alpha)) = \begin{cases} 0 & \text{if } y = f(x,\alpha) \\ 1 & \text{if } y \neq f(x,\alpha). \end{cases}$$

  For this loss function *R(α)* provides probability of error.

- Similarly for **Regression Estimation**, and **Density estimation, …**

- **The General Setting of The Learning Problem**

  Let the probability measure *P(z)* be defined on the space Z. Consider the set of functions *Q(z,α)* , $\alpha \in \Lambda$. The goal is: to minimize the risk functional

$$R(\alpha) = \int Q(z,\alpha)\, dP(z), \qquad \alpha \in \Lambda$$

  If the probability measure ***P(z)*** is unknown but and i.i.d sample $\boldsymbol{z_1}$, … , $\boldsymbol{z_l}$ *is given*

Statistical Learning Theory

- *What kind of function f to consider?*

  At first , all possible functions ,

$$f(x, \alpha) , \alpha \in \Lambda_{all}=\{\alpha \,|f(., \alpha): X \rightarrow Y \}$$

- *At the time of training, it* ***is impossible*** **to compute minimum possible risk** *,also* ***risk of f cannot be computed*** **without knowledge of** P**.**

Statistical Learning Theory

- What we can do is ***to count the mistakes***.
- Empirical Risk Minimization(ERM) Induction Principle
  - **Expected risk functional $R(\alpha)$** *is replaced by the* ***empirical risk functional***

$$R_{\mathrm{emp}}(\alpha) = \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z, \alpha)$$

On the basis of training set z_1 , … , z_l

  - **Approximate** the function $Q(x,\alpha_0)$ **which minimizes risk** $R(\alpha)$ by the function $Q(x,\alpha_l)$ **which minimizes** $R_{emp}(\alpha)$

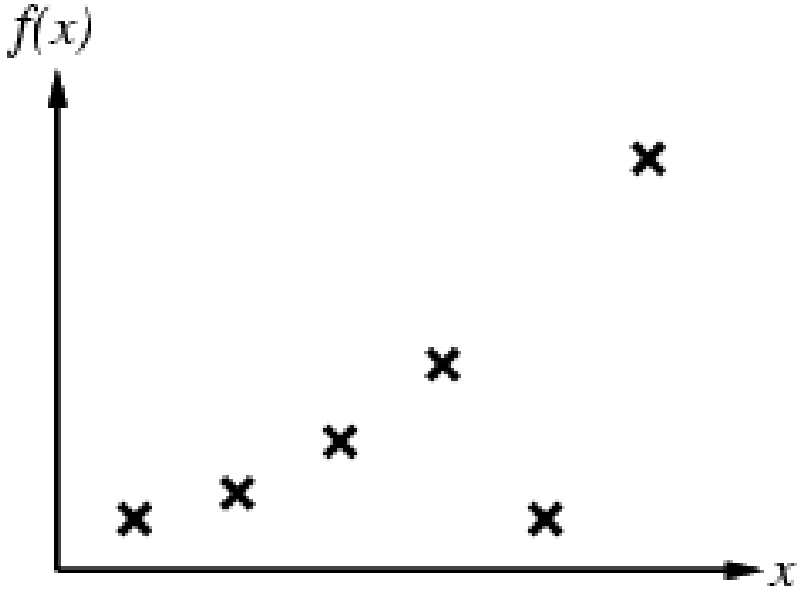
Generalization And Consistency

- Does a function ***f*** which makes ***few errors on the training set*** also makes ***few errors on the rest of the space*** X , that is whether it has a small overall risk $R(\alpha)$?
- ***Generality:*** ***A classifier*** f_n ***is Generalizes well if the difference*** $|R(\alpha_n)-R_{emp}(\alpha_n)|$ ***is small .***
- ***Consistency : an algorithm, when presented more and more training samples ,eventually converge to an optimal solution.***

Definition 11 (Consistency). *An algorithm is consistent if for any probability measure* P,

$$\lim_{n\to\infty} R(g_n) = R^* \ \textit{almost surely.}$$

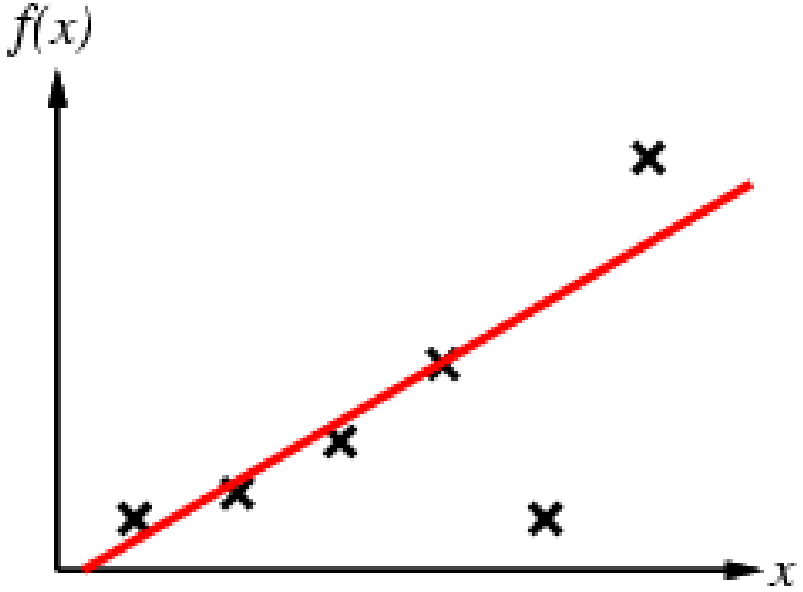
Bias-Variance Dilemma



- Given empirical observations, $(x_1, y_1), \ldots, (x_l, y_l)$, nearly linear

R(a) ? The problem is that we cannot compute this risk form training data.

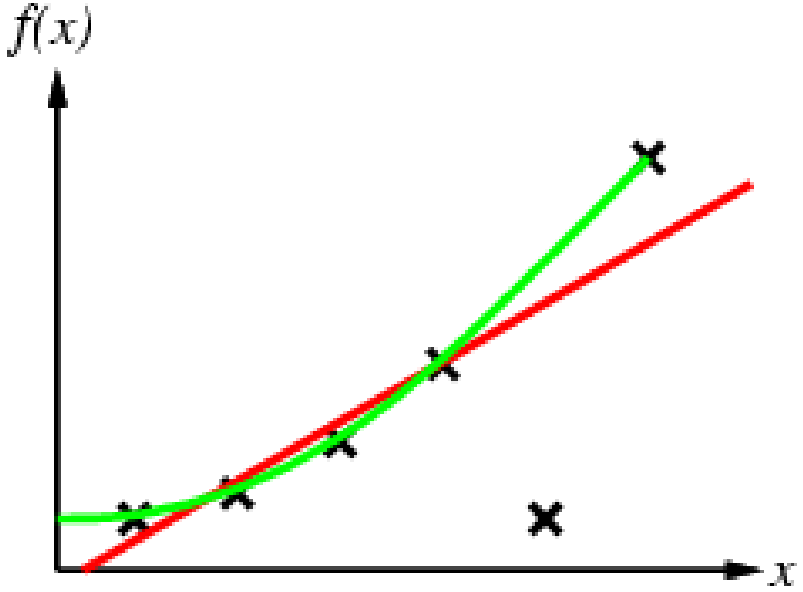
Bias-Variance Dilemma



- Given empirical observations, $(x_1, y_1), \ldots, (x_l, y_l)$, nearly linear
- Straight line ,does not explain the data. Blue line, fits training data perfectly, with zero training error.

R(a) ? The problem is that we cannot compute this risk form training data.

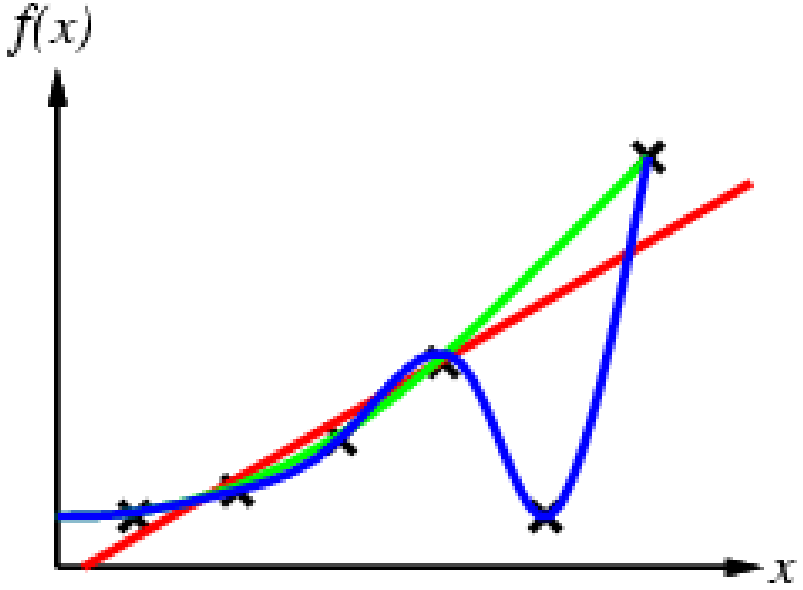
Bias-Variance Dilemma



- Given empirical observations,$(x_1 , y_1),\ldots , (x_l , y_l)$,nearly linear
- Straight line ,does not explain the data. Blue line, fits training data perfectly, with zero training error.

R(a) ? The problem is that we cannot compute this risk form training data.

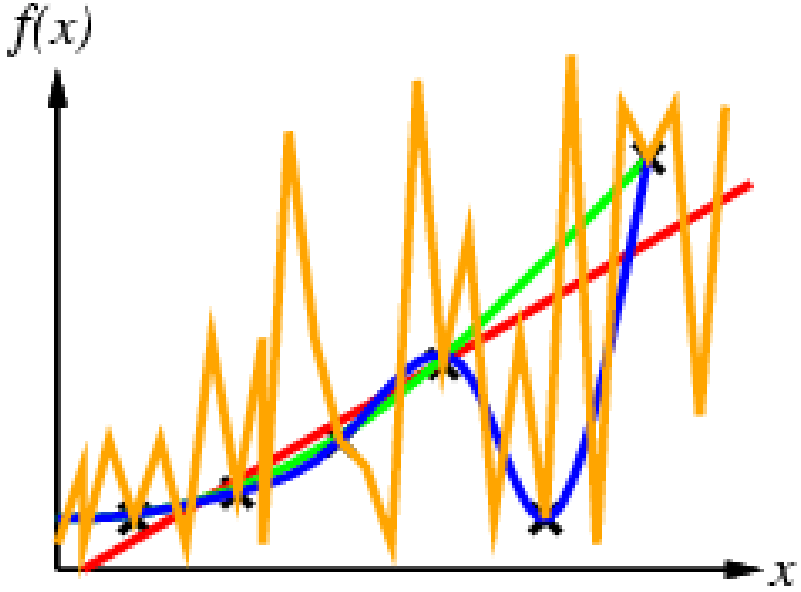
Bias-Variance Dilemma



- Given empirical observations, $(x_1, y_1), \ldots, (x_l, y_l)$, nearly linear
- Straight line ,does not explain the data. Blue line, fits training data perfectly, with zero training error.

R(α) ? The problem is that we cannot compute this risk form training data.

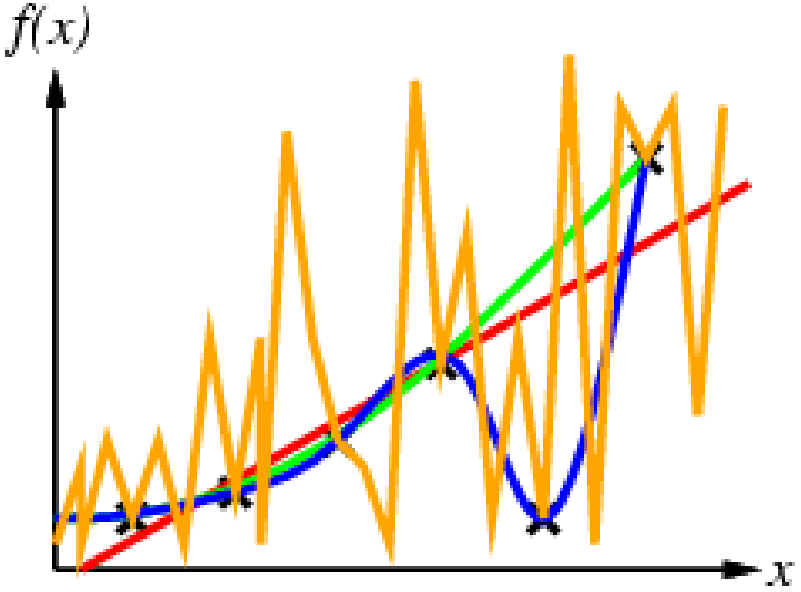
Bias-Variance Dilemma



- Given empirical observations, $(x_1, y_1), \ldots, (x_l, y_l)$, nearly linear
- Straight line ,does not explain the data. Blue line, fits training data perfectly, with zero training error.
- What is the true risk $R(\alpha)$? ***The problem is that we cannot compute this risk from training data.***

Bias-Variance Dilemma

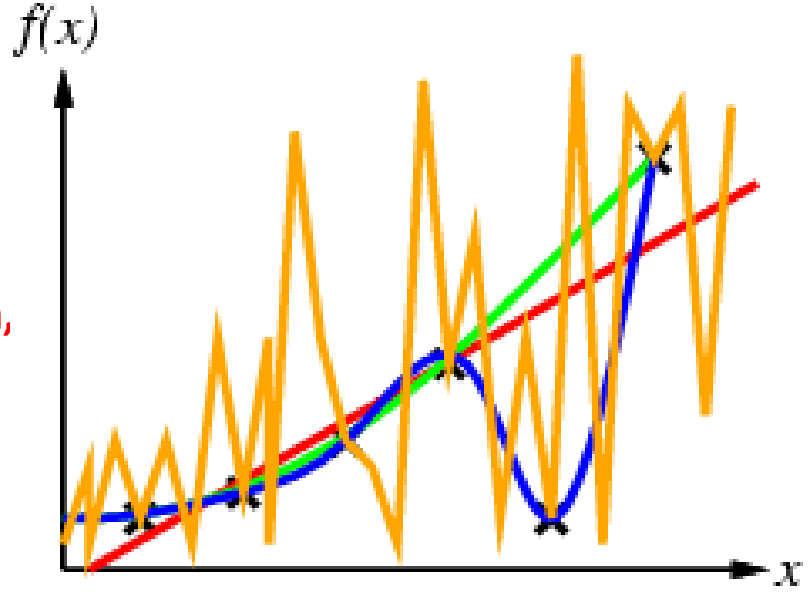
- Bias-Variance Dilemma



  - What is the degree of fitness?
  - **A complex function with low error or a simple function with high error?**
  - Physicists accept the straight line with a large error, and of course they attribute the error to the measurement rather than to the model.
  - The question is, how much learning error should be tolerated by presenting a simple model?

Bias-Variance Dilemma

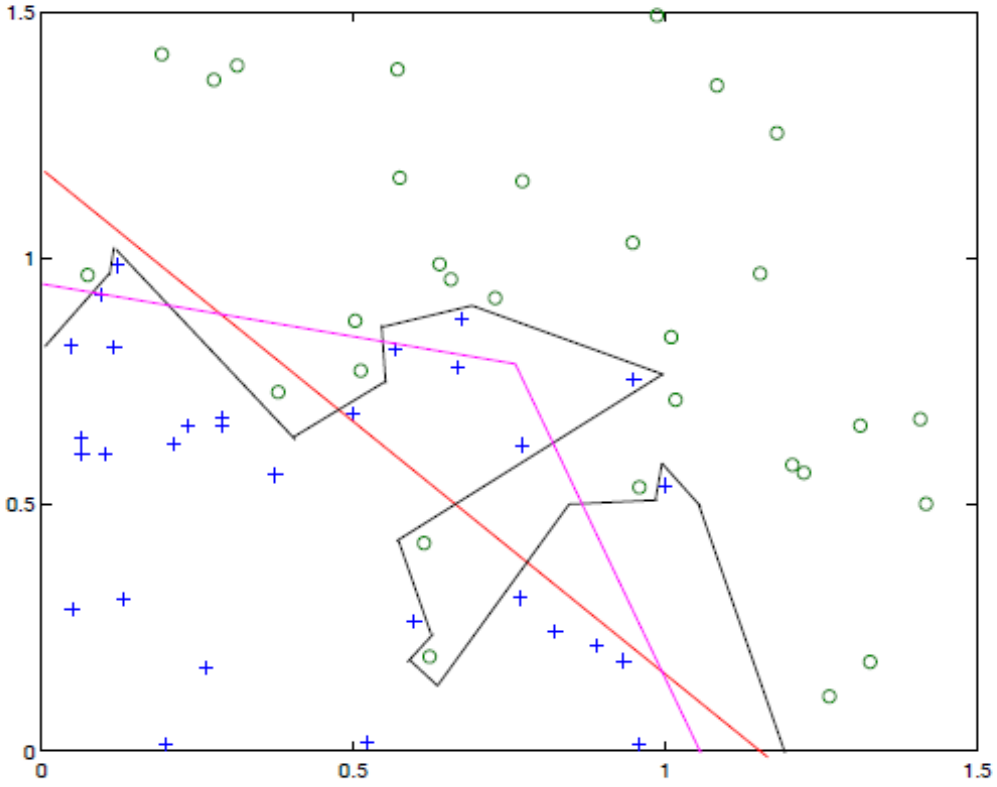
- Bias-Variance Dilemma,
  - **Approximation, Estimation error**
  - **Overfitting, Underfitting**



- If we fit a line, we have imposed our view of simplicity on the model: high **bias**
- If we fit a high-order polynomial, we have tracked all the changes in the data and therefore the model has a lot of variance. (i.e. it tracks a lot of changes):high **variance**
- This problem is known as ***Bias-Variance dilemma*.**

Bias-Variance Dilemma

- Bias-Variance Dilemma,
  - **Approximation, Estimation error**
  - **Overfitting, Underfitting**



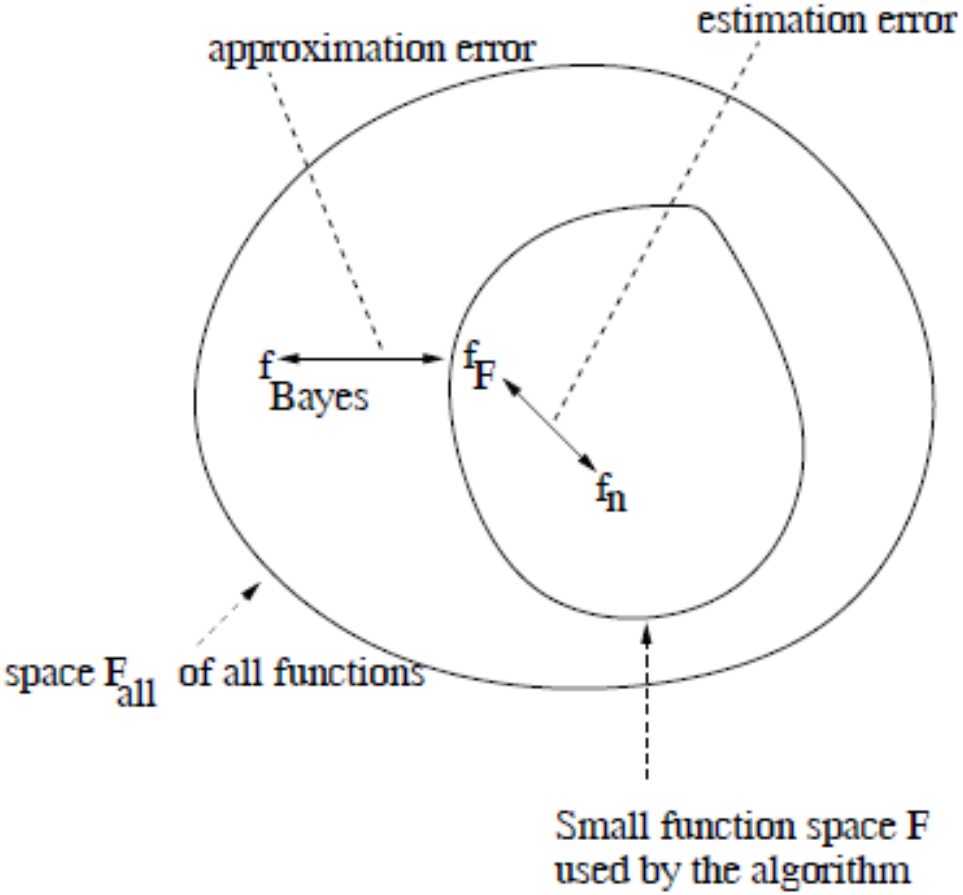
Bias-Variance Dilemma

- Could we just choose Λ as the space Λ_{all} of ***all functions*** $f(x, \alpha)$, $\alpha \in \Lambda_{all}=\{\alpha \mid f(., \alpha): X \rightarrow Y \}$, define the classifier

  $f(x, \alpha_n)$, $\alpha_n = argmin_{\Lambda all} R_{emp}(\alpha)$ and obtain consistency? **NO**
- **In other words, consistency is not achieved by minimizing empirical risk on all functions.**
- **It is proven that if** Λ_{all} ***is so large*** *that it contains*

  Optimal Classifier for all different distributions P ,
- *this will lead* ***to inconsistency****. So if we want to learn successfully , we need* **to work with a smaller function class Λ .**

If the set of alpha parameters is chosen large enough, of course, considering the possible distributions for the learning data in such a way that the optimal classifier for all possible data distributions is included in it, the result will lead to inconsistency. According to the No Free Lunch Theorem, there is no classifier that is responsive (optimal) for all possible distributions, and this statement now means that make the range of alphas large enough to be responsive. In a sense, we are making the system memory-intensive, so it is inconsistent.



$$R(f_n) - R(f_{Bayes}) = \underbrace{\Big(R(f_n) - R(f_{\mathcal{F}})\Big)}_{\text{estimation error}} + \underbrace{\Big(R(f_{\mathcal{F}}) - R(f_{Bayes})\Big)}_{\text{approximation error}}$$

Theory of Consistency of LP

- Is empirical risk minimization is always consistent? No, Consider

$$Y = \begin{cases} -1 & \text{if } X < 0.5 \\ 1 & \text{if } X \geq 0.5. \end{cases}$$

$$f_n(X) = \begin{cases} Y_i & \text{if } X = X_i \text{ for some } i = 1, \ldots, n \\ 1 & \text{otherwise.} \end{cases}$$

- $R_{emp}(f_n) = 0$
- f_n does not learn anything, the classifier **just memorize** the training labels and otherwise simply predicts the label 1.
- The selection function minimizes the empirical risk only by being memory-based and has a constant value at all other points.

Theory of Consistency of LP

- Why empirical risk minimization can be inconsistent?

$$Y = \begin{cases} -1 & \text{if } X < 0.5 \\ 1 & \text{if } X \geq 0.5. \end{cases}$$

$$f_n(X) = \begin{cases} Y_i & \text{if } X = X_i \text{ for some } i = 1, \ldots, n \\ 1 & \text{otherwise.} \end{cases}$$

– **The numerical values of fn at points x1, …, xl do not contain information about the values of the function at other points, so they do not relate the empirical risk to the true risk. So they are not consistent.**

Theory of Consistency of LP

- **Why empirical risk minimization can be inconsistent?**

$$Y = \begin{cases} -1 & \text{if } X < 0.5 \\ 1 & \text{if } X \geq 0.5. \end{cases}$$

$$f_n(X) = \begin{cases} Y_i & \text{if } X = X_i \text{ for some } i = 1, \ldots, n \\ 1 & \text{otherwise.} \end{cases}$$

- **If restrictions are imposed on the space of possible functions, one can hope for generalizability.**
- **In SLT, a concept called function space capacity is introduced to create constraints, or in essence, the size of the version space is considered.**

Theory of Consistency of LP

- The theory of consistency is an **asymptotic** theory.
- *Necessary and sufficient conditions for convergence* of the solutions to the best possible as the number of observations is increased.
- **Any analysis of the convergence properties** of the **ERM** must be a ***worst case analysis***.

If little data is available, the empirical risk is very different from the actual risk. It should also be performed on all possible functions and in the worst case (sup).

The Key Theorem: Let $Q(z,\alpha), \alpha \in \Lambda$ be a set of functions that has a bounded loss for probability measure $P(z)$

$$A \leq \int Q(z,\alpha) dP(z) \leq B \qquad \forall \alpha \in \Lambda.$$

Then for the ERM principle to be consistent it is necessary and sufficient that the empirical risk $R_{\text{emp}}(\alpha)$ converge *uniformly* to the actual risk $R(\alpha)$ over the set $Q(z,\alpha), \alpha \in \Lambda$ as follows:

$$\lim_{\ell \to \infty} \text{Prob}\left\{ \sup_{\alpha \in \Lambda} (R(\alpha) - R_{\text{emp}}(\alpha)) > \varepsilon \right\} = 0, \quad \forall \varepsilon. \quad (11)$$

Theory of Consistency of LP

- **Necessary condition for consistency** (not only the *sufficient condition*) depends on whether or not the deviation for the ***Worst function*** over **the given set of functions**

$$\Delta(\alpha_{\text{worst}}) = \sup_{\alpha \in \Lambda} (R(\alpha) - R_{\text{emp}}(\alpha))$$

- **Converge** in probability **to zero**.

The Key Theorem: Let $Q(z, \alpha), \alpha \in \Lambda$ be a set of functions that has a bounded loss for probability measure $P(z)$

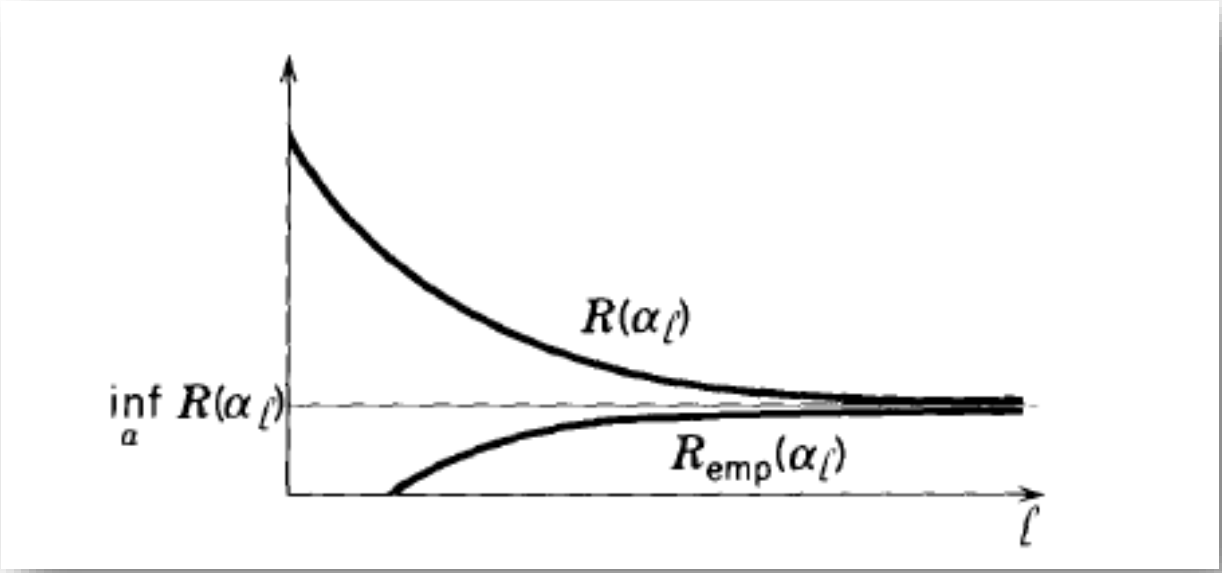
$$A \leq \int Q(z, \alpha) dP(z) \leq B \qquad \forall \alpha \in \Lambda.$$

Then for the ERM principle to be consistent it is necessary and sufficient that the empirical risk $R_{\text{emp}}(\alpha)$ converge *uniformly* to the actual risk $R(\alpha)$ over the set $Q(z, \alpha), \alpha \in \Lambda$ as follows:

$$\lim_{\ell \to \infty} \text{Prob}\left\{ \sup_{\alpha \in \Lambda} (R(\alpha) - R_{\text{emp}}(\alpha)) > \varepsilon \right\} = 0, \quad \forall \varepsilon. \quad (11)$$

Theory of Consistency of LP

- This type of convergence is called ***uniform convergence***.
- This theorem considers number of samples in infinity.
- But what when number of samples is finite and number of different functions is very large or infinite.



Theory of Consistency of LP

- ***Uniform Convergence*** *in the* ***simplest case****:*
  - let Q(x, α) , α ϵ Λ be a set of indicator functions.
  - What is the necessary and sufficient conditions for uniform two-sided convergence?

$$P\left\{\sup_{\alpha\in\Lambda}\left|\int Q(z,\alpha)\,dF(z)-\frac{1}{\ell}\sum_{i=1}^{\ell}Q(z_i,\alpha)\right|>\varepsilon\right\}\xrightarrow[\ell\to\infty]{}0$$

Theory of Consistency of LP

- ***Uniform Convergence*** *in the* ***simplest case****:*
- Rewrite as

$$P\left\{\sup_{\alpha\in\Lambda}|P\{Q(z,\alpha)>0\}-\nu_\ell\{Q(z,\alpha)>0\}|>\varepsilon\right\}\underset{\ell\to\infty}{\longrightarrow}0, \qquad (3.18)$$

where $P\{Q(z,\alpha\}>0\}$ are probabilities of the events $A_\alpha=\{z:Q(z,\alpha)>0\}$, $\alpha\in\Lambda$, and $\nu_\ell\{Q(z,\alpha)>0\}$ are frequencies of these events obtained on the given data $z_1,\ldots,z_\ell$.

The probability of a loss opposite to zero tends towards the most frequent loss value. The meaning behind this statement is that the statistical average tends towards the numerical average, even when there is an infinite number of data.

Theory of Consistency of LP

- ***Uniform Convergence*** *in the* ***simplest case:***

The convergence rate using Chernoff inequality shows that the upper bound of the difference between empirical and statistical risk is different instead of zero in different volumes of learning data, which in infinite learning data is zero as the upper bound.

According to the Bernoulli theorem for any fixed event $A^* = \{z : Q(z, \alpha^*) > 0\}$, the frequencies converge to the probability when the number of observations tends to infinity. The inequality

$$P\{|P\{Q(z, \alpha^*) > 0\} - \nu_\ell\{Q(z_i, \alpha^*) > 0\}| > \varepsilon\} \leq 2\exp\{-2\varepsilon^2\ell\} \qquad (3.19)$$

(Chernoff inequality) describes the rate of convergence.

Theory of Consistency of LP

- ***Uniform Convergence*** *in the* ***simplest case:***

The Simplest Model. Let our set of events contain a finite number N of events $A_k = \{z : Q(z, \alpha_k) > 0\}$, $k = 1, 2, .., N$. For this set of events, uniform convergence does hold. Indeed, the following sequence of inequalities is valid:

$$P\{\max_{1 \le k \le N} |P\{Q(z, \alpha_k) > 0\} - \nu_\ell\{Q(z_i, \alpha_k) > 0\}| > \varepsilon\}$$

$$\le \sum_{k=1}^{N} P\{|P\{Q(z, \alpha_k) > 0\} - \nu_\ell\{Q(z_i, \alpha_k) > 0\}| > \varepsilon\}$$

$$\le 2N \exp\{-2\varepsilon^2 \ell\} \qquad (3.20)$$

$$= 2 \exp\left\{\left(\frac{\ln N}{\ell} - 2\varepsilon^2\right)\ell\right\}. \qquad (3.21)$$

- *This inequality suggest that in order to obtain uniform convergence for any ε, the expression*

$$\frac{\ln N}{\ell} \xrightarrow[\ell \to \infty]{} 0$$

number of models N
number of learning samples L

has to be true.

Theory of Consistency of LP

Necessary and sufficient conditions for *uniform convergence*

- ***Consider a set of functions*** $f(x, \alpha)$, $\alpha \in \Lambda = \{\alpha \mid f(., \alpha): X \rightarrow Y\}$
- ***Even if the set of functions is infinite,***

The number of different output of these functions ***for the sample*** $(z_1, \ldots, z_l)$

is finite .

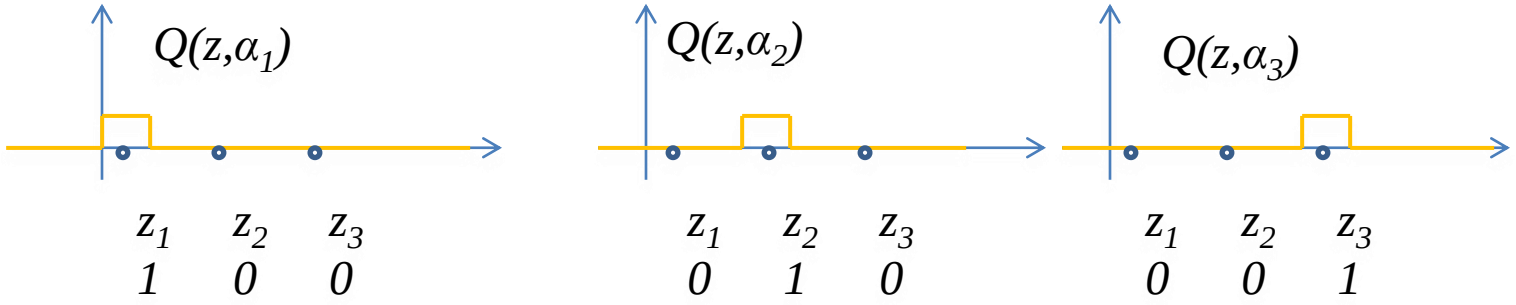
- ***In fact it is at most*** 2^l **Number of outputs on l examples**
- ***Define:***
- $N^{\Lambda}(z_1, \ldots, z_l)$ **: Diversity of the set of functions** $Q(z,\alpha)$ on the given sample that represents ***the number of different separations of this sample*** that can be obtained using functions from the given set of indicator functions .

Number of different shattering of learning sample $N^{\Lambda}(z_1, \ldots, z_l)$ ***with separating function Q with Alpha parameter***

Theory of Consistency of LP

Necessary and sufficient conditions for *uniform convergence*

- ***Entropy of the set of functions:***
- $N^{\Lambda}(z_1, \ldots, z_l)$ **: Diversity of this set of functions** $Q(z,\alpha)$ on the given sample that represents ***the number of different separations of this sample*** that can be obtained using functions from the given set of indicator functions .
- for example : consider one dimensional Indicator functions



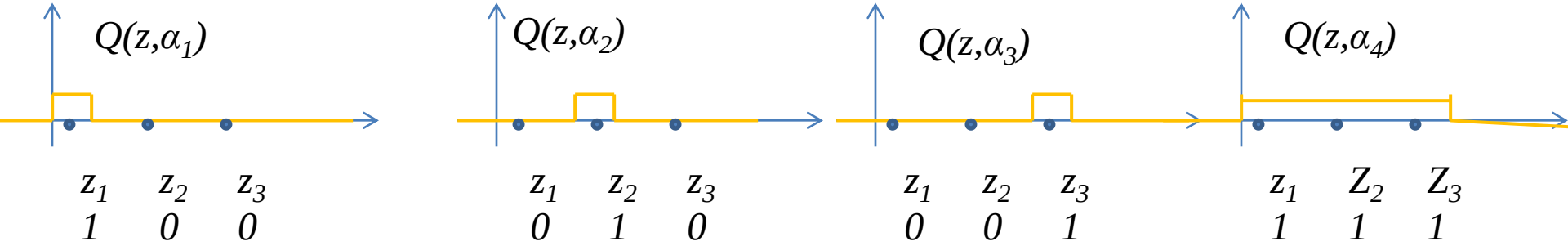
- It is not possible to have an output (0,1 ,1) of functions $Q(z,\alpha_1)$, $Q(z,\alpha_2)$, $Q(z,\alpha_3)$ on the above defined (z_1,z_2,z_3)
- So from the 2^3 possible output on the samples (z_1,z_2,z_3) , we can have only 3 of them.

Theory of Consistency of LP

Necessary and sufficient conditions for *uniform convergence*

- ***Entropy of the set of functions:***
- $N^{\Lambda}(z_1, \ldots, z_l)$ **: Diversity of this set of functions** $Q(z,\alpha)$ on the given sample that represents ***the number of different separations of this sample*** that can be obtained using functions from the given set of indicator functions .

- if we add a function $Q(z,\alpha_4)$ to the set of functions

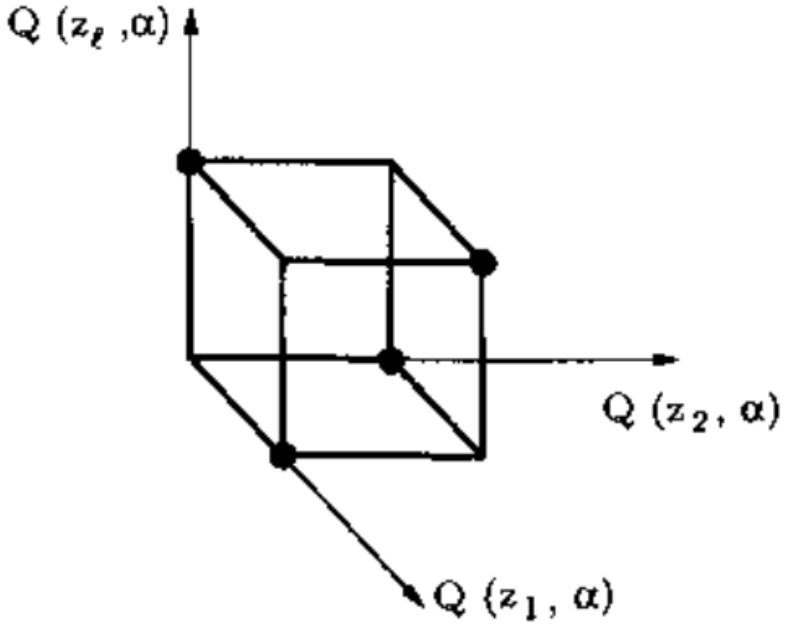


- Then we can have (1,1,1) but not (1,0,1)
- So from the 2^3 possible output on the samples (z_1,z_2,z_3) , we can have only 4 of them.

Theory of Consistency of LP

Necessary and sufficient conditions for *uniform convergence*

- ***Entropy of the set of functions:***
- *In another way , consider the set of l dimensional vectors,*
  - $q(\alpha)=(Q(z_1,\alpha), \dots, Q(z_l,\alpha))$, $\alpha \in \Lambda$ *that can be obtained when α takes various values from* Λ. $N^{\Lambda}(z_1, \dots, z_l)$ *is the number of different vertices of the l-dimensional cube that can be obtained on the basis of sample* $(z_1, \dots, z_l)$ *and the set of functions* $Q(x,\alpha)$, $\alpha \in \Lambda$.



Theory of Consistency of LP

Necessary and sufficient conditions for *uniform convergence*

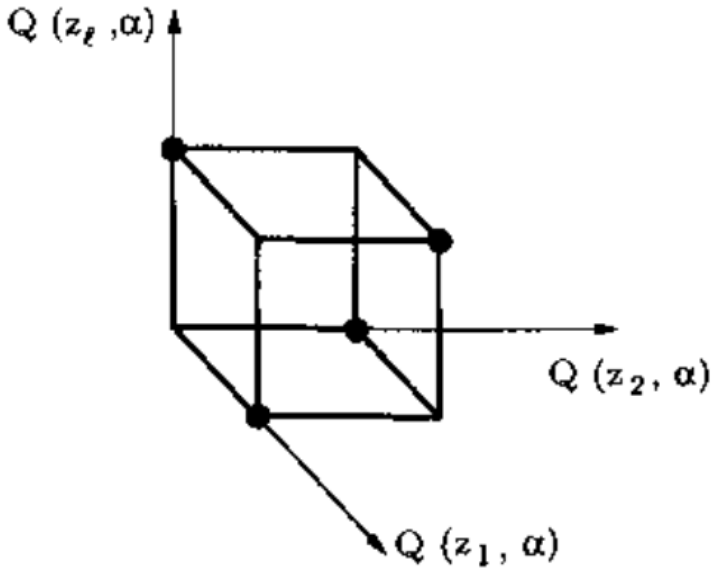
- ***Entropy of the set of functions*** ; ***Random entropy:***
  - *Describe the* ***diversity of the set of functions*** *on the given data.*

$$H^{\Lambda}(z_1, \ldots, z_l) = \ln N^{\Lambda}(z_1, \ldots, z_l)$$

  - *It's a random variable, since it was constructed using random i.i.d data*
  - *Expectation of random entropy over the joint distribution function* $P(z_1, \ldots, z_l)$

$$H^{\Lambda}(l) = E[\ln N^{\Lambda}(z_1, \ldots, z_l)]$$

Statistical average over all states and values on the learning data



Theory of Consistency of LP

Necessary and sufficient conditions for *uniform convergence*

Theorem 3.3. *In order that uniform convergence*

$$P\left\{\sup_{\alpha\in\Lambda}\left|\int Q(z,\alpha)\,dF(z)-\frac{1}{\ell}\sum_{i=1}^{\ell}Q(z_i,\alpha)\right|>\varepsilon\right\}\xrightarrow[\ell\to\infty]{}0$$

over the set of indicator functions $Q(z,\alpha),\alpha\in\Lambda$ *be valid it is necessary and sufficient that the condition*

$$\frac{H^{\Lambda}(\ell)}{\ell}\xrightarrow[\ell\to\infty]{}0 \qquad (3.24)$$

be satisfied.

- Difference between this equation and $\frac{\ln N}{\ell}\xrightarrow[\ell\to\infty]{}0$

is only in ***characterizing the capacity*** of the set of functions.

Nonfalsifiability

- Since the era of ancient philosophy, two models of reasoning have been accepted:
  - Deductive, which means moving from general to particular using a system of axioms and inference rules.
  - Inductive, which means moving from particular to general, consists of the formation of general judgments from particular assertions.
  - However, general judgments from true particular assertions are **not always true**. Nevertheless, it is assumed that there exist such cases of inductive inference for which generalization assertions are justified

Nonfalsifiability

- Inductive inference
  - Inductive, which means moving from particular to general, consists of the formation of general judgments from particular assertions.
  - However, general judgments from true particular assertions are **not always true**. Nevertheless, it is assumed that there exist such cases of inductive inference for which generalization assertions are justified
  - In other words, can the observation that: All swans are white: which has been true for all our observations so far, be the basis for the argument that: All swans are white?

Nonfalsifiability

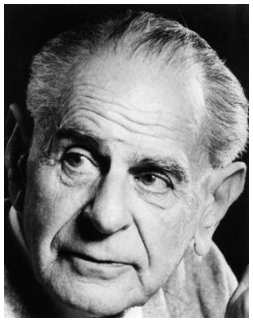
- Since the era of ancient philosophy, two models of reasoning have been accepted:
  - In other words, can the observation that: All swans are white: which has been true for all our observations so far, be the basis for the argument that: All swans are white? این

    The proposition is a falsifiable proposition, meaning that it is falsified by the observation of a black swan case.
  - However, there are cases where it is valid to generalize from specific cases (inductive inference).
  - **What is the difference between a valid inductive inference and an invalid inductive inference? This has been a very important issue in the history of philosophy, a problem called the Demarcation Problem. The distinction between valid and invalid inductive inference.**

Nonfalsifiability

- The **demarcation problem**, originally proposed by **kant**, is a central question of inductive theory.
  - ❖ **What is the difference between the cases with a justified inductive step and those for which the inductive step is not justified?**
- In history of science, there have been both true theories that reflect reality (say **chemistry**) and false ones (say **alchemy**) that do not.
  - ➢ **Is there a formal way to distinguish between true and false theories?**
    - ➢ Complexity,? Predictive ability? Use of mathematics? Level of formality?
    - ➢ None of the above gives clear advantage to either of theories.

Nonfalsifiability



- In 1930, **Sir Karl R. Popper**, the great philosopher, suggested his famous criterion for **demarcation problem:**

  - **A necessary condition for justifiability of a theory is the feasibility of its falsification.**
  - By falsification of a theory, popper means the existence of a collection of particular assertions which cannot be explained by the given theory although they fall into its domain. If the given theory can be falsified, it satisfies the necessary condition of a justifiable theory.

- Example:
  - In the New York area, both a tropical storm and snowfall can happen in one hour.
  - According to meteorology it is impossible. But according to astrology, it can be explained using elements in starry sky.

Theorems about Nonfalsifiability

- Vapnik shows that if uniform two-sided convergence does not take place, then the method of empirical risk minimization is nonfalsifiable.
- Case of Complete Nonfalsifiability
  - Suppose for the set of indicator functions $Q(x, \alpha)$, $\alpha \in \Lambda$, the following equality is true:

$$\lim_{\ell \to \infty} \frac{H^{\Lambda}(\ell)}{\ell} = \ln 2.$$

  - Intuitively, it is clear that the ratio of the entropy to the number of observations monotonically decreases when the number of observations increases.(it's also a theorem).Thus if this equation happened, then for any finite number *l*, *we have* the following equality holds true.

$$\frac{H^{\Lambda}(\ell)}{\ell} = \ln 2$$

Theorems about Nonfalsifiability

- Case of Complete Nonfalsifiability, Continued
  - According to definition of entropy, this means that for almost all samples $z_1, \ldots, z_l$, the following equality is valid:

$$N^{\Lambda}(z_1, \ldots, z_\ell) = 2^\ell$$

  - In other words, the set of functions of the learning machine is such that almost any sample can be separated in all possible ways by functions of this set.
  - This implies that the minimum of empirical risk for this machine equals zero.

Theorems about Nonfalsifiability

- Case of Complete Nonfalsifiability, Continued
  - We call this learning machine **nonfalsifialble** because it can give a general explanation (function) for almost any data.

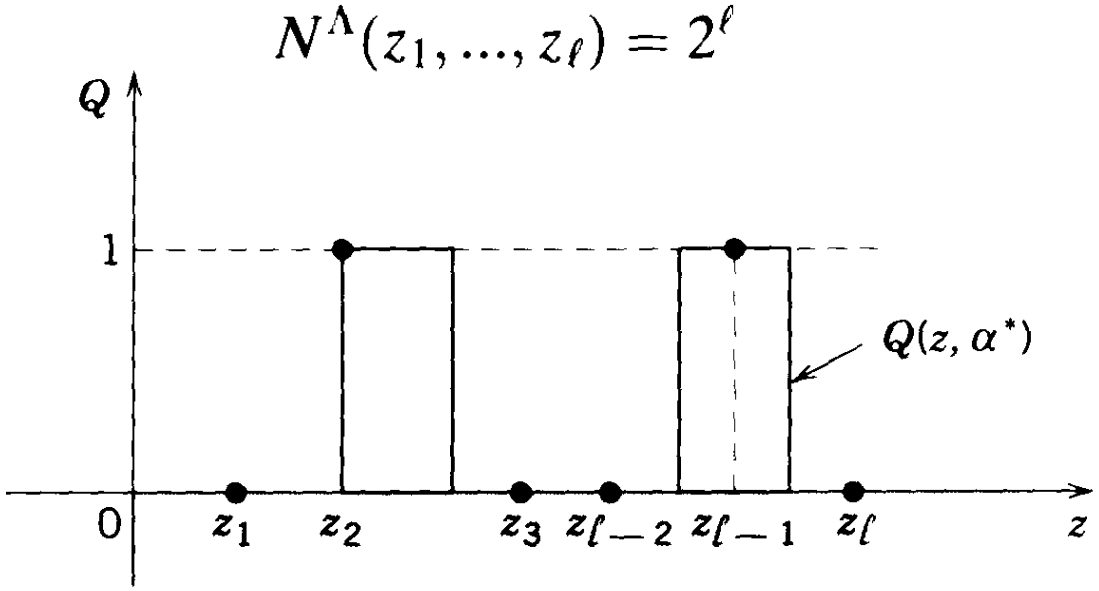


FIGURE 3.6. A learning machine with the set of functions $Q(z, \alpha)$, $\alpha \in \Lambda$, is *nonfalsifiable* if for almost all samples $z_1, \ldots, z_\ell$ given by the generator of examples and for any possible labels $\delta_1, \ldots, \delta_\ell$ for these z's, the machine contains a function $Q(z, \alpha^*)$ that provides equalities $\delta_i = Q(x_i, \alpha)$, $i = 1, \ldots, \ell$.

Theorems about Nonfalsifiability

- Theorem About Partial Nonfalsifiability

Theorem 3.6. *For the set of indicator functions* $Q(z, \alpha), \alpha \in \Lambda$, *let the convergence*

$$\lim_{\ell \to \infty} \frac{H^{\Lambda}(\ell)}{\ell} = c > 0$$

be valid. Then there exists a subset Z^* *of the set* Z *such that*

(a) $$P(Z^*) = c$$

and (b) for the subset $z_1^*, \ldots, z_k^* = (z_1, \ldots, z_\ell) \cap Z^*$

of almost any training set $z_1, \ldots, z_\ell$

that belongs to Z^* *and for any given sequence of binary values*

$$\delta_1, \ldots, \delta_k, \qquad \delta_i \in \{0, 1\}$$

there exists a function $Q(z, \alpha^*)$ *for which the equalities*

$$\delta_i = Q(z_i^*, \alpha^*), \qquad i = 1, 2, \ldots, k$$

hold true.

Theorems about Nonfalsifiability

- Theorem About Partial Nonfalsifiability
- This theorem shows that if conditions of uniform covergence fail, then there exists some subspace where the learning machine is nonfalsifiable.

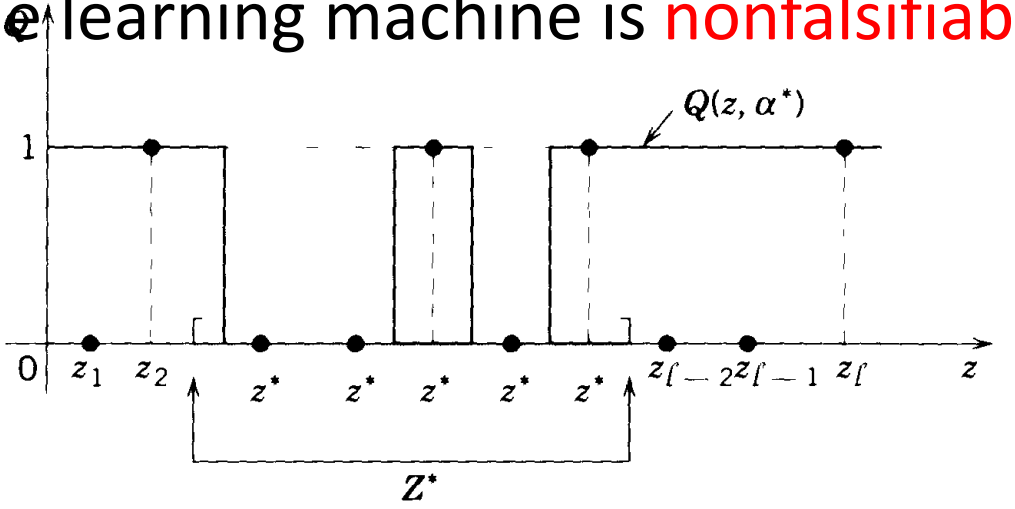


FIGURE 3.7. A learning machine with the set of functions $Q(z, \alpha)$, $\alpha \in \Lambda$, is *partially nonfalsifiable* if there exists a region $Z^* \in Z$ with nonzero measure such that for almost all samples $z_1, \ldots, z_\ell$ given by the generator of examples and for any labels $\delta_1, \ldots, \delta_\ell$ for these z's, the machine contains a function $Q(z, \alpha^*)$ that provides equalities $\delta_i = Q(z_i, \alpha)$ for all z_i belonging to the region Z^*.

Theory of Consistency of LP

Necessary and sufficient conditions for *uniform convergence*

- Three milestones in Learning Theory
  - Entropy for sets of indicator functions

$$H^{\Lambda}(l)=E[\ln N^{\Lambda}(z_1, \ldots, z_l)]$$

  - *The annealed VC-entropy:*

$$H_{ann}^{\Lambda}(l)=\ln E[N^{\Lambda}(z_1, \ldots, z_l)]$$

  - *Growth function*

$$G^{\Lambda}(l)=\ln \sup_{z1, \ldots, zl} N^{\Lambda}(z_1, \ldots, z_l)$$

  - $H^{\Lambda}(l) \le H_{ann}^{\Lambda}(l) \le G^{\Lambda}(l)$

Theory of Consistency of LP

The Three Milestone in learning theory

- *The first Milesotne:*

$$\lim_{\ell \to \infty} \frac{H^{\Lambda}(\ell)}{\ell} = 0$$

  - *Describes the **necessary and sufficient condition** for **consistency of the ERM principle**.*
  - *Any machine minimizing empirical risk should satisfy it.*

Theory of Consistency of LP

The Three Milestone in learning theory

- The second milestone :

$$\lim_{\ell \to \infty} \frac{H^{\Lambda}_{\text{ann}}(\ell)}{\ell} = 0$$

  - Describes the sufficient condition for fast convergence.
  - Recently it has been proved that it's also is necessary . In 1998 it's necessity was an open question
  - Fast convergence:
    - ***Asymptotic rate of convergence is fast*** *if for any* $l > l_0$ *and some constant* $c>0$ *. the followina exponential holds true*

$$P\{R(\alpha_\ell) - R(\alpha_0) > \varepsilon\} < e^{-c\varepsilon^2 \ell}$$

Theory of Consistency of LP

The Three Milestone in learning theory

- The Third milestone:
  - Under **what conditions is the ERM principle consistent** and **rapidly converging** , ***independently of the probability measure****?*
  - *The necessary and sufficient conditions for consistency of ERM for any probability measure:*

$$\lim_{\ell \to \infty} \frac{G^{\Lambda}(\ell)}{\ell} = 0.$$

  - *It describes the conditions under which the learning machine implementing ERM Principle has an asymptotic high rate of convergence independently of the problem to be solved.*

Rate of Convergence of Learning Process

- To estimate **the quality of the ERM method for a given sample size**, it's necessary to obtain **non-asymptotic bounds** on the rate of uniform convergence.
- **The growth function** has a remarkable property that can be used for this purpose

Theorem: Any growth function either satisfies the equality

$$G^{\Lambda}(\ell) = \ell \ln 2$$

or is bounded by the inequality

$$G^{\Lambda}(\ell) < h\left(\ln \frac{\ell}{h} + 1\right)$$

where h is an integer for which

$$G^{\Lambda}(h) = h \ln 2$$
$$G^{\Lambda}(h+1) \neq (h+1) \ln 2.$$

In other words the growth function will be either a linear function or will be bounded by a logarithmic function. (For example, it cannot be of the form $G^{\Lambda}(\ell) = c\sqrt{\ell}$).

Rate of Convergence of Learning Process

- Growth function:

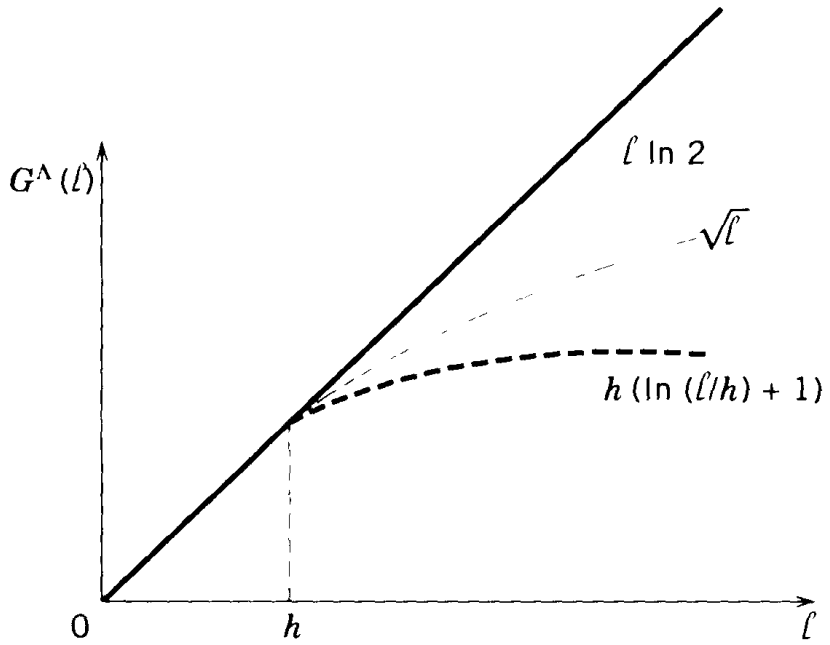
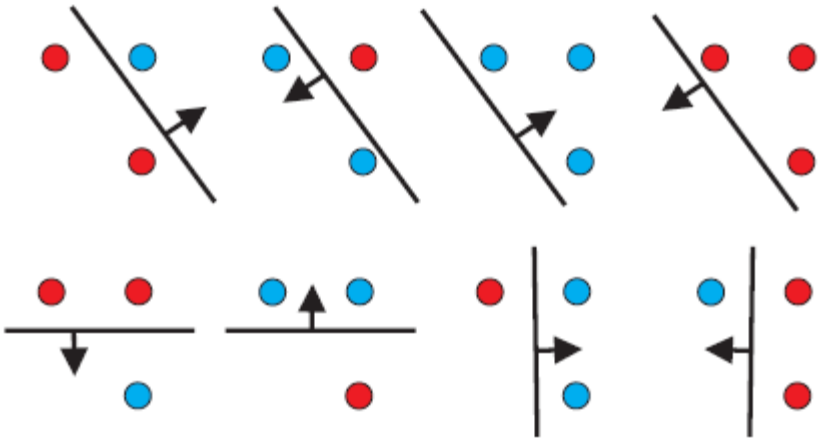


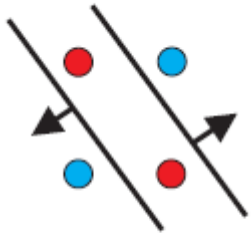
FIGURE 4.1. The growth function is either linear or bounded by a logarithmic function. It cannot, for example, behave like the dashed line.

Rate of Convergence of Learning Process

- h is called the **VC-dimension** of Learning Machine



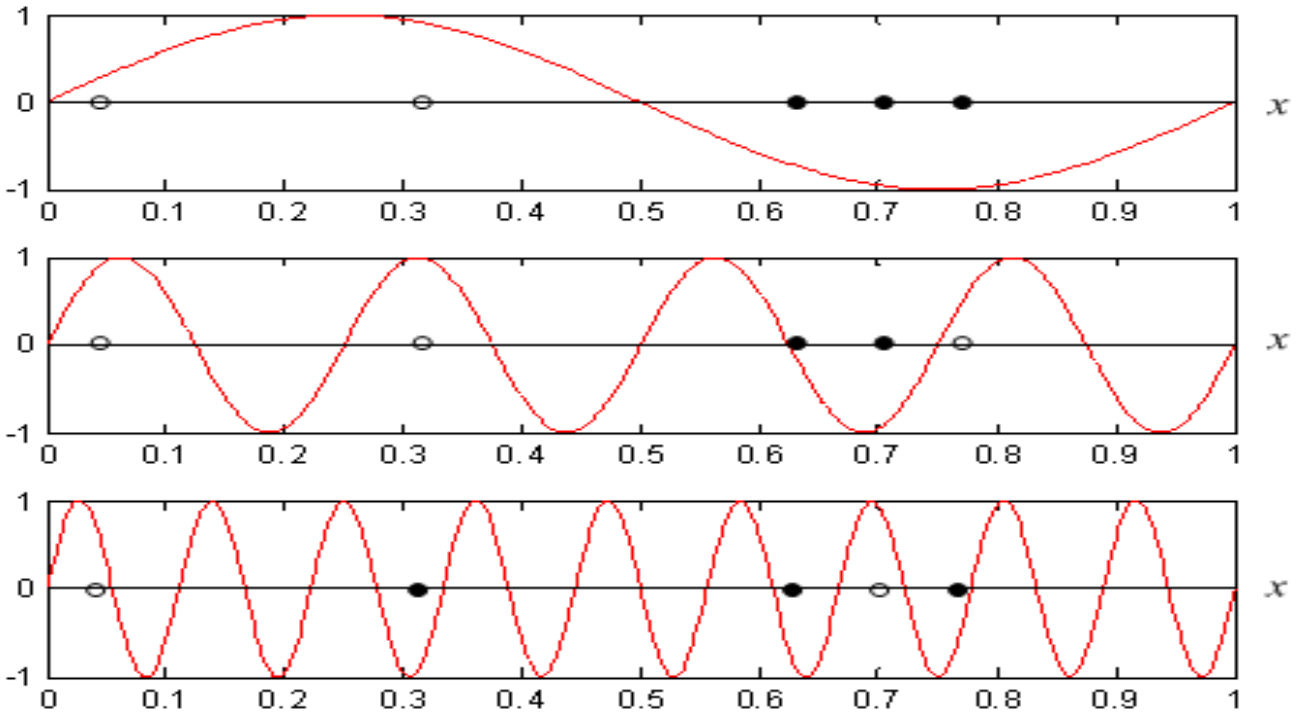
3 points, shattered



4 points, undivisible

Rate of Convergence of Learning Process

- A set of indicator functions $y = I(\sin(\omega x))$ has infinite VC dimension



Rate of Convergence of Learning Process

VC-dimension

- **VC-dimension** of the set of indicator functions is **infinite** if the **Growth function** for this set of functions is **linear**.
- Theorem :

Theorem 6 *Empirical risk minimization is consistent with respect to $\mathcal{F}$ if and only if $VC(\mathcal{F})$ is finite.*

- *Finiteness of VC-dimension also implies fast convergence.*

Rate of Convergence of Learning Process
VC-dimension

- *Equivalent Definition of the VC Dimension*
  - The ***VC-dimension*** *of a* ***set of indicator functions*** $Q(z,\alpha)$, $\alpha \in \Lambda$ is the **maximum number *h* of vectors** z_1 , ... , z_l **which can be separated in all 2^h** possible ways using functions of this set.
  - *If* ***for any n, there*** **exists a set of *n* vectors** which **can be shattered by the set $Q(z,\alpha)$, $\alpha \in \Lambda$** then the VC-dimension is equal **to infinity**.

Rate of Convergence of Learning Process

VC-dimension

- Example:

The VC-dimension of the set of *linear indicator functions*

$$Q(z,\alpha) = \theta\left\{\sum_{p=1}^{n} \alpha_p z_p + \alpha_0\right\}$$

in n-dimensional coordinate space $Z = (z_1, \cdots, z_n)$ is equal to $h = n+1$, since using functions of this set one can shatter at most $n+1$ vectors. Here $\theta\{\cdot\}$ is the step function, which takes value one, if the expression in the brackets is positive and takes value zero otherwise.

Rate of Convergence of Learning Process

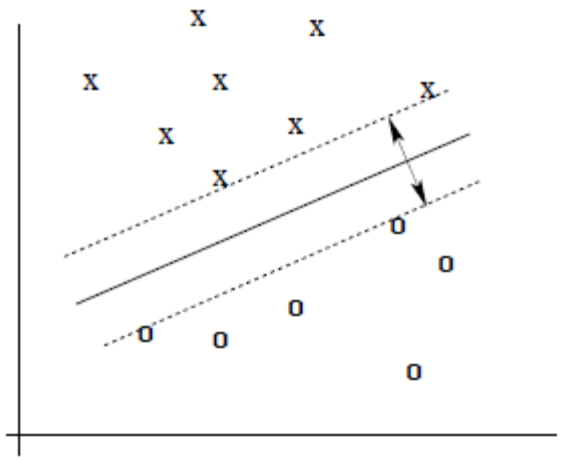
VC-dimension

- Δ-margin separating hyper plane is

$$(w^* \cdot x) - b = 0, \qquad |w^*| = 1$$

- If it classifies vectors x as

$$y = \begin{cases} 1, & \text{if } (w^* \cdot x) - b \geq \Delta \\ -1, & \text{if } (w^* \cdot x) - b \leq -\Delta. \end{cases}$$

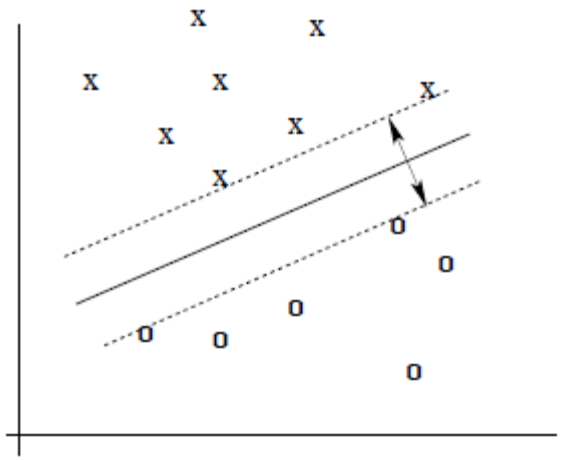


Rate of Convergence of Learning Process

VC-dimension

Theorem: Let vectors $x \in X$ belong to a sphere of radius R. Then the set of Δ-margin separating hyperplanes has the VC dimension h bounded by the inequality

$$h \leq \min\left(\left[\frac{R^2}{\Delta^2}\right], n\right) + 1.$$



Bounds On The Risk

Theorem 4.2. *For any ℓ the inequality*

$$P\left\{\sup_{\alpha\in\Lambda}\frac{R(\alpha)-R_{\text{emp}}(\alpha)}{\sqrt{R(\alpha)}}>\varepsilon\right\}<4\exp\left\{\left(\frac{H^{\Lambda}_{\text{ann}}(2\ell)}{\ell}-\frac{\varepsilon^2}{4}\right)\ell\right\} \tag{4.31}$$

holds true.

Corollary. *For the existence of nontrivial exponential bounds on uniform relative convergence it is sufficient that*

$$\lim_{\ell\to\infty}\frac{H^{\Lambda}_{\text{ann}}(\ell)}{\ell}=0. \tag{4.32}$$

Bounds On The Risk

- For the set of totally Bounded functions

Theorem: With probability at least $1-\eta$, the inequality

$$R(\alpha) \le R_{\text{emp}}(\alpha) + \frac{B\varepsilon}{2}\left(1+\sqrt{1+\frac{4R_{\text{emp}}(\alpha)}{B\varepsilon}}\right) \qquad (*)$$

holds true simultaneously for all functions of the set

$$0 \le Q(z,\alpha) \le B, \qquad \alpha \in \Lambda.$$

where

$$\varepsilon = 4\,\frac{h\left(\ln\frac{2\ell}{h}+1\right)-\ln\eta}{\ell}.$$

Bounds On The Risk

Theorem 4.4. *For a set of indicator functions $Q(z, \alpha)$, $\alpha \in \Lambda$, with finite VC dimension h the following two inequalities hold true:*

1. *The inequality estimating the rate of two-sided uniform convergence:*

$$P\left\{\sup_{\alpha\in\Lambda}\left|\int Q(z,\alpha)\,dF(z) - \frac{1}{\ell}\sum_{i=1}^{\ell} Q(z_i,\alpha)\right| > \varepsilon\right\}$$
$$< 4\exp\left\{\left(\frac{h(1+\ln(2\ell/h))}{\ell} - \varepsilon_*^2\right)\ \ell\right\}, \qquad (4.46)$$

   where $\varepsilon^* = (\varepsilon - 1/\ell)$, and

2. *The inequality estimating the rate of relative uniform convergence:*

$$P\left\{\sup_{\alpha\in\Lambda}\frac{\int Q(z,\alpha)\,dF(z) - \frac{1}{\ell}\sum_{i=1}^{\ell} Q(z_i,\alpha)}{\sqrt{\int Q(z,\alpha)\,dF(z)}} > \varepsilon\right\}$$
$$< 4\exp\left\{\left(\frac{h(1+\ln(2\ell/h))}{\ell} - \frac{\varepsilon^2}{4}\right)\ell\right\}. \qquad (4.47)$$

Theory of Constructing Learning Algorithms

- SVM
  - Optimal hyperplane (maximal margin) , the hyperplane that can separate data without error and **the distance** between the closest vector and the hyperplane is **maximal.**
  - **Map** the input vectors into **a very high-dimensional** feature space Z through some nonlinear mapping
  - Construct Optimal separating hypeplane
  - **VC-dimension is R^2/ Δ^2**
  - **To generalize well**, **decrease the VC dimension** by constructing an **optimal separating hyperplane** (that **maximizes the margin**)

Occam's Razor vs VC-Dimension

- Occam's Razor:
  - *Entities should not be multiplied beyond necessity*

  Or
  - **The simplest explanation is the best**
- Vapnik proves it is wrong, SRM Theory:
  - **The explanation by the machine with the smallest capacity (VC-dimension) is the best.**
  - Naïve notion of complexity (e.g. number of parameters) do not necessarily reflect capacity properly

Statistical Learning Theory

- Four Parts of Learning Theory (LT)
  - LT has to address the following four questions
    - What are the **conditions for consistency** of the **ERM principle**?
    - How **fast** does the sequence of smallest **empirical risk values converge to the smallest actual risk**?
    - How can one **control the rate of convergence** of the learning machine?
    - How can we construct algorithms that can control the **rate of generalization**?
  - The answers to this questions form the four parts of LT
    - The theory of **consistency of Learning Process** (LP)
    - The non-asymptotic theory of the **rate of convergence** of LP
    - The theory of controlling the **generalization** of LP
    - The theory of **constructing** Learning algorithms

Refrences

- Statistical learning theory, vapnik, 1998
- The nature of statistical learning theory, 1995
- An Overview of statistical learning theory, 1999
- Introduction to statistical learning theory, Bousquet, Boucheron, Lugosi, 2003
- Statistical learning theory: models, Concepts and results, Luxburg, Scholkpof
- Elements of statistical learning theory, smola, sholkpof
- A tutorial on Support Vector Machines for Pattern recognition, C.J. Burges , 1998

Rademacher Complexity

- It depends on the underlying distribution and usually lends to much sharper bounds than VC-Dimension bounds.
- Let $\sigma_1, \sigma_2, \ldots$ *independent random variables which attain two values +1 and -1 with probability 0.5 each.*
- *Rademacher Complexity defined as*

$$\mathcal{R}(\mathcal{F}) := E \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(X_i).$$

From a mathematical point of view, the Rademacher complexity is convenient to work with. One can prove generalization bounds of the following form: with probability at least $1 - \delta$,

$$R(f) \leq \mathrm{R}_{\mathrm{emp}}(f) + 2\mathcal{R}(\mathcal{F}) + \sqrt{\frac{\log(1/\delta)}{2n}}$$

- The Discovery that the generalization ability of the learning machine depends on the capacity of the set of functions implemented by the learning machine which differ from the number of free parameters is one of the most important achievements of the new theory.
- The main principle of inference from a small sample size:
  - If you possess a restricted amount of information for solving some problem, try to solve the problem directly and never solve a more general problem as an intermediate step. It is possible that the available information is sufficient for a direct solution but is insufficient for solving a more general intermediate problem.
  - Inductive inference: from particular to general
  - Deductive inference: from general to particular
  - Transductive inference: the direct estimation of values of a function only at points of interest using a given set of functions forms a new type of inference which can be called transductive inference. From particular to particular.
  - Convergence in probability and almost sure covergence.

Theory for Controlling The Generalization of Learning Machine

- Goal is to specify methods which are appropriate for a given sample size.
- **ERM principle** is intended for **dealing with a large sample size**. When *l/h is large the second summand on the right hand side of inequality (*) becomes small. The actual risk is then close to empirical risk.*
- But if *l/h is small,* ***a small*** $R_{emp}(\alpha)$ *does not guarantee a small value of risk.*
- *Minimization of R(α) requires a new principle based on the* ***simultaneous minimization*** *of the two terms in (*).*
- *The first term depend on the value of* ***the empirical risk***
- *The second term depends on the* ***VC-dimension*** *of the learning machine.*

Theory for Controlling The Generalization of Learning Machine

- ***Structural Risk Minimization induction Principle***

  Let S the set of functions $Q(z,\alpha)$, $\alpha \in \Lambda$, be provided with a *structure; so that is composed of the* ***nested subsets*** *of* functions $S_k=\{Q(z,\alpha) , \alpha \in \Lambda_k \}$, such that

$$s_1 \subset s_2 \subset \ldots \subset s_n \subset \ldots \quad and$$

$$s^* = \cup_k s_k$$

Theory for Controlling The Generalization of Learning Machine

- Structural Risk Minimization induction Principle, continued

  An *admissible structure is one satisfying the following three* properties.

  1) The set S* is everywhere dense in S
  2) The VC-dimension h_k of each set S_k of functions is finite.
  3) Any element S_k of the structure contains totally bounded functions

$$0 \leq Q(z,\alpha) \leq B_k \, , \, \alpha \in \Lambda_k$$

  The **SRM principle** suggests that for a given set of observations z_1 , ... , z_l , choose **the element of structure S_n** , where *n=n(l)* and choose **the particular function** from S_n , for which the guaranteed risk (*) is minimal.

In topology and related areas of mathematics, a subset *A* of a topological space *X* is called **dense** (in *X*) if any point *x* in *X* belongs to *A* or is a limit point of *A*.[1] Informally, for every point in *X*, the point is either in *A* or arbitrarily "close" to a member of *A* - for instance, every real number is either a rational number or has one arbitrarily close to it (see Diophantine approximation).

Theory for Controlling The Generalization of Learning Machine

- SRM Principle

Theorem: For admissible structures the method of structural risk minimization provides approximations $Q(z, \alpha_\ell^{n(\ell)})$ for which the sequence of risks $R(\alpha_\ell^{n(\ell)})$ converge to the best one $R(\alpha_0)$ with asymptotic rate of convergence[10]

$$V(\ell) = r_{n(\ell)} + B_{n(\ell)} \sqrt{\frac{h_{n(\ell)} \ln \ell}{\ell}} \tag{25}$$

if the law $n = n(\ell)$ is such that

$$\lim_{\ell \to \infty} \frac{B_{n(\ell)}^2 h_{n(\ell)} \ln \ell}{\ell} = 0. \tag{26}$$

In (25) B_n is the bound for functions from S_n and $r_n(\ell)$ is the rate of approximation

$$r_n = \inf_{\alpha \in \Lambda_n} \int Q(z, \alpha)\, dP(z) - \inf_{\alpha \in \Lambda} \int Q(z, \alpha)\, dP(z).$$

Theory of Constructing Learning Algorithms

- To implement SRM Principle, two factors in inequality (*), must be controlled
  1) the value of **empirical risk**
  2) the **capacity factor**
- We consider, Methods of separating Hyperplanes
- Is not flexible enough to provide low empirical risk for many real-life problems
- To increase the flexibility of the sets of functions:
  - 1) to use **a richer set of indicator functions** which are superpositions of **linear indicator functions** -> leads to neural networks
  - 2) to map the input vectors **in high dimensional feature space** and construct in this space a **Δ-margin separating hypeplane** -> **leads to SVM**

Theory for Controlling The Generalization of Learning Machine

- SRM Principle
  - Actually **suggests a tradeoff** between **the quality of the approximation** and the **complexity of the approximation** function.

Theorem: For any distribution function the SRM method provides convergence to the best possible solution with probability one.

In other words SRM method is universally strongly consistent.