Class 8

Generalization error, VC-dimension and VC bounds

Federico Girosi

Class 8

- Estimation error;
- Uniform convergence in probability;
- VC-dimension;
- VC bounds;
- approximation and estimation bounds;

Notations

- $I[f] = \int_{X \times Y} (y - f(\mathbf{x}))^2 P(\mathbf{x}, y) d\mathbf{x} dy$

- $I_{\text{emp}}[f] = \frac{1}{l} \sum_{i=1}^{l} (y_i - f(\mathbf{x}_i))^2$

- $f_0 = \arg\min_{f \in \mathcal{T}} I[f]$

- $f_n = \arg\min_{f \in H_n} I[f]$

- $\hat{f}_{n,l} = \arg\min_{f \in H_n} I_{\text{emp}}[f]$

- $I[f_n] \leq \epsilon(n)$

- With probability $1 - \delta$

$$|I[f] - I_{\text{emp}}[f]| \leq \Omega(n, l, \delta) \quad \forall f \in H_n$$

A general statement

With probability $1-\delta$:

$$E[(f_0(\mathbf{x}) - \widehat{f}_{n,l}(\mathbf{x}))^2] \leq 2\Omega(n,l,\delta) + \epsilon(n)$$

Generalization = estimation +

approximation

Empirical Risk Minimization

$P(\mathbf{x}, y)$ is unknown $\Rightarrow$ we cannot minimize $I[f]$.

However we can approximate $I[f]$ by the *empirical risk*

$$I_{\mathrm{emp}}[f] = \frac{1}{l} \sum_{i=1}^{l} (y_i - f(\mathbf{x}_i))^2$$

and find an approximate solution:

$$\hat{f}_{n,l}(\mathbf{x}) = \arg \min_{f \in H_n} I_{\mathrm{emp}}[f] \ .$$

This is the so called *empirical risk minimization principle*.

Empirical Risk Minimization

The Empirical risk minimization principle makes sense only if

$$\text{minimum of } I_{\text{emp}}[f]$$

$$P \Downarrow l \to \infty$$

$$\text{minimum of } I[f]$$

From the law of large numbers we know that, **for a given** $f \in H_n$

$$I_{\text{emp}}[f]$$

$$P \Downarrow l \to \infty$$

$$I[f]$$

but this is not enough to guarantee convergence of the minimum of I_{emp} to the minimum of $I[f]$.

Example

$h(x) = (1-x)^2 \ \Rightarrow \min_{x \in R} h(x) = 0$

$h_\sigma(x) = -2e^{-\frac{x^2}{\sigma^2}} + h(x)$

$\min_{x \in R} h_\sigma(x) = -2 \ \forall \sigma > 0$

$$\lim_{\sigma \to 0} \|h_\sigma - h\|_{L_2} = \lim_{\sigma \to 0} \int \ dx \ 4e^{-\frac{2x^2}{\sigma^2}} = 0$$

Therefore

$$\lim_{\sigma \to 0} h_\sigma = h \quad (in \ L_2)$$

but

$$\lim_{\sigma \to 0} \ \min_{x \in R} h_\sigma \neq \min_{x \in R} h$$

Convergence of minima to minima

- The previous examples shows that convergence in L_2 does **not** imply convergence of minima to minima.

- The same statement is true for convergence in probability.

- It is intuitive that convergence in L_∞ **does** imply convergence of minima to minima.

- Convergence in probability $\Rightarrow$ **uniform** convergence in probability?

The elements of H_n are usually functions of the form $f = f(\mathbf{x}, \alpha)$, where $\alpha \in \Lambda$ is a vector of n parameters. We make the dependence on α explicit by writing

$$I[\alpha] = \int (y - f(\mathbf{x}, \alpha))^2 P(\mathbf{x}, y) d\mathbf{x} dy$$

Defining $\mathbf{z} = (\mathbf{x}, y)$ and

$$Q(\mathbf{z}, \alpha) = (y - f(\mathbf{x}, \alpha))^2$$

we rewrite the expected risk as

$$I[\alpha] = \int \; Q(\mathbf{z}, \alpha) P(\mathbf{z}) d\mathbf{z}$$

From now on we will consider functionals of this form, where Q is "any" function.

Under the new definition of expected risk, the empirical risk can be written

$$I_{\text{emp}}[\alpha] = \frac{1}{l} \sum_{i=1}^{l} Q(\mathbf{z}_i, \alpha)$$

where the $\mathbf{z}_i$ are drawn from $P(\mathbf{z})$, which is unknown.

The minimum of the empirical risk is achieved for a value α_l of the parameters:

$$\alpha_l = \arg\min_{\alpha \in \Lambda} I_{\text{emp}}[\alpha]$$

Consistency of the empirical risk minimization method

The empirical risk minimization method is said to be **consistent** for the set $\{Q(\mathbf{z},\alpha), \alpha \in \Lambda\}$ and for the probability distribution $P(\mathbf{z})$ if

$$\lim_{l \to \infty} I_{\text{emp}}[\alpha_l] \stackrel{P}{=} \inf_{\alpha \in \Lambda} I[\alpha]$$

and

$$\lim_{l \to \infty} I[\alpha_l] \stackrel{P}{=} \inf_{\alpha \in \Lambda} I[\alpha]$$

Key theorem of learning theory (Vapnik and Chervonenkis, 1989)

Let $Q(\mathbf{z}, \alpha)$ be such that

$$A \leq Q(\mathbf{z}, \alpha) \leq B$$

The **necessary** and **sufficient** condition for consistency is

$$\lim_{l \to \infty} P\{\sup_{\alpha \in \Lambda}(I[\alpha] - I_{\text{emp}}[\alpha]) > \epsilon\} = 0 \quad \forall \epsilon > 0$$

This notion of convergence is called **one sided uniform convergence in probability**.

Necessary and sufficient conditions for consistency

For a *given*, fixed probability distribution $P(\mathbf{z})$, the necessary and sufficient conditions for consistency are obtained in terms of **random VC entropy** of the set of functions $Q(\mathbf{z}, \alpha)$.

VC entropy for indicator functions

We assume now that $Q(\mathbf{z}, \alpha)$ is indicator function, that is a function with binary values, 0 and 1.

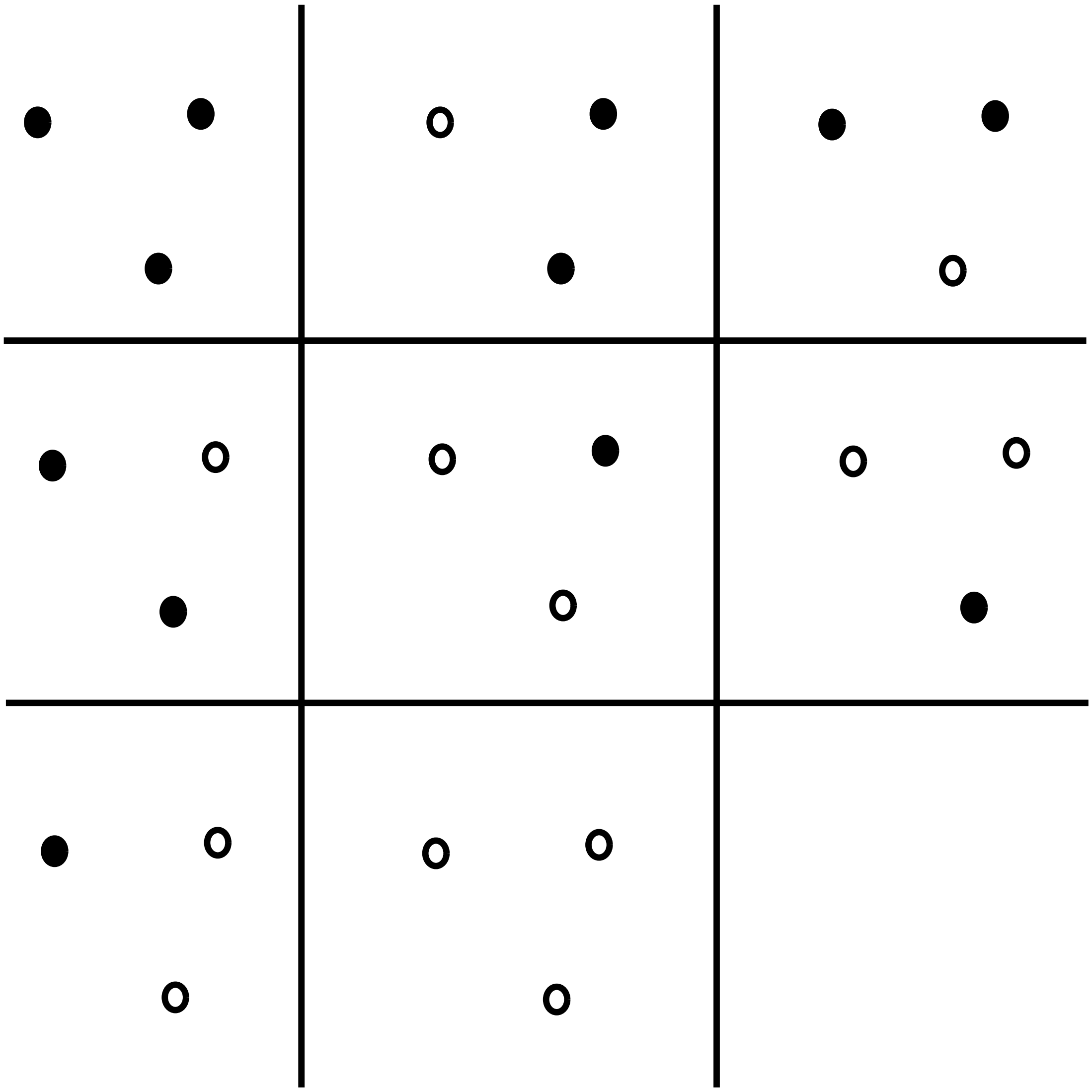
For a given value of α and a given a sample $\mathbf{z}_1, \ldots, \mathbf{z}_l$, the binary vector

$$q(\alpha) = (Q(\mathbf{z}_1, \alpha), \ldots, Q(\mathbf{z}_l, \alpha))$$

is a vertex of the l-dimensional cube $[0, 1]^l$.

$q(\alpha)$ represents a particular labeling of the sample $\mathbf{z}_1, \ldots, \mathbf{z}_l$.

The 8 possible labelings of 3 points in 2 dimensions.



VC-entropy for indicator functions

As the parameters α vary in the set Λ, different labelings can be implemented (or, alternatively, different vertices of the $[0,1]^l$ cube are spanned).

We define

$$N^{\Lambda}(\mathbf{z}_1, \ldots, \mathbf{z}_l)$$

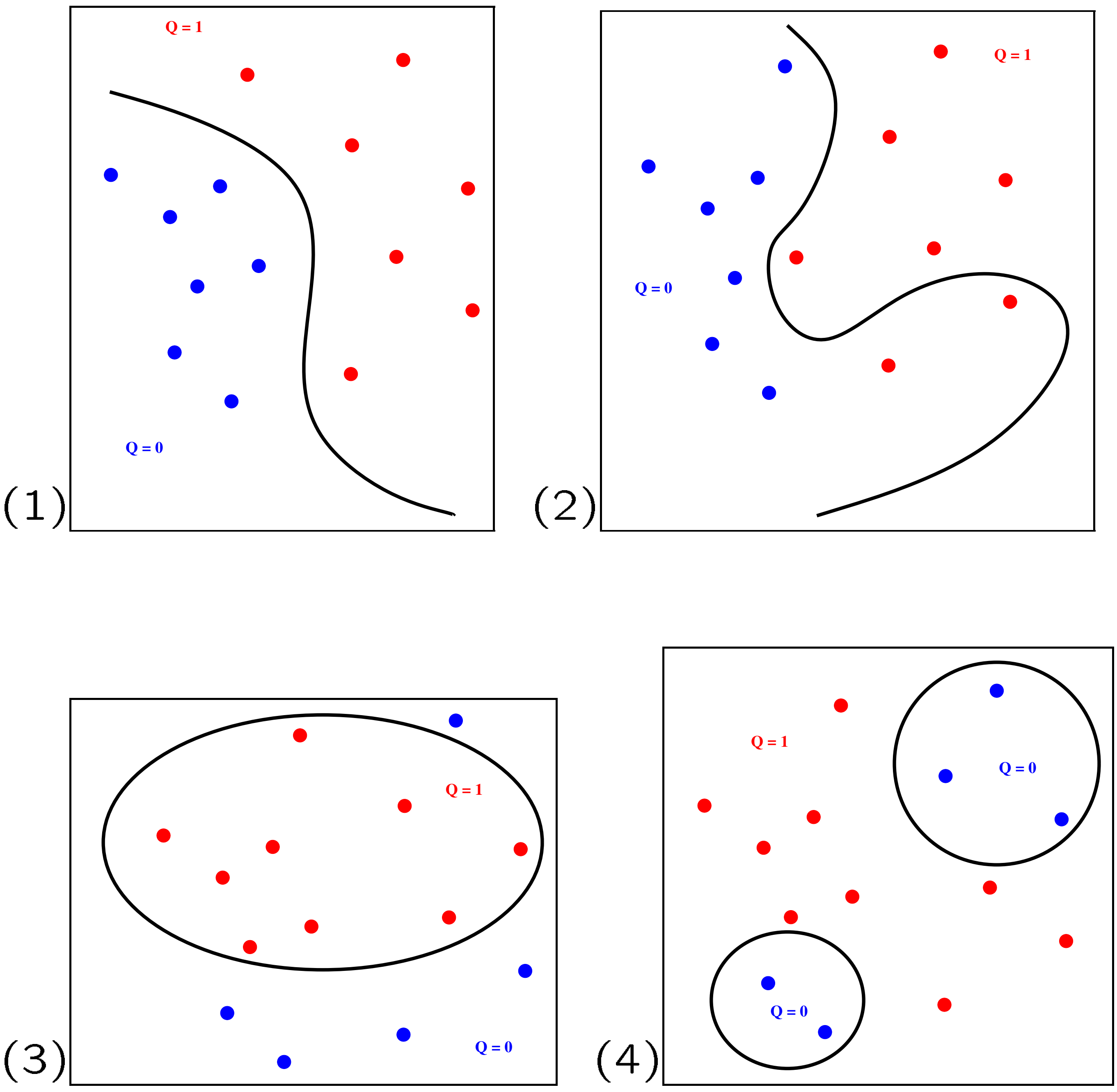
as the number of labelings that can be implemented as α varies in Λ.

$N^{\Lambda}(\mathbf{z}_1, \ldots, \mathbf{z}_l)$ depends on $(\mathbf{z}_1, \ldots, \mathbf{z}_l)$ and Q.

An obvious property of $N^{\Lambda}(\mathbf{z}_1, \ldots, \mathbf{z}_l)$ is:

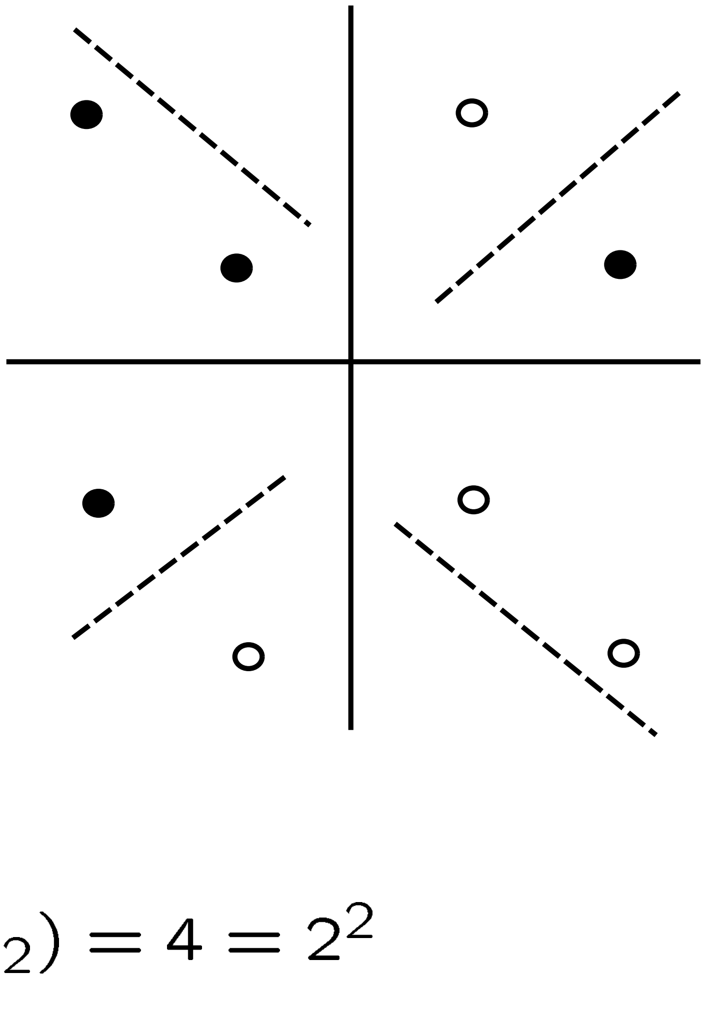
$$N^{\Lambda}(\mathbf{z}_1, \ldots, \mathbf{z}_l) \leq 2^l$$

Implementation of different labelings



Example

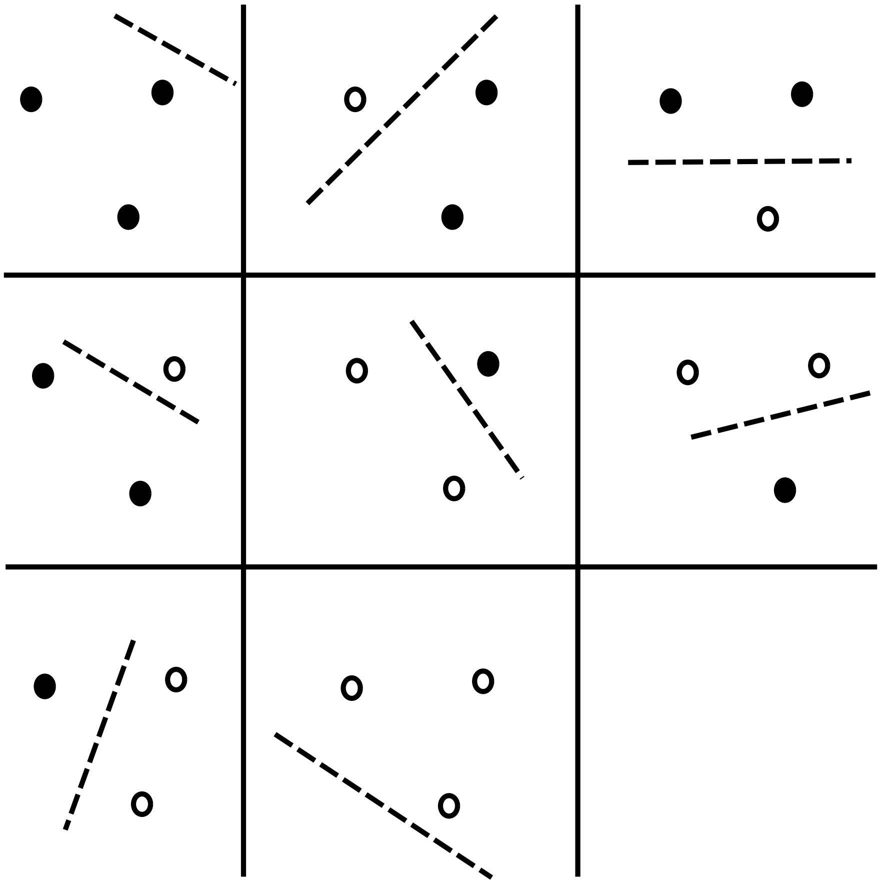
$Q(\mathbf{z}, \alpha) = \theta(\mathbf{z} \cdot \mathbf{w} + b)$, with $\alpha = (\mathbf{w}, b)$ and $l = 2$.



$$N^{\Lambda}(\mathbf{z}_1, \mathbf{z}_2) = 4 = 2^2$$

Example

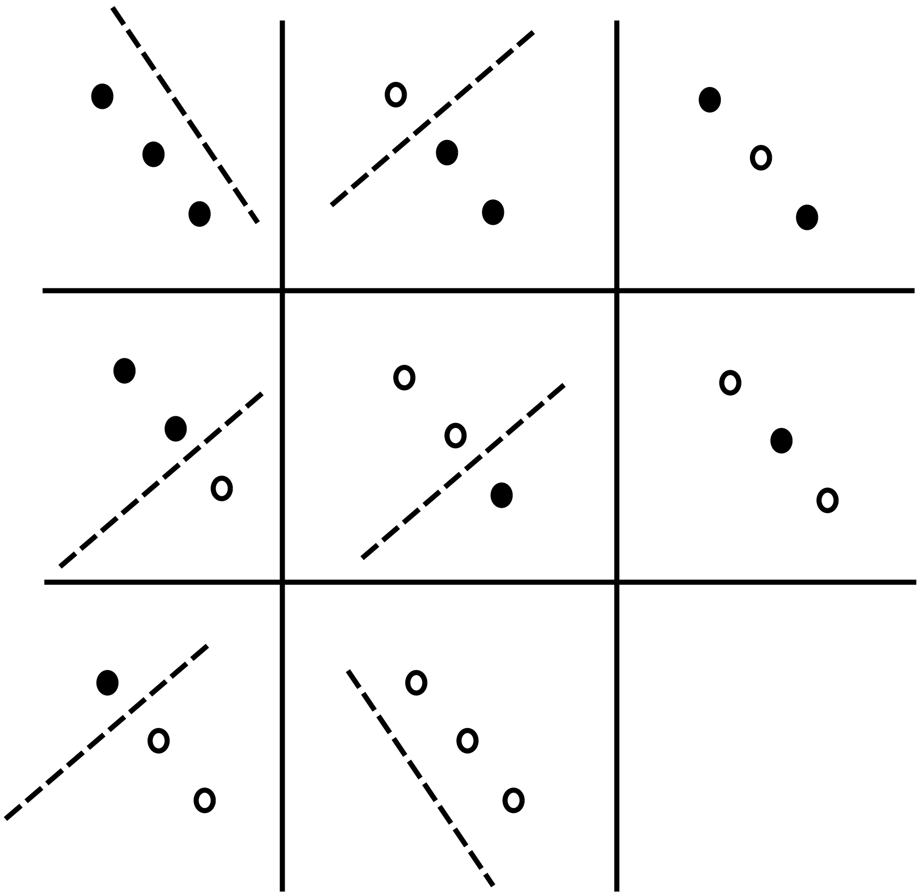
$Q(\mathbf{z}, \alpha) = \theta(\mathbf{z} \cdot \mathbf{w} + b)$, with $\alpha = (\mathbf{w}, b)$ and $l = 3$.



$$N^{\Lambda}(\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3) = 2^3 = 8$$

Example

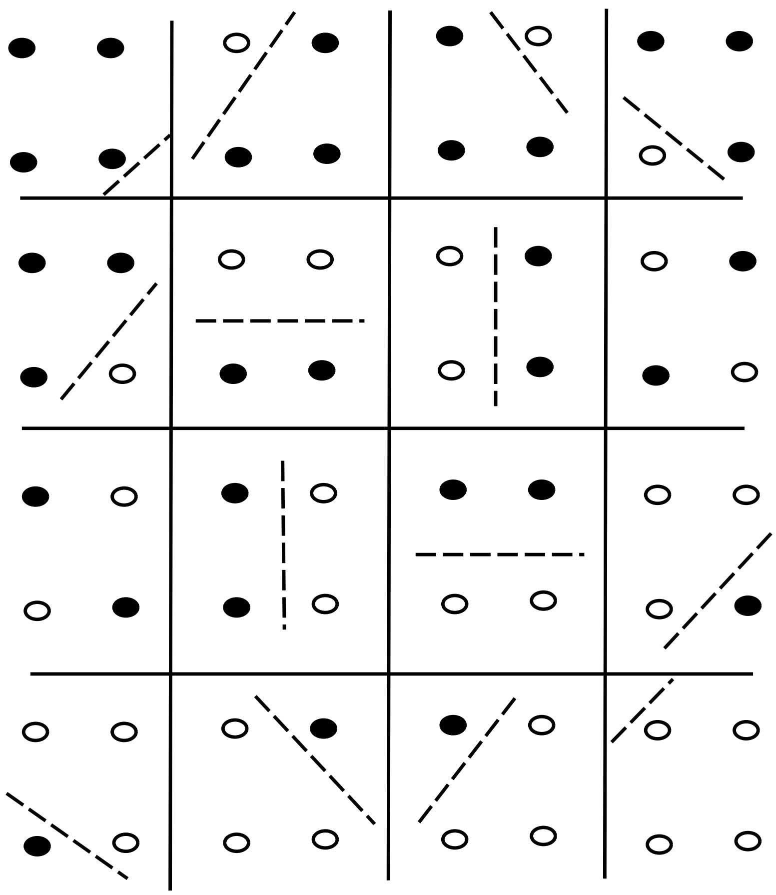
$Q(\mathbf{z}, \alpha) = \theta(\mathbf{z} \cdot \mathbf{w} + b)$, with $\alpha = (\mathbf{w}, b)$ and $l = 3$.



$$N^{\Lambda}(\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3) = 6 < 2^3 = 8$$

Example

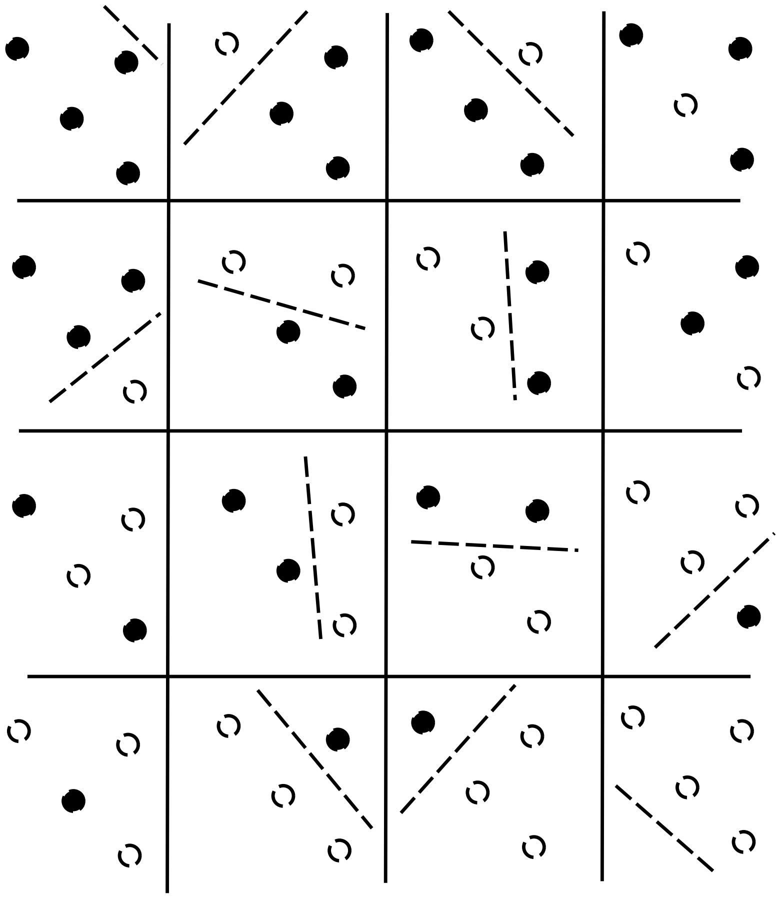
$Q(\mathbf{z}, \alpha) = \theta(\mathbf{z} \cdot \mathbf{w} + b)$, with $\alpha = (\mathbf{w}, b)$ and $l = 4$.



$$N^{\Lambda}(\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3, \mathbf{z}_4) = 14 < 2^4 = 16$$

Example

$Q(\mathbf{z}, \alpha) = \theta(\mathbf{z} \cdot \mathbf{w} + b)$, with $\alpha = (\mathbf{w}, b)$ and $l = 4$.



$$N^{\Lambda}(\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3, \mathbf{z}_4) = 12 < 2^4 = 16$$

VC-entropy for indicator functions

The VC-entropy for a set of indicator functions $Q(\mathbf{z}, \alpha)$ and a given probability distribution $P(\mathbf{z})$ is defined as

$$H^{\Lambda}(l) = E[\ln N^{\Lambda}(\mathbf{z}_1, \ldots \mathbf{z}_l)]$$

where $E[\cdot]$ represents the expectation over $P(\mathbf{z}_1)P(\mathbf{z}_2)\ldots P(\mathbf{z}_l)$.

$H^{\Lambda}(l)$ is a measure of the diversity of the set of functions $Q(\mathbf{z}, \alpha)$.

The VC-entropy has the property that

$$0 \leq H^{\Lambda}(l) \leq l \ln 2$$

Two-sided uniform convergence in probability

The following kind of convergence is called **two-sided uniform convergence in probability**

$$\lim_{l\to\infty} P\{\sup_{\alpha\in\Lambda} |I[\alpha] - I_{\text{emp}}[\alpha]| > \epsilon\} = 0 \quad \forall\epsilon > 0$$

Two-sided uniform convergence is only **sufficient** condition for consistency.

Two-sided uniform convergence in probability

Necessary and sufficient condition for two-sided uniform convergence in probability is

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

Notice that this is a **sufficient** condition for consistency.

This result can be extended to prove necessary and sufficient conditions for *bounded* functions $Q(\mathbf{z}, \alpha)$.

Consistency and VC-entropy

The fact that in order to achieve consistency

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

means that the VC-entropy should not grow very fast, that is the class of functions $Q(\mathbf{z}, \alpha)$ should not be "too rich".

If all the possible labelings can always be implemented, that is if

$$N^{\Lambda}(\mathbf{z}_1, \ldots \mathbf{z}_l) = 2^l \quad \forall l$$

Then

$$\lim_{l \to \infty} \frac{H^{\Lambda}(l)}{l} = 2$$

and consistency is not guaranteed: no learning takes place.

Beyond the VC-entropy

Studying the VC-entropy is not enough:

- The VC-entropy depends on the specific probability distribution $P(\mathbf{z})$, which is usually unknown;

- Although uniform convergence (and consistency) is guaranteed, the rate of convergence might be arbitrarily slow;

It is therefore important to study under which conditions consistency is achieved for **arbitrary** *probability distributions, with* **fast** *rate of convergence.*

Annealed VC-entropy

The **annealed VC-entropy** is defined as

$$H_{\text{ann}}^{\Lambda}(l) = \ln E[N^{\Lambda}(\mathbf{z}_1, \ldots \mathbf{z}_l)]$$

The condition

$$\lim_{l \to \infty} \frac{H_{\text{ann}}^{\Lambda}(l)}{l} = 0$$

is **sufficient** to guarantee fast rate of convergence for a fixed probability distribution $P(\mathbf{z})$.

Note that

$$H^{\Lambda}(l) \leq H_{\text{ann}}^{\Lambda}(l)$$

A necessary and sufficient condition for consistency, for any probability distributions $P(\mathbf{z})$ and at a fast rate of convergence is:

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

where we have defined the **growth function** $G^{\Lambda}(l)$ as

$$G^{\Lambda}(l) = \ln \sup_{\mathbf{z}_1,\ldots,\mathbf{z}_l} N^{\Lambda}(\mathbf{z}_1, \ldots, \mathbf{z}_l)$$

Notice that the definition of $G^{\Lambda}(l)$ does not depend on the probability distribution $P(\mathbf{z})$: $G^{\Lambda}(l)$ only depends on the class of functions $Q(\mathbf{z}, \alpha)$.

Estimation error bounds

For indicator functions $Q(\mathbf{z}, \alpha)$ and for any probability measure $P(\mathbf{z})$:

$$P\{\sup_{\alpha \in \Lambda} |I[\alpha] - I_{\mathrm{emp}}[\alpha]| > \epsilon\} \leq 4 e^{\left(\frac{G^{\Lambda}(2l)}{l} - \epsilon^2\right) l}$$

(Vapnik, 1982)

Estimation error bounds

Alternatively one can say that, with probability $1-\delta$, the following inequality holds:

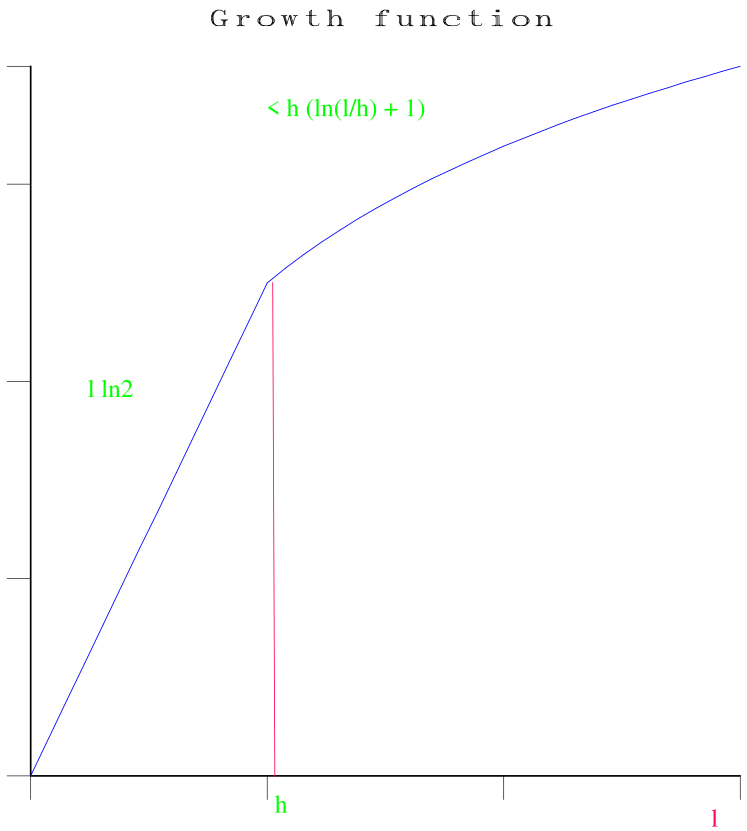
$$\sup_{\alpha \in \Lambda} |I[\alpha] - I_{\text{emp}}[\alpha]| \leq \sqrt{\frac{G^{\Lambda}(2l) - \ln\left(\frac{\delta}{4}\right)}{l}}$$

However, this is inequality is not very useful unless we know how to compute $G^{\Lambda}(l)$ for the class of functions $Q(\mathbf{z}, \alpha)$ of interest.

We need to study the behavior of $G^{\Lambda}(l)$

A remarkable theorem (Vapnik, 1982)

The growth function $G^{\Lambda}(l)$ has always the following shape



where h is an integer (possibly infinite) that is called the **VC-dimension** of the set of functions $Q(\mathbf{z}, \alpha)$.

VC-dimension

The VC-dimension of a set of indicator functions $Q(\mathbf{z}, \alpha)$ is h if and only if

- There is **at least one set of h points** that can be labeled in all possible ways;

- there is **no set of $h+1$ points** that can be labeled in all possible ways;

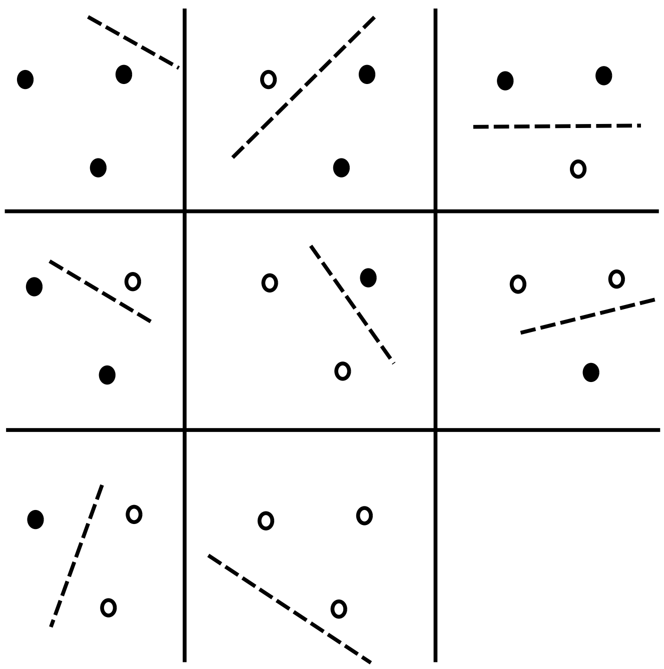
Necessary and sufficient condition for consistency of the empirical risk minimization method, for arbitrary probability distribution $P(\mathbf{z})$ and fast rate of convergence:

The VC-dimension of the set of functions $Q(\mathbf{z}, \alpha)$ is finite

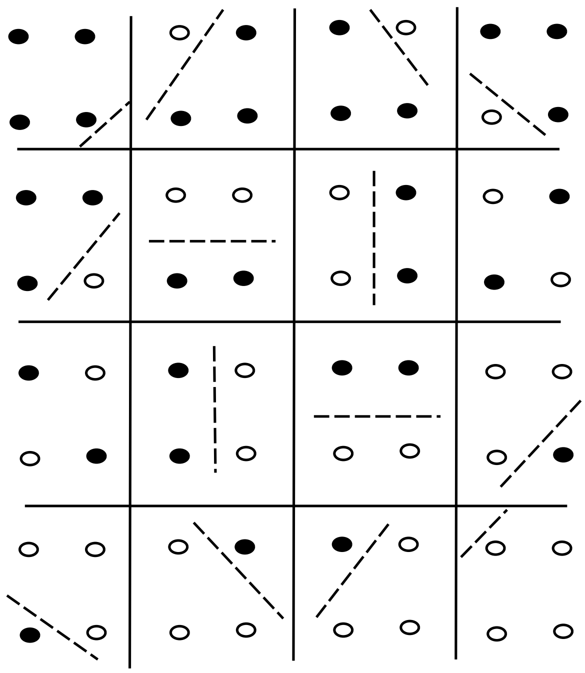
Example

VC-dimension of hyperplanes:

$$Q(\mathbf{z}, \alpha) = \theta(\mathbf{z} \cdot \mathbf{w} + b)$$



all the possible labelings



not all the possible labelings

VC-dimension = 3

VC-dimension

The VC-dimension of a set of indicator functions $Q(\mathbf{z}, \alpha)$ is a measure of the complexity of the set, and it is related to the number of free parameters α.

If $Q(\mathbf{z}, \alpha)$ has the form

$$Q(\mathbf{z}, \alpha) = \theta \left(\sum_{k=1}^{n} \alpha_k \phi_k(\mathbf{z}) \right)$$

where $\phi_k(\mathbf{z})$ are linearly independent functions and $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n)$, then $h = n$, that is the VC-dimension is equal to the number of free parameters.

Estimation error bounds

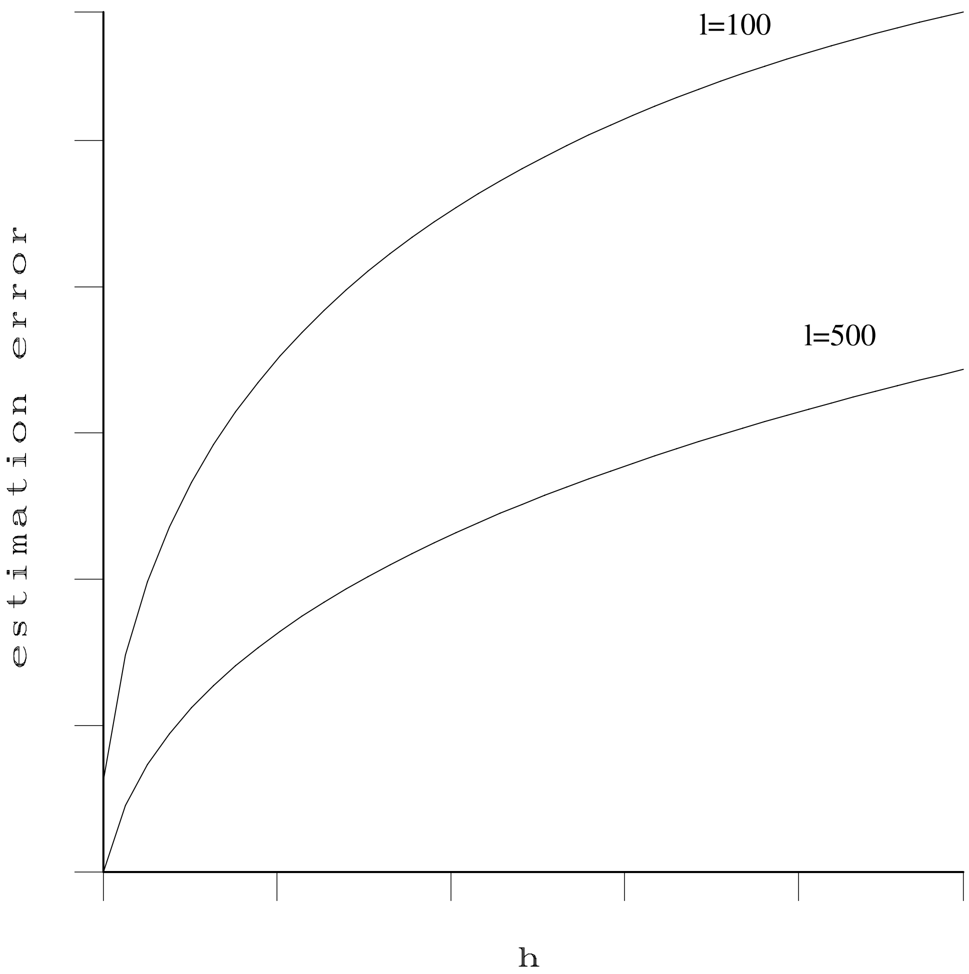
Using the fact that

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

for $l > h$, a typical uniform VC bound for indicator functions looks like:

$$\sup_{\alpha\in\Lambda} |I[\alpha] - I_{\mathrm{emp}}[\alpha]| \leq \sqrt{\frac{h\left(\ln \frac{2l}{h} + 1\right) - \ln\left(\frac{\delta}{4}\right)}{l}}$$

where the bound holds with probability $1 - \delta$.



Estimation error as a function of the VC dimension for different numbers of data points.

Estimation error bounds

The intuition behind the VC bound:

$$\sup_{\alpha \in \Lambda} |I[\alpha] - I_{\text{emp}}[\alpha]| \leq \sqrt{\frac{h\left(\ln \frac{2l}{h} + 1\right) - \ln\left(\frac{\delta}{4}\right)}{l}}$$

is that the uniform deviation between expected risk and empirical risk decreases with the number of data points l, but **increases with the VC-dimension**, and therefore with the complexity of the set of indicator functions $Q(\mathbf{z}, \alpha)$.

VC-dimension for real functions

When $Q(\mathbf{z}, \alpha)$ is not an indicator function the usual definition of VC-dimension does not apply.

In this case the VC-dimension of the set of $Q(\mathbf{z}, \alpha)$ is defined as the VC-dimension of the indicator function $Q^*(\mathbf{z}, \alpha)$

$$Q^*(\mathbf{z}, \alpha^*) = \theta(Q(\mathbf{z}, \alpha) - \beta)$$

where $\alpha^* = (\alpha, \beta)$.

VC-dimension and approximation

In function approximation we are interested in an expected risk of the form:

$$I[\alpha] = \int (y - f(\mathbf{x}, \alpha))^2 P(\mathbf{x}, y) d\mathbf{x} dy$$

and therefore $Q(\mathbf{z}, \alpha) = (y - f(\mathbf{x}, \alpha))^2$ with $\mathbf{z} = (\mathbf{x}, y)$.

It can be shown that the VC-dimension of $Q(\mathbf{z}, \alpha)$ is equal to the VC-dimension of the indicator function:

$$Q^*(\mathbf{z}, \alpha^*) = \theta(f(\mathbf{x}, \alpha) - \beta)$$

The VC-dimension is therefore a complexity measure of the hypothesis space H_n.

The VC-dimension is proportional, but not necessarily equal, to the number of parameters.

- For Multilayer Perceptrons with hard thresholds $h \propto n \ln n$ (Maass, 1994);

- For Multilayer perceptrons with standard sigmoid thresholds $h \propto n^2$ (Koiran and Sontag, 1995);

- For functions of the form $f(z, \alpha) = \theta(\sin(\alpha z))$ the VC-dimension is infinite;

- For the set of monotonic functions
$$f(z, \alpha) = \sum_{i=1}^{n} c_i \sigma(w_i z + b_i), \quad c_i \geq 0$$
the VC-dimension is $h = 1$, independently of n.

Generalization error bound

We have already shown that, with probability $1-\delta$:

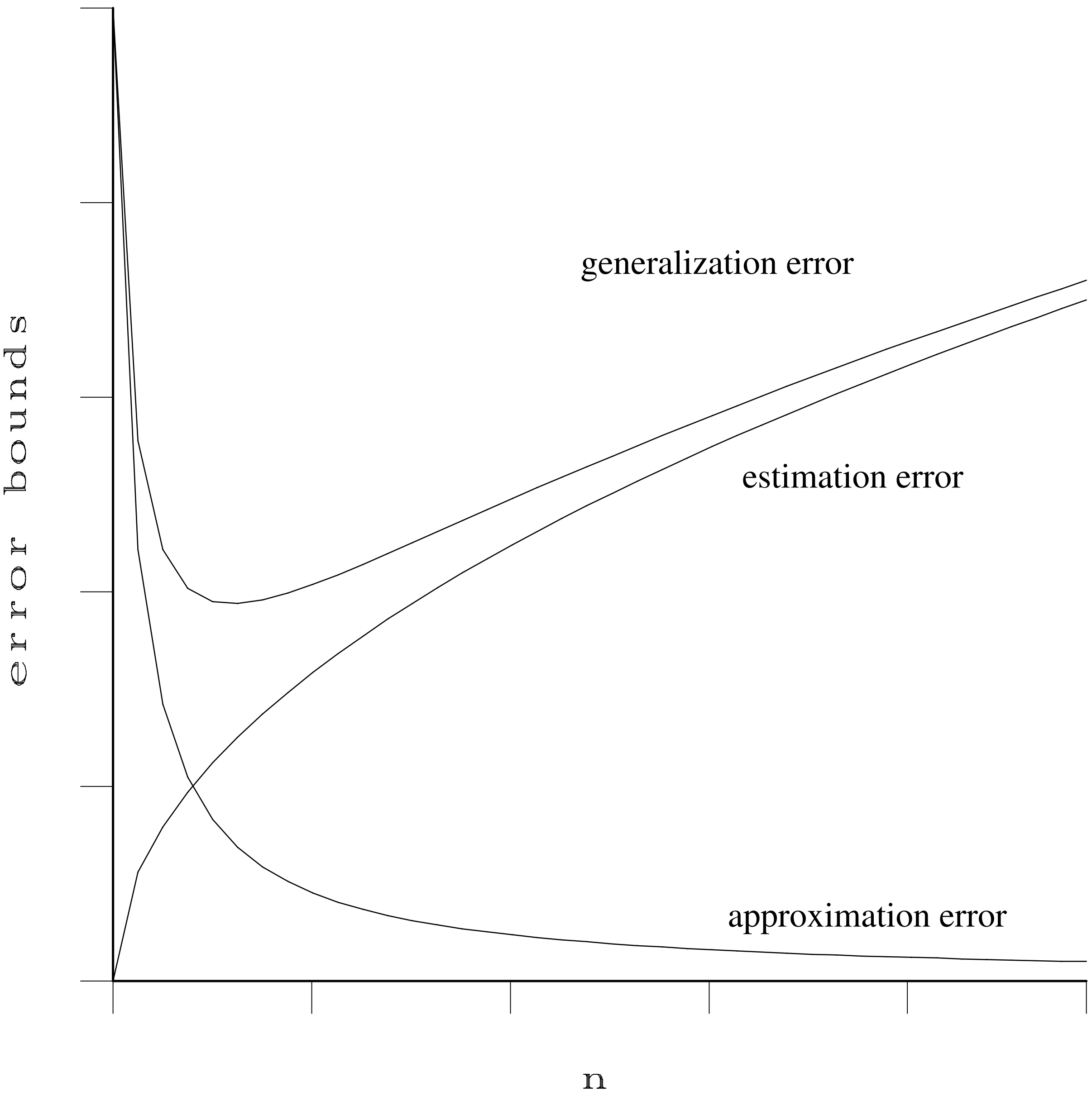
$$E[(f_0(\mathbf{x}) - \hat{f}_{n,l}(\mathbf{x}))^2] \leq 2\Omega(n, l, \delta) + \epsilon(n)$$

where Ω is a VC bound and $\epsilon(n)$ is a bound on the approximation error. Assuming that:

- approximation error $\propto n^{-\alpha}$;
- VC-dimension $\propto n^{\beta}$;

and omitting logarithimic factors, a typical generalization error bound is:

$$E(n, l) \equiv E[(f_0(\mathbf{x}) - \hat{f}_{n,l}(\mathbf{x}))^2] \leq \frac{1}{n^{\alpha}} + \sqrt{\frac{n^{\beta}}{l}}$$


generalization error
estimation error
approximation error
error bounds
n

Generalization error bound

Assuming that

$$E(n,l) \leq \frac{1}{n^{\alpha}} + \sqrt{\frac{n^{\beta}}{l}}$$

for any given l the optimal number of parameters is given by

$$n \propto l^{\frac{1}{\beta+2\alpha}}$$

and therefore the generalization error has a rate of convergence which is at most

$$E(n,l) \leq l^{-\frac{1}{2+\frac{\beta}{\alpha}}}$$

This suggests that $l^{-\frac{1}{2}}$ is the best possible rate.

Theorem (Barron, 1991)

Let $\hat{f}_{n,l}$ be a Multilayer Perceptron with n hidden units trained on l examples, subject to the constraint that the parameters of the network belong to a uniform grid of appropriate spacing. If the target function h is such that

$$\int_{R^d} d\mathbf{s}\ \|\mathbf{s}\||\tilde{h}(\mathbf{s})| < +\infty \ .$$

then

$$E[(h - \hat{f}_{n,l})^2] \leq O\left(\frac{1}{n}\right) + O\left(\frac{nd\ln(l)}{l}\right)$$

The optimal number of parameters is:

$$n^*(l) \propto \left[\frac{l}{d\ln l}\right]^{\frac{1}{2}}$$

and the rate of convergence of the generalization error is $O(l^{\frac{1}{2}})$.

Theorem (Niyogi and Girosi, 1993)

Let $h \in H^{m,1}(R^d)$, where $m > d$. Let $\hat{f}_{n,l}$ a superposition of n Gaussians, whose coefficients, centers and variances have been estimated using l data points drawn from h. Then with probability greater than $1-\delta$ the following inequality is valid:

$$E[(h - \hat{f}_{n,l})^2] \leq O\left(\frac{1}{n}\right) + O\left(\left[\frac{nd \ln(nl) - \ln\delta}{l}\right]^{1/2}\right)$$

The optimal number of parameters is:

$$n^*(l) \propto \left[\frac{l}{d \ln l}\right]^{\frac{1}{3}}$$

Generalization error