Class 6

Generalization error
and
approximation theory

Federico Girosi

Plan of the class

1. Learning and generalization error;
2. the approximation problem and rates of convergence;
3. complexity of spaces of functions, n-widths, metric entropy;
4. rates of convergence "independent of dimension";

Learning from examples

Let X and Y be two sets of random, non independent variables.

A data set $D = \{(\mathbf{x}_i, y_i) \in X \times Y\}_{i=1}^{l}$ is collected as following:

`For i=1` $\rightarrow$ `l`

1) a point $\mathbf{x}_i \in X$ is chosen, with probability $P(\mathbf{x}_i)$;

2) the corresponding variable $y_i \in Y$ is observed, with probability $P(y_i|\mathbf{x}_i)$;

The probability of collecting a data point $(\mathbf{x}, y)$ is therefore

$$P(\mathbf{x}, y) = P(\mathbf{x})P(y|\mathbf{x})$$

Learning means being able to infer the relationship between X and Y from the set D;

$$\Downarrow$$

for every $\mathbf{x} \in X$, to be able to predict the corresponding value of y with the minimum error;

$$\Downarrow$$

to produce an *hypothesis* $f(\mathbf{x})$ wich minimizes some risk function;

For every $\mathbf{x}$ and y drawn from $P(\mathbf{x}, y)$, we can measure the error of an hypothesis f (the *risk* associated to it) as

$$(y - f(\mathbf{x}))^2$$

The average error is

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

$I[f]$ is the so called *expected risk*

In this framework the goal of *learning* is to find an hypothesis f_{min} that minimizes the expected risk:

$$f_{\mathrm{min}}(\mathbf{x}) = \arg\min_{f \in \mathcal{H}} I[f]$$

where $\mathcal{H}$ is some set of functions, called the *hypothesis space*.

`This, of course, cannot be done, since we do not how to compute the expected risk` $I[f]$ `because we do not know` $P(\mathbf{x}, y)$ `...`

Notice that the expected risk $I[f]$ can be decomposed as follows:

$$I[f] = \int_X (f_0(\mathbf{x}) - f(\mathbf{x}))^2 P(\mathbf{x}) d\mathbf{x} + \int_{X \times Y} (y - f_0(\mathbf{x}))^2 P(\mathbf{x}, y) d\mathbf{x} dy$$

where

$$f_0(\mathbf{x}) = \int_Y y P(y|\mathbf{x}) dy$$

is the conditional mean of the output variables, the so called *regression function*.

1. The goal of learning is to find:

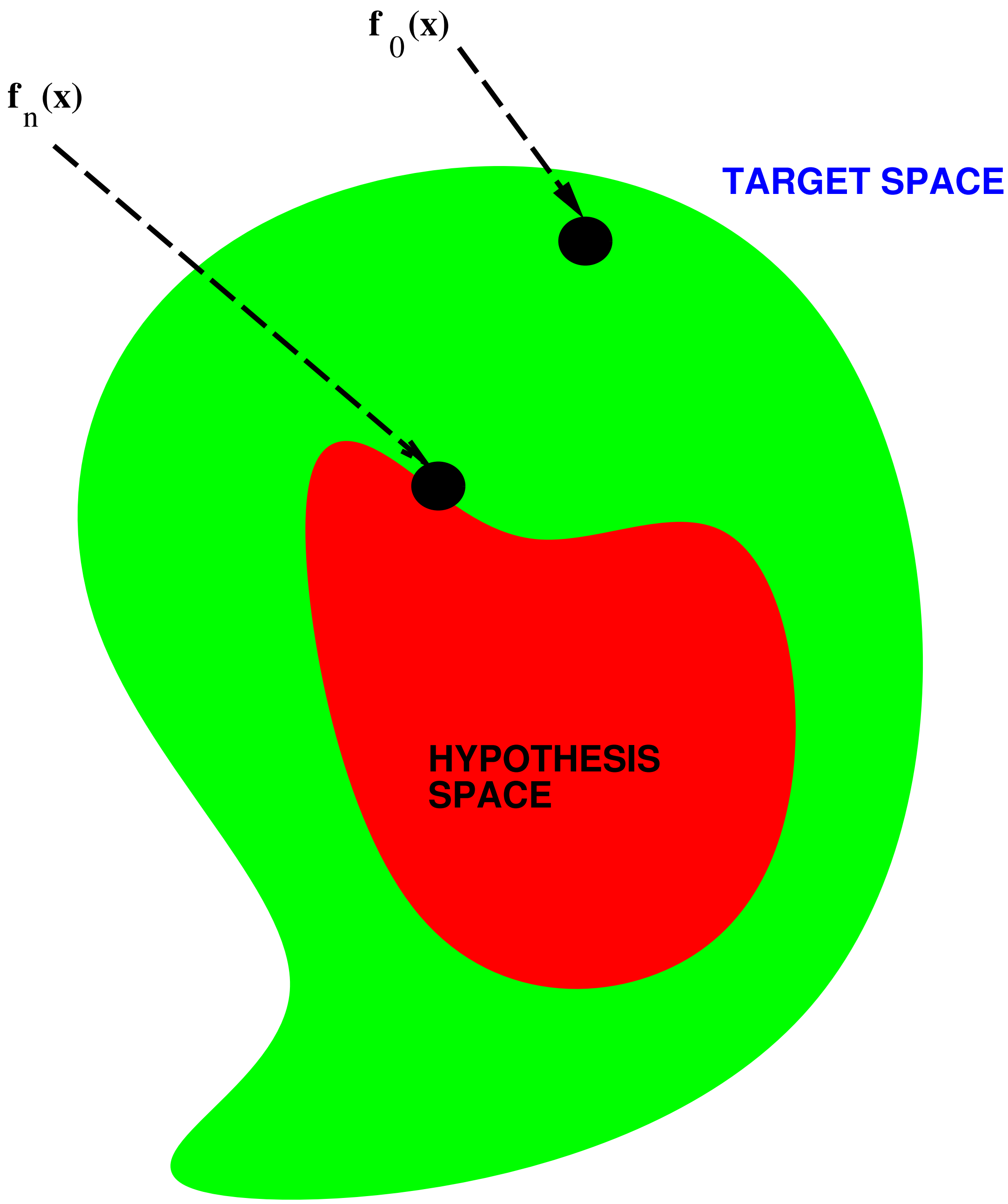
$$f_n = \arg \min_{f \in H_n} I[f]$$

2.

$$\arg \min_{f \in H_n} I[f] = \arg \min_{f \in H_n} \int_X (f_0(\mathbf{x}) - f(\mathbf{x}))^2 P(\mathbf{x}) d\mathbf{x}$$

3. Learning $\equiv$ to find the function $f_n \in H_n$ that best approximates the regression function $f_0(\mathbf{x})$, which is an unknown function belonging to some large space of functions, also called the *target space* $\mathcal{T}$.

   The target space $\mathcal{T}$ is usually a much larger set than the hypothesis space H_n.



$\mathbf{f}_0(\mathbf{x})$
$\mathbf{f}_n(\mathbf{x})$
TARGET SPACE
HYPOTHESIS SPACE

Examples of hypothesis spaces $\mathcal{H}$

- $\mathcal{H}$ = set of polynomials of degree n:

$$\mathcal{H} = \{g \; | g(x) = a_0 + a_1 x + a_2 x^2 + \ldots + a_n x^n\}$$

- $\mathcal{H}$ = truncated Fourier series

$$\begin{aligned}\mathcal{H} = \{g | g(x) = \; & a_0 + a_1 \sin(x) + b_1 \cos(x) + \\ & + a_2 \sin(2x) + b_2 \cos(2x) + \\ & + \ldots a_n \sin(nx) + b_n \cos(nx)\}\end{aligned}$$

- $\mathcal{H}$ = sphere in a Sobolev space

$$\mathcal{H} = \{f \in L_2 \mid f' \in L_2, \quad \|f'\|^2 \leq \mu\}$$

Examples of hypothesis spaces $\mathcal{H}$

- $\mathcal{H}$ = any *linear* basis function expansion:

  $$\mathcal{H} = \{g | g(\mathbf{x}) = \sum_{\alpha=1}^{n} c_\alpha G_\alpha(\mathbf{x})\}$$

  where $\{G_\alpha(\mathbf{x})\}_{\alpha=1}^{n}$ is a fixed set of basis functions;

- $\mathcal{H}$ = any *nonlinear* basis function expansion:

  $$\mathcal{H} = \{g | g(\mathbf{x}) = \sum_{\alpha=1}^{n} c_\alpha G(\mathbf{x}; \mathbf{w}_\alpha)\}$$

  where the c_α and the $\mathbf{w}_\alpha$ are free parameters and G is some given function;

Hypothesis spaces

$$f(x) = \sum_{k=1}^{n} c_k x^k$$

$$f(x) = \sum_{k=0}^{n} a_k \sin(kx) + b_k \cos(kx)$$

$$f(x) = \frac{\sum_{k=1}^{n} a_k x^k}{\sum_{k=1}^{m} b_k x^k}$$

$$f \in L_2 \ , \ f' \in L_2, \quad \|f'\|^2 \leq n$$

$$f(\mathbf{x}) = \sum_{\alpha=1}^{n} c_\alpha G(\|\mathbf{x} - \mathbf{t}_\alpha\|)$$

$$f(\mathbf{x}) = \sum_{\alpha=1}^{n} c_\alpha \sigma(\mathbf{x} \cdot \mathbf{w}_\alpha + \theta_\alpha)$$

$$f(\mathbf{x}) = \frac{\sum_{\alpha=1}^{n} c_\alpha G(\|\mathbf{x} - \mathbf{t}_\alpha\|)}{\sum_{\alpha=1}^{m} d_\alpha \sigma(\mathbf{x} \cdot \mathbf{w}_\alpha + \theta_\alpha)}$$

$$\ldots$$

Target spaces

- $\mathcal{T}$ = continuous function with s continuous derivatives in d variables;
- $\mathcal{T}$ = square integrable functions in d variables with s square integrable derivates (Sobolev space);
- $\mathcal{T}$ = integrable functions in d variables with s integrable derivates (Sobolev space);
- $\mathcal{T}$ = band-limited functions;
- $\mathcal{T}$ = functions with integrable Fourier transform;
- Nikol'ski-Besov spaces;
- Lizorkin-Triebel spaces;
- Marcinkiewicz spaces;
- Orlicz spaces;
- weighted spaces of Kipriyanov;

Empirical Risk Minimization

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

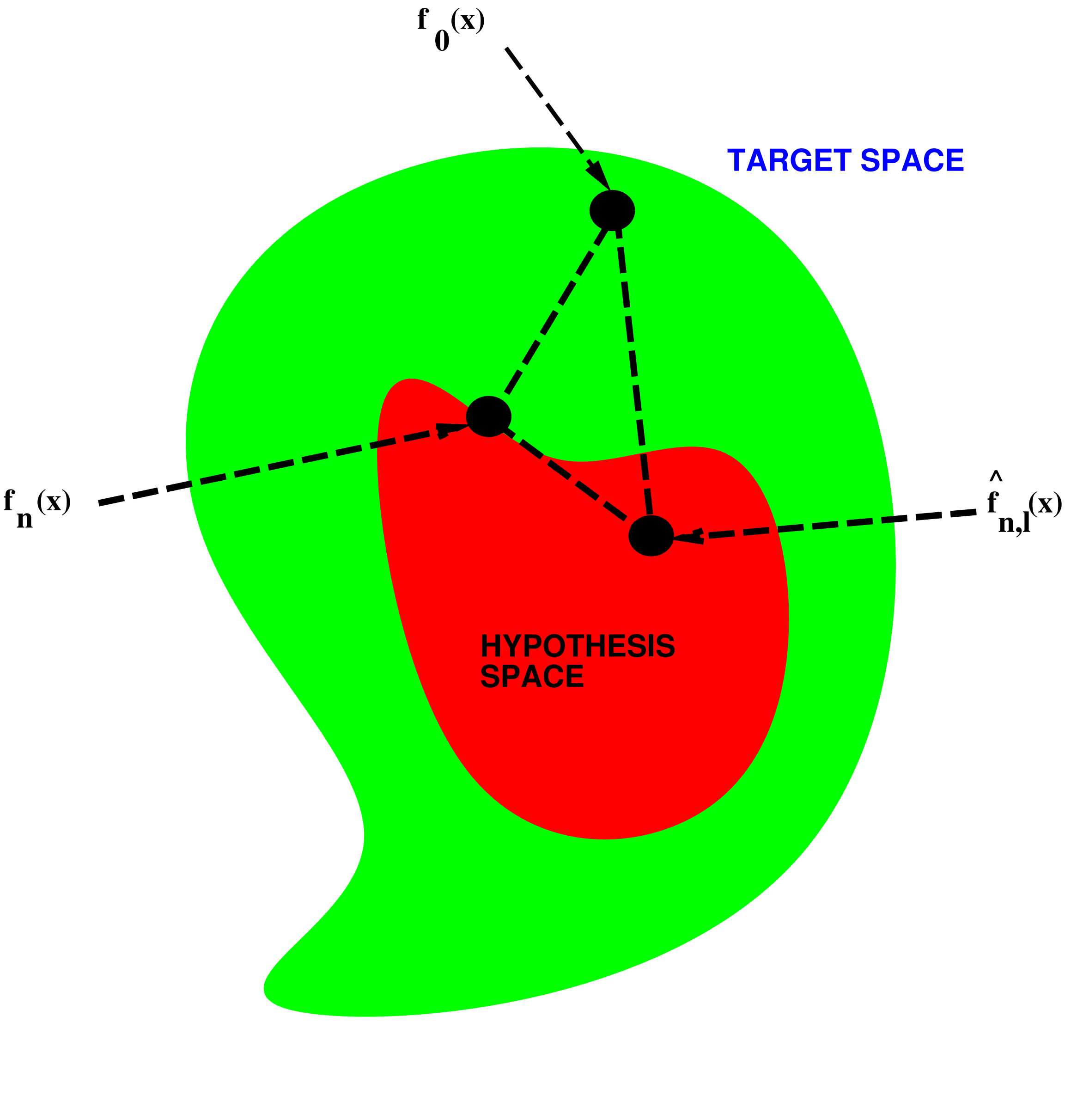
However we can approximate $I[f]$ by the *empirical risk*

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

and find an approximate solution:

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

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



$f_0(x)$
TARGET SPACE
$f_n(x)$
$\hat{f}_{n,l}(x)$
HYPOTHESIS SPACE

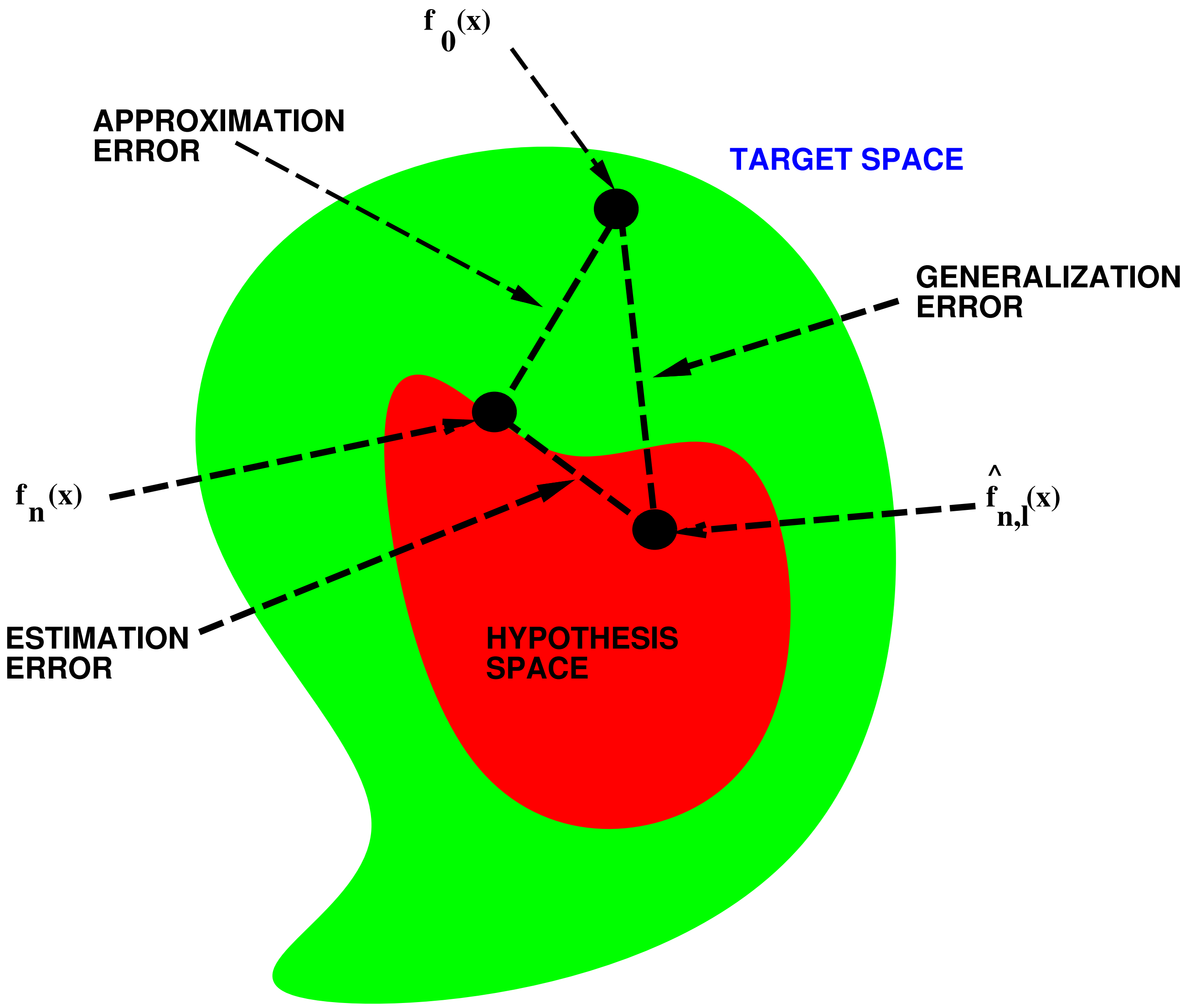
- The best we can possibly do: to find $f_0(\mathbf{x})$

- what we would be happy to do: to find $f_n(\mathbf{x})$

- what we end up doing: to find $\widehat{f}_{n,l}$ (the result of learning)

The goodness of the resulting hypothesis $\widehat{f}_{n,l}$ is measured by

$$I[\widehat{f}_{n,l}] \equiv \textbf{generalization error}$$



$f_0(x)$
APPROXIMATION ERROR
TARGET SPACE
GENERALIZATION ERROR
$f_n(x)$
$\hat{f}_{n,l}(x)$
ESTIMATION ERROR
HYPOTHESIS SPACE

Notations

$E[\cdot]$ stands for the average over $\mathbf{x}$ and y.

$$E[(f_0(\mathbf{x}) - f(\mathbf{x}))^2] = \int_X (f_0(\mathbf{x}) - f(\mathbf{x}))^2 P(\mathbf{x}) d\mathbf{x}$$

$$E[(f_0(\mathbf{x}) - y)^2] = \sigma^2 = \int_{X \times Y} (y - f_0(\mathbf{x}))^2 P(\mathbf{x}, y) d\mathbf{x} dy$$

$$I[f] = E[(f_0(\mathbf{x}) - f(\mathbf{x}))^2] + \sigma^2$$

- The best we can possibly do: to find $f_0(\mathbf{x})$

- what we would be happy to do: to find $f_n(\mathbf{x})$

- what we end up doing: to find $\widehat{f}_{n,l}$

VERY IMPORTANT: to estimate

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

and to prove in what sense

$$\lim_{n,l\to\infty} \widehat{f}_{n,l}(\mathbf{x}) = f_0(\mathbf{x})$$

Bounding the generalization error

$$E[(f_0(\mathbf{x}) - \hat{f}_{n,l}(\mathbf{x}))^2] =$$

$$= E[(f_0(\mathbf{x}) - \hat{f}_{n,l}(\mathbf{x}))^2] - E[(f_0(\mathbf{x}) - f_n(\mathbf{x}))^2] +$$

$$+\ E[(f_0(\mathbf{x}) - f_n(\mathbf{x}))^2]$$

and since

$$I[f] = E[(f_0(\mathbf{x}) - f(\mathbf{x}))^2] + \sigma^2$$

then

$$E[(f_0(\mathbf{x}) - \hat{f}_{n,l}(\mathbf{x}))^2] = \ (I[\hat{f}_{n,l}] - I[f_n])\ +$$

$$+E[(f_0(\mathbf{x}) - f_n(\mathbf{x}))^2]$$

Approximation error

From approximation theory we have bounds of the type:

$$E[(f_0(\mathbf{x}) - f_n(\mathbf{x}))^2] \leq \epsilon(n)$$

where $\epsilon(n)$ depends on:

- the space $\mathcal{T}$ to which the regression function $f_0(\mathbf{x})$ belongs;
- the set H_n

Usually $\bigcup_{n=0}^{\infty} H_n$ is dense in $\mathcal{T}$, and therefore

$$\lim_{n \to \infty} \epsilon(n) = 0$$

Bounding the generalization error

We have seen that

$$E[(f_0(\mathbf{x}) - \hat{f}_{n,l}(\mathbf{x}))^2] = \; (I[\hat{f}_{n,l}] - I[f_n]) + \\ + E[(f_0(\mathbf{x}) - f_n(\mathbf{x}))^2]$$

THEREFORE

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

Estimation error

We do not minimize I, but I_{emp}, and find $\widehat{f}_{n,l}$ rather than f_n.

If we want $\widehat{f}_{n,l}$ to converge to f_n we must impose *uniform convergence in probability*:

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

The theory of uniform convergence in probability (Vapnik, Dudley, Pollard, ...) provides bounds of the form:

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

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

- $I[\widehat{f}_{n,l}] =$ how well we do with n parameters and l data

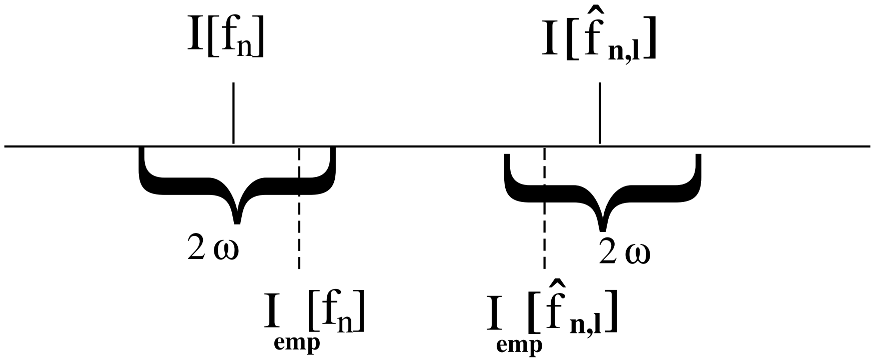
- $I[f_n] =$ how well we can do with n parameters

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

- $I[f_n] \leq I[\widehat{f}_{n,l}]$

- $I_{\text{emp}}[\widehat{f}_{n,l}] \leq I_{\text{emp}}[f_n]$

A useful inequality



If, with probability $1 - \delta$

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

then

$$|I[\widehat{f}_{n,l}] - I[f_n]| \leq 2\Omega(n, l, \delta)$$

A general statement

With probability $1-\delta$:

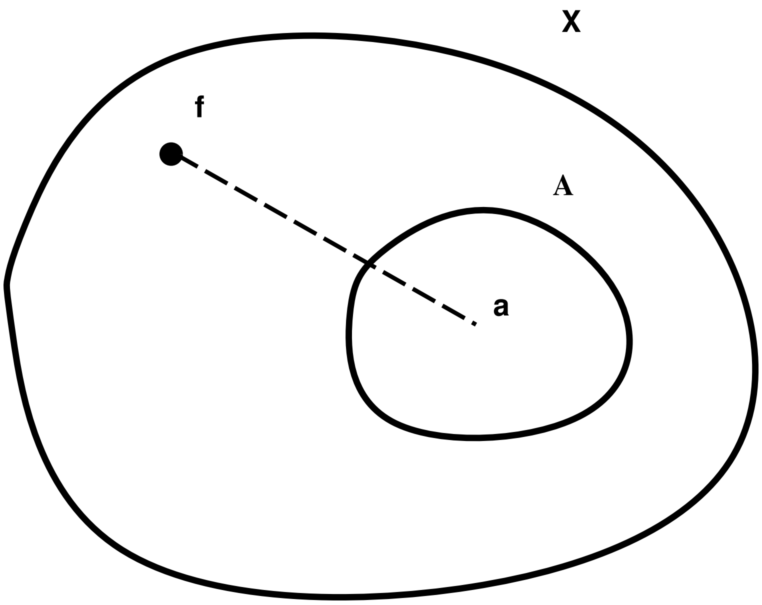
$$E[(f_0(\mathbf{x}) - \widehat{f}_{n,l}(\mathbf{x}))^2] \leq 2\Omega(n,l,\delta) + \epsilon(n)$$

Generalization = estimation +

approximation

- **Approximation error:** it depends critically on the space of function $\mathcal{T}$ that is approximated, and less critically on the approximating class H_n;

- **estimation error:** it does **not** depend critically on the space of function $\mathcal{T}$ that is approximated, but it **does** depend critically on the approximating class H_n



The fundamental objects are

$$\mathcal{T}\ ,\quad \|\cdot\|\ ,\quad H$$

Usually we have

$$H = \{H_n\}_{n=1}^{\infty}$$

where

$$H_1 \subset H_2 \subset \ldots H_n \subset \ldots$$

H is the *approximation scheme*

Examples of the target space $\mathcal{T}$

- $\mathcal{T}$ = continuous function with s continuous derivatives in d variables;

- $\mathcal{T}$ = square integrable functions in d variables with s square integrable derivates;

- $\mathcal{T}$ = integrable functions in d variables with s integrable derivates;

- $\mathcal{T}$ = band-limited functions;

- $\mathcal{T}$ = functions with integrable Fourier transform;

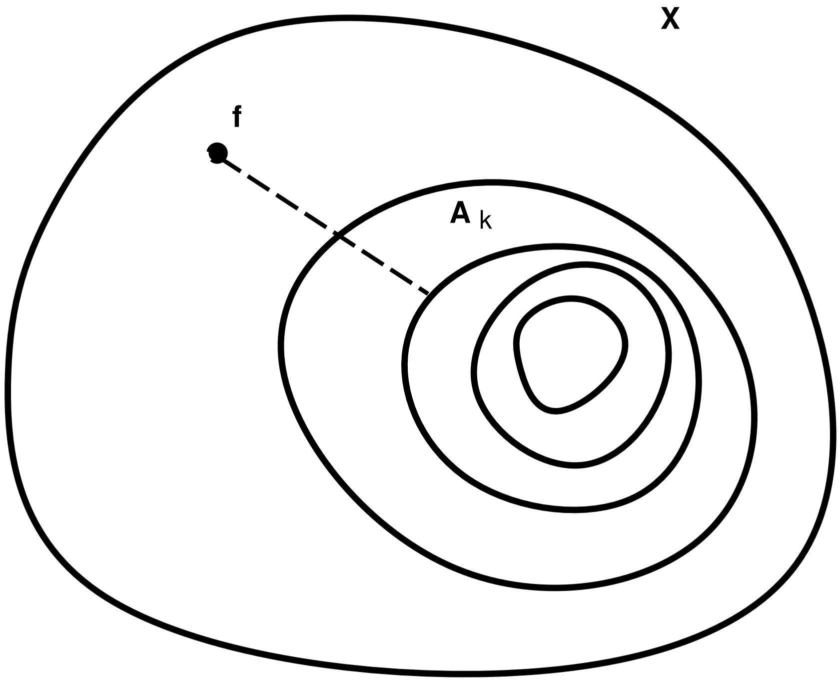
Examples of hypothesis spaces H_n

- H_n = polynomials of degree n in one variable;

- H_n = truncated Fourier series with n terms;

- H_n = Radial Basis Functions with n basis functions;

- H_n = Multilayer Perceptron with n hidden units;



Degree of approximation

$$d_{\mathcal{T}}(f, H_n) \equiv \inf_{h \in H_n} \|f - h\|$$

$d_{\mathcal{T}}(f, H_n)$ is the smallest error that we can do if we approximate $f \in \mathcal{T}$ with an element of H_n.

Degree of approximation

Typically

$$\lim_{n\to\infty} d_{\mathcal{T}}(f, H_n) = 0$$

It means that we can approximate functions of $\mathcal{T}$ arbitrarily well with elements of $\{H_n\}_{n=1}^{\infty}$

For example: $\mathcal{T}$ = continuous functions on compact sets, and H_n = polynomials of degree at most n.

The interesting question is:

How fast does $d_{\mathcal{T}}(f, H_n)$ go to zero?

The rate of convergence to zero is a measure of the relative complexity of $\mathcal{T}$ with respect to the approximation scheme H.

The rate of convergence depends on

$\mathcal{T}$, $\| \cdot \|$, f, and H

and usually has the form

$$d_{\mathcal{T}}(f, H_n) = C_{\mathcal{T}}(f) \left(\frac{1}{n}\right)^{r(\mathcal{T},H)}$$

$\frac{1}{r(\mathcal{T},H)}$ is a measure of the complexity of $\mathcal{T}$ with respect to H.

The curse of dimensionality

Usually

$$r(\mathcal{T}, H) = \frac{c(\mathcal{T}, H)}{d}$$

where d is the dimension. Therefore, if ϵ is the desired error:

$$n \propto \left(\frac{1}{\epsilon}\right)^{\frac{d}{c(\mathcal{T},H)}}$$

Example

Consider the class of functions

$$\Lambda^d_{s,\alpha} \equiv \Lambda^d_{s,\alpha}(M_0, \ldots, M_{s+1})$$

defined on R^d such that

1. Derivatives up to order s exist and are continuous;

2. $\|D^j f\|_{L_\infty} \leq M_j \qquad j = 0, 1, \ldots, s \quad ;$

3. Derivatives of order s belong to $Lip_{M_{s+1}}\alpha\ (R^d)$;

Let $f \in \Lambda^d_{s,\alpha}$, and let P_n the set of polynomials of degree n. Then there exists a constant M such that

$$d_{\Lambda^d_{s,\alpha}}(f, P_n) = M \left(\frac{1}{n}\right)^{\frac{s+\alpha}{d}}$$

It is a general fact that rates of convergence for functions in d dimensions and with a smoothness of order s have the form

$$O\left(\left(\frac{1}{n}\right)^{\frac{s}{d}}\right)$$

Rates of convergence of this form are called of **Jackson type**

Metric Entropy

X a Banach Space

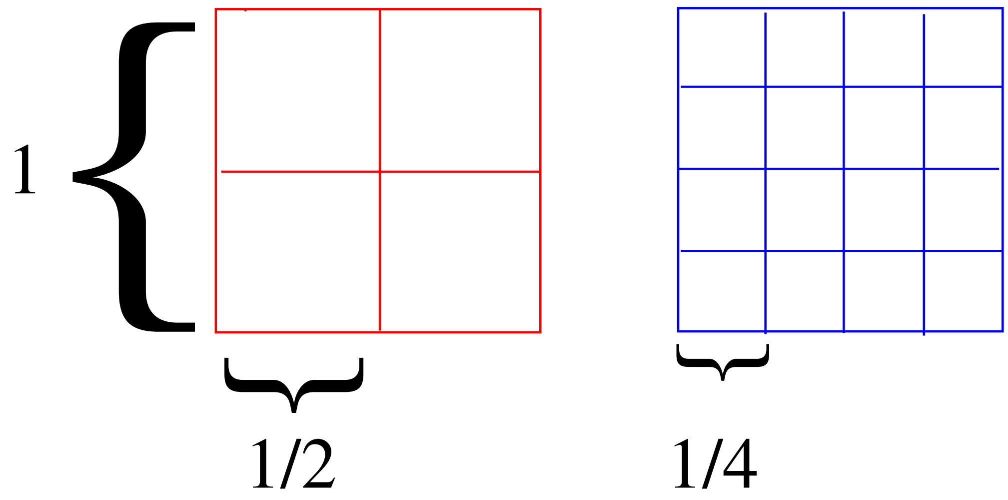
$\phi \subset X$, ϕ compact.

The **metric entropy** $H_\epsilon(\phi)$ is

$$H_\epsilon(\phi) = \log N_\epsilon(\phi)$$

where $\log N_\epsilon(\phi)$ is the minimum number of balls of radius ϵ that cover the set ϕ.

Metric Entropy



$$\epsilon = \frac{1}{k} \Rightarrow N_\epsilon = k^2 = \left(\frac{1}{\epsilon}\right)^2$$

$$H_\epsilon = 2\log\left(\frac{1}{\epsilon}\right)$$

Metric Entropy

$$H_\epsilon(\Lambda^d_{s,\alpha}) \approx \left(\frac{1}{\epsilon}\right)^{\frac{d}{s+\alpha}}$$

It seems that $\frac{d}{s+\alpha}$ is the measure of complexity.

Metric Entropy

How many pages are needed in order to tabulate with an accuracy ϵ a function of a certain set ϕ?

$$\# \text{ of pages} \propto H_\epsilon(\phi)$$

$$\Downarrow$$

How many pages are needed in order to tabulate with an accuracy ϵ a function of d variables with a degree of smoothness s?

$$\# \text{ of pages} \propto \left(\frac{1}{\epsilon}\right)^{\frac{d}{s}}$$

N-widths

Let $\mathcal{T}$ be a Banach space of functions, ψ a subset of $\mathcal{T}$, and H_n a n-dimensional subspace of $\mathcal{T}$, that is a set of functions of the form

$$f = \sum_{i=1}^{n} c_i \phi_i$$

Define the **n-width** of ψ in $\mathcal{T}$ as

$$d_n(\psi) = \inf_{\phi_1, \ldots \phi_n} \sup_{f \in \psi} \inf_{c_1, \ldots c_n} \|f - \sum_{i=1}^{n} c_i \phi_i\|$$

It is a measure of how well a linear technique can approximate a subset ψ of $\mathcal{T}$.

Example

$$d_n(\Lambda^d_{s,\alpha}) \approx \left(\frac{1}{n}\right)^{\frac{s+\alpha}{d}}$$

It means that there is at least on function in $\Lambda^d_{s,\alpha}$that *no linear technique* can approximate at a better rate.

It also can be generalized to **nonlinear techniques**!!!

Linear vs. Nonlinear

- For many natural choices of target spaces, *polynomial* and *rational* approximation have the same rate of convergence;

- For certain functions, piecewise constant approximation with *moving* and *fixed* centers have very different rates of convergence.

Generalized Translation Networks (H. Mhaskar)

Consider networks of the form:

$$f(\mathbf{x}) = \sum_{k=1}^{n} a_k \phi(A_k \mathbf{x} + \mathbf{b}_k)$$

where $\mathbf{x} \in R^d$, $\mathbf{b}_k \in R^m$, $1 \leq m \leq d$, A_k are $m \times d$ matrices, $a_k \in R$ and ϕ is some given function.

For $m = 1$ this is a Multilayer Perceptron .

For $m = d$, A_k diagonal and ϕ radial this is a Radial Basis Functions network.

Theorem (Mhaskar, 1994)

Let $W_r^p(R^d)$ be the space of functions whose derivatives up to order r are p-integrable in R^d. Under very general assumptions on ϕ one can prove that there exists $d \times m$ matrices $\{A_k\}_{k=1}^n$ such that, for any $f \in W_r^p(R^d)$, one can find $\mathbf{b}_k$ and a_k such that:

$$\|f - \sum_{k=1}^{n} a_k \phi(A_k \mathbf{x} + \mathbf{b}_k)\|_p \leq c n^{-\frac{r}{d}} \|f\|_{W_r^p}$$

Moreover, the coefficients a_k are linear functionals of f.

This rate is optimal.

Classes of functions in d dimensions with smoothness of order s have an *intrinsic complexity* characterized by the ratio $\frac{s}{d}$:

- the curse of dimensionality is the d factor;

- the blessing of smoothness is the s factor;

We cannot expect to find an approximation technique that "beats the curse of dimensionality", *unless we let the smoothness s change with the dimension d.*

Theorem (Barron, 1991)

Let f be a function such that its Fourier transform satisfies

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

and let Ω be a bounded domain in R^d. Then we can find a neural network with n coefficients, n weights and n biases such that

$$\|f - \sum_{i=1}^{n} c_i \sigma(\mathbf{x} \cdot \mathbf{w}_i + \theta_i)\|^2_{L_2(\Omega)} < \frac{c}{n}$$

The rate of convergence is **independent of the dimension**.

The space of functions such that

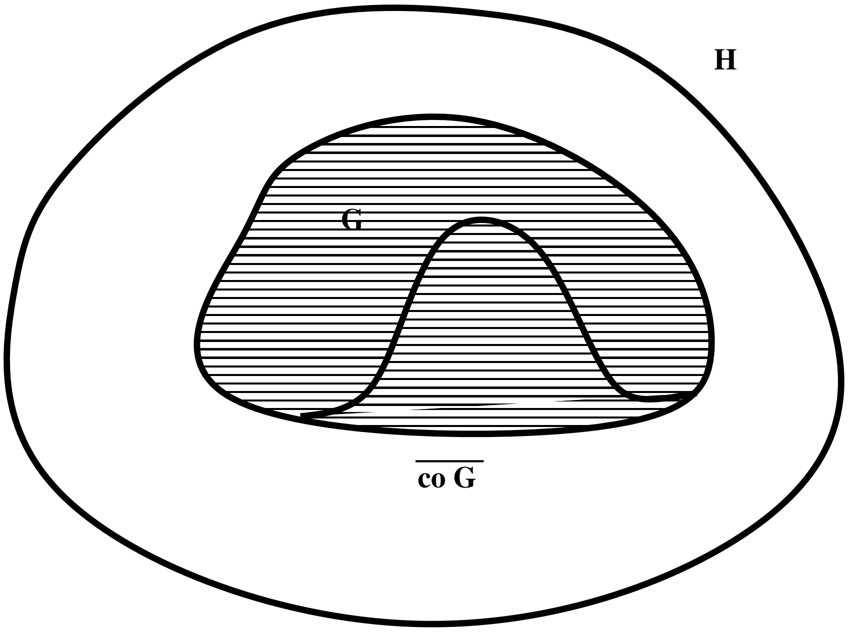
$$\int_{R^d} d\mathbf{s} \; \|\mathbf{s}\| |\tilde{f}(\mathbf{s})| < +\infty \; .$$

is the space of functions that can be written as

$$f = \frac{1}{\|\mathbf{x}\|^{d-1}} * \lambda$$

where λ is any function whose Fourier Transform is integrable.

Notice how the space becomes more constrained as the dimension increases.



Let H be an Hilbert space and $G \subset H$ such that $\|g\| \leq b \ \ \forall g \in G$. Let $f \in \overline{\text{coG}}$.

Then $\forall c > b^2 - \|f\|^2$ we can find n numbers $\alpha_k > 0$, such that $\sum_{i=1}^{n} \alpha_k = 1$, and n elements $g_k \in G$ such that

$$\|f - \sum_{k=1}^{n} \alpha_k g_k\|^2 \leq \frac{c}{n}$$

(Maurey, 1981; Jones, 1990; Barron, 1991)

Let $f_n = \sum_{k=1}^{n} \alpha_k g_k$ be the approximating function in Jones' Lemma. The sequence $\{f_n\}_{n=0}^{\infty}$ has the following structure:

$$f_{n+1} = \alpha_n f_n + (1 - \alpha_n) g_n$$

where $\alpha_n \in I \equiv [0, 1]$ and g_n "approximately solve" the following minimization problem:

$$\inf_{\alpha_n \in I, g_n \in \mathcal{G}} \|f - \alpha_n f_n - (1 - \alpha_n) g_n\|$$

"approximately solve" means that it is sufficient at each step to reach a distance from the infimum of order $O(\frac{1}{n^2})$.

Example

$$G = \{c\sigma(\mathbf{x} \cdot \mathbf{w} + \theta) \mid |c| \leq b, \theta \in R, \mathbf{w} \in R^d\}$$

The target function f should belong to Barron's space (that is $\overline{\text{coG}}$). Then, starting with $f_0 = 0$, we have

$$f_1(\mathbf{x}) = (1 - \alpha_0)c_0\sigma(\mathbf{x} \cdot \mathbf{w}_0 + \theta_0)$$

and we have to solve

$$\min_{\alpha_0, |c_0| \leq b, \theta_0, \mathbf{w}_0} \|f(\mathbf{x}) - (1 - \alpha_0)c_0\sigma(\mathbf{x} \cdot \mathbf{w}_0 + \theta_0)\|$$

Next step

We define

$$f_2(\mathbf{x}) = \alpha_1 f_1(\mathbf{x}) + (1 - \alpha_1) c_1 \sigma(\mathbf{x} \cdot \mathbf{w}_1 + \theta_1)$$

and we have to solve

$$\min_{\alpha_1, |c_1| \le b, \theta_1, \mathbf{w}_1} \| f(\mathbf{x}) - \alpha_1 f_1(\mathbf{x}) - (1 - \alpha_1) c_1 \sigma(\mathbf{x} \cdot \mathbf{w}_1 + \theta_1) \|$$

and so on ...

Jones' lemma: (my) instructions for use

"Whenever" a function f admits an integral representation as

$$f(\mathbf{x}) = \int_{R^n} d\mu(\mathbf{t})\ G(\mathbf{x};\mathbf{t})$$

where $\mu(\mathbf{t})$ is a signed measure and $G(\mathbf{x};\mathbf{t})$ is some parametric function, then f can be approximated as

$$f^*(\mathbf{x}) = \sum_{\alpha=1}^{n} c_\alpha G(\mathbf{x};\mathbf{t}_\alpha)$$

and the approximation error is bounded by $O(\frac{1}{\sqrt{n}})$.

A version of Jones' lemma*

Let $G(\mathbf{x};\mathbf{t})$ be a function defined over $R^k \times R^d$ such that $\|G(\mathbf{x};\mathbf{t})\|_{L_2} \leq b$ for some $b > 0$ and for all $\mathbf{t} \in R^d$. Let f be a function defined over R^k, such that

$$f(\mathbf{x}) = \int_{R^d} d\mathbf{t}\ \mu(\mathbf{t}) G(\mathbf{x};\mathbf{t})$$

where $\|\mu(\mathbf{t})\|_{L_1} = 1$. Then it is possible to find n Boolean parameters $\{c_i\}_{i=1}^n$ with values in $\{-1, 1\}$ and n vectors $\{\mathbf{t}_i\}_{i=1}^n$ such that:

$$\left\| f - \frac{1}{n}\sum_{i=1}^{n} c_i G(\mathbf{x};\mathbf{t}_i) \right\|_{L_2(P)}^2 \leq \frac{b^2 - \|f\|_{L_2}^2}{n}$$

*(Girosi, 1996)

Proof

We start with μ being positive. Notice that, for an arbitrary choice of the vectors $\{\mathbf{t}_i\}_{i=1}^n$, the following equality holds:

$$\left\| f - \frac{1}{n}\sum_{i=1}^n G(\mathbf{x};\mathbf{t}_i) \right\|_{L_2}^2 = \|f\|_{L_2}^2 +$$

$$+\frac{1}{n^2}\sum_{i=1}^n \sum_{j=1}^n \int d\mathbf{x}\ G(\mathbf{x};\mathbf{t}_i)G(\mathbf{x};\mathbf{t}_j) +$$

$$-\frac{2}{n}\sum_{i=1}^n \int d\mathbf{x}\ f(\mathbf{x})G(\mathbf{x};\mathbf{t}_i)$$

$$\left\| f - \frac{1}{n} \sum_{i=1}^{n} G(\mathbf{x}; \mathbf{t}_i) \right\|_{L_2}^2 = \|f\|_{L_2}^2 +$$

$$+ \frac{1}{n^2} \sum_{i=1}^{n} \int d\mathbf{x} \ G^2(\mathbf{x}; \mathbf{t}_i) +$$

$$+ \frac{1}{n^2} \sum_{i \neq j}^{n} \int d\mathbf{x} \ G(\mathbf{x}; \mathbf{t}_i) G(\mathbf{x}; \mathbf{t}_j) +$$

$$- \frac{2}{n} \sum_{i=1}^{n} \int d\mathbf{x} \ f(\mathbf{x}) G(\mathbf{x}; \mathbf{t}_i)$$

We now consider the variables $\{\mathbf{t}_i\}_{i=1}^n$ as random variables, that we assume identically distributed with probability distribution $\mu(\mathbf{t})$, and define the following expected value

$$E[\cdot] = \int d\mathbf{t}_1 \ \mu(\mathbf{t}_1) \ \ldots \int d\mathbf{t}_n \ \mu(\mathbf{t}_n)$$

Taking the expected value of the equality above, exchanging the order of integration, and using the integral representation of f we obtain

$$E\left[\left\| f - \frac{1}{n}\sum_{i=1}^{n} G(\mathbf{x};\mathbf{t}_i)\right\|_{L_2}^2\right] = \|f\|_{L_2}^2 +$$

$$+\frac{1}{n}\int d\mathbf{x}d\mathbf{t}\ \mu(\mathbf{t})G^2(\mathbf{x};\mathbf{t}) +$$

$$+\frac{1}{n^2}n(n-1)\|f\|_{L_2}^2 - \frac{2}{n}n\|f\|_{L_2}^2 = \frac{1}{n}(A^2 - \|f\|_{L_2}^2)$$

where we have defined

$$A^2 = \int d\mathbf{x}d\mathbf{t}\ \mu(\mathbf{t})G^2(\mathbf{x};\mathbf{t}) \leq b^2$$

Therefore there exists at least a set of vectors $\{\mathbf{t}_i\}_{i=1}^n$ such that

$$\left\| f - \frac{1}{n}\sum_{i=1}^{n} G(\mathbf{x};\mathbf{t}_i) \right\|_{L_2}^2 \leq \frac{A^2 - \|f\|_{L_2}^2}{n} \leq \frac{b^2 - \|f\|_{L_2}^2}{n}$$

If the function μ is not positive we can always write

$$\mu(\mathbf{t}) = sign(\mu(\mathbf{t}))|\mu(\mathbf{t})|$$

and therefore substitute $G(\mathbf{x};\mathbf{t}_i)$ with $sign(\mu(\mathbf{t}_i))G(\mathbf{x};\mathbf{t}_i)$. Denoting $sign(\mu(\mathbf{t}_i))$ by c_i the theorem follows $\square$.

Theorem (Girosi and Anzellotti, 1992)

Let $f \in H^{m,1}(R^d)$, where $H^{m,1}(R^d)$ is the space of functions whose partial derivatives up to order m are integrable, and let $G_m(\mathbf{x})$ be the Bessel-Macdonald kernel, that is the Fourier transform of

$$\tilde{G}_m(\mathbf{s}) = \frac{1}{(1+\|\mathbf{s}\|^2)^{\frac{m}{2}}} \qquad m > 0 \ .$$

If m > d and m is even, we can find a Radial Basis Functions network with n coefficients c_α and n centers $\mathbf{t}_\alpha$ such that

$$\|f - \sum_{\alpha=1}^{n} c_\alpha G_m(\mathbf{x} - \mathbf{t}_\alpha)\|^2_{L_\infty} < \frac{c}{n}$$

Theorem (Girosi, 1992)

Let $f \in H^{m,1}(R^d)$, where $H^{m,1}(R^d)$ is the space of functions whose partial derivatives up to order m are integrable. **If $\mathbf{m} > \mathbf{d}$** and m is even, we can find a Gaussian basis function network with n coefficients c_α, n centers $\mathbf{t}_\alpha$ and n variances σ_α such that

$$\|f - \sum_{\alpha=1}^{n} c_\alpha e^{-\frac{(\mathbf{x}-\mathbf{t}_\alpha)^2}{2\sigma_\alpha^2}}\|^2_{L_\infty} < \frac{c}{n}$$

Same rate of convergence: $O(\frac{1}{\sqrt{n}})$

Function space	Norm	Approximation scheme
$\int_{R^d} \; d\mathbf{s} \; \lvert\tilde{f}(\mathbf{s})\rvert < +\infty$	$L_2(\Omega)$	$f(\mathbf{x}) = \sum_{i=1}^n c_i \sin(\mathbf{x}\cdot\mathbf{w}_i + \theta_i)$
(Jones)		
$\int_{R^d} \; d\mathbf{s} \; \lVert\mathbf{s}\rVert\lvert\tilde{f}(\mathbf{s})\rvert < +\infty$	$L_2(\Omega)$	$f(\mathbf{x}) = \sum_{i=1}^n c_i \sigma(\mathbf{x}\cdot\mathbf{w}_i + \theta_i)$
(Barron)		
$\int_{R^d} \; d\mathbf{s} \; \lVert\mathbf{s}\rVert^2\lvert\tilde{f}(\mathbf{s})\rvert < +\infty$	$L_2(\Omega)$	$f(\mathbf{x}) = \sum_{i=1}^n c_i \lvert\mathbf{x}\cdot\mathbf{w}_i + \theta_i\rvert_+ +$
(Breiman)		$+\mathbf{a}\cdot\mathbf{x} + b$
$\tilde{f}(\mathbf{s}) \in C_0^k, \; 2k > d$	$L_\infty(R^d)$	$f(\mathbf{x}) = \sum_{\alpha=1}^n c_\alpha e^{-\lVert\mathbf{x}-\mathbf{t}_\alpha\rVert^2}$
(Girosi and Anzellotti)		
$H^{2m,1}(R^d), \; 2m > d$	$L_\infty(R^d)$	$f(\mathbf{x}) = \sum_{\alpha=1}^n c_\alpha e^{-\frac{\lVert\mathbf{x}-\mathbf{t}_\alpha\rVert^2}{\sigma_\alpha^2}}$
(Girosi)		