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, related by a probabilistic relationship:

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

Let $D = \{(\mathbf{x}_i, y_i) \in X \times Y\}_{i=1}^{l}$ a set of examples drawn from $P(\mathbf{x}, y)$.

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

The conditional probability $P(y|\mathbf{x})$ describes a *probabilistic* relationship between the input variable $\mathbf{x}$ and the output variable y: the same input $\mathbf{x}$ can generate different outputs y.

Examples of $P(y|\mathbf{x})$:

- $P(y|\mathbf{x}) \propto e^{-\beta(y-h(\mathbf{x}))^2}$: sampling a function in presence of noise;

- $P(y|\mathbf{x}) = \delta(y - h(\mathbf{x}))$: sampling a function in absence of noise.

$P(\mathbf{x}, y)$	unknown
$\Downarrow$	*sampling*
$D = \{(\mathbf{x}_i, y_i)\}_{i=1}^{l}$	data set
$\Downarrow$	*learning*
$(\mathbf{x}, f(\mathbf{x}))$	$\mathbf{x} \notin D$

The function f is the result of learning, and it is called an *estimate*

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

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

where $E[\cdot]$ stands for the mathematical expectation.

In this framework *learning* means being able to find an estimate f^* that minimizes the expected risk:

$$f^*(\mathbf{x}) = \arg\min_{f \in \mathcal{F}} I[f]$$

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

Notice that

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

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*.

Notations

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

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

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

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

Therefore

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

σ^2 is an intrinsic limitation, that measures the spread of the conditional probability $P(y|\mathbf{x})$.

The best possible estimator is the regression function $f_0(\mathbf{x})$, that we assume to belong to some given class of functions A.

From now on the class $\mathcal{F}$ will be a class of functions H_n parametrized by a number of parameters proportional to n. Examples:

- H_n = class of one dimensional polynomials of degree n;

- $$H_n = \{f | f(\mathbf{x}) = \sum_{i=1}^{n} c_i G(\|\mathbf{x} - \mathbf{t}_i\|) \ , c_i \in R \ , \ \mathbf{t}_i \in R^d\}$$

- $$H_n = \{f | f(\mathbf{x}) = \sum_{i=1}^{n} c_i \sigma(\mathbf{x} \cdot \mathbf{w}_i + \theta_i) \ , c_i \in R, \ \theta_i \in R, \ \mathbf{w}_i \in R^d\}$$

$P(\mathbf{x}, y)$ 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 technique*.

- 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 $\hat{f}_{n,l}$

VERY IMPORTANT: to estimate

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

and to prove in what sense

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

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

Usually $\bigcup_{n=0}^{\infty} H_n$ is dense in A, 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:

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

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

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

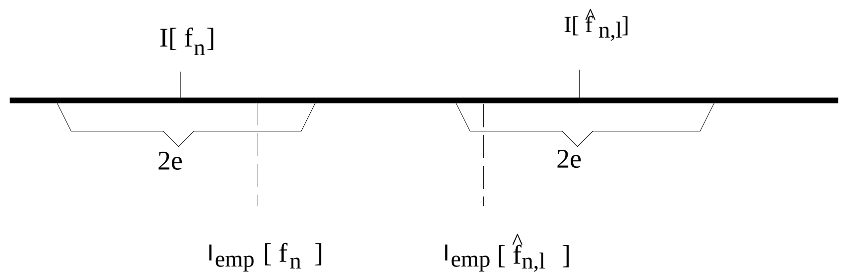
- $I[f_n] =$ how well we can do with n parameters

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

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

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

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 $1-\delta$:

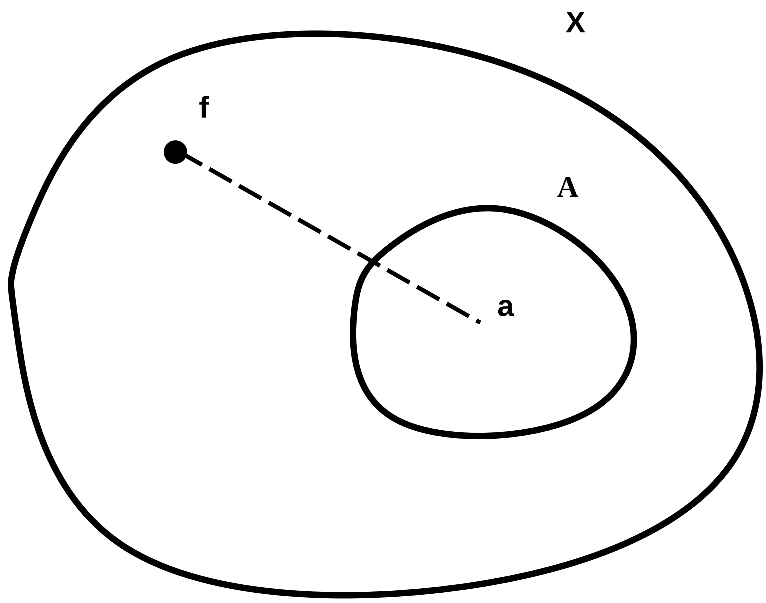
$$E[(f_0(\mathbf{x}) - \widehat{f}_{n,l}(\mathbf{x}))^2] \leq 2\Omega(n,l,\delta) + \epsilon(n)$$

Generalization = estimation +

approximation

- **Approximation error:** it depends critically on the space of function A that is approximated, and less critically on the approximating class H_n;

- **estimation error:** it does **not** depend critically on the space of function A that is approximated, but it **does** depend critically on the approximating class H_n



The fundamental objects are

$$X \ , \quad \|\cdot\| \ , \quad A$$

Usually we have

$$A = \{A_n\}_{n=1}^{\infty}$$

where

$$A_1 \subset A_2 \subset \ldots A_n \subset \ldots$$

A is the *approximation scheme*

Examples of the space X (functions to be approximated)

- $X =$ continuous function with s continuous derivatives in d variables;

- $X =$ square integrable functions in d variables with s square integrable derivates;

- $X =$ integrable functions in d variables with s integrable derivates;

- $X =$ band-limited functions;

- $X =$ functions with integrable Fourier transform;

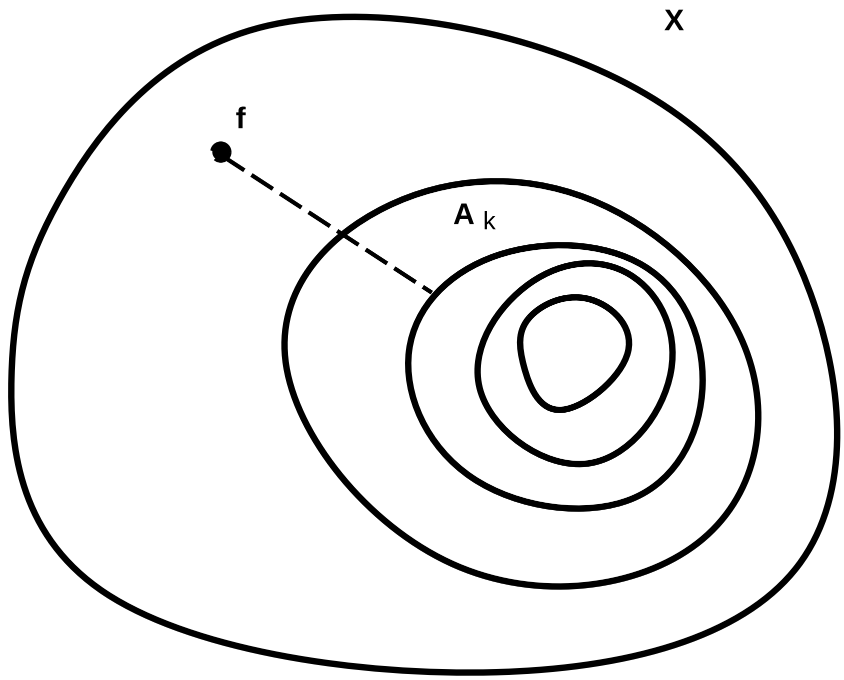
Examples of the space A_n (approximating functions)

- $A_n =$ polynomials of degree n in one variable;

- $A_n =$ truncated Fourier series with n terms;

- $A_n =$ Radial Basis Functions with n basis functions;

- $A_n =$ Multilayer Perceptron with n hidden units;



Degree of approximation

$$d_X(f, A_k) \equiv \inf_{a \in A_k} \|f - a\|$$

$d_X(f, A_k)$ is the smallest error that we can do if we approximate $f \in X$ with an element of A_k.

Degree of approximation

Typically

$$\lim_{n\to\infty} d_X(f, A_n) = 0$$

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

For example: $X =$ continuous functions on compact sets, and $A_n =$ polynomials of degree at most n.

The interesting question is:

How fast does $d_X(f, A_n)$ go to zero?

The rate of convergence to zero is a measure of the relative complexity of X with respect to the approximation scheme A.

The rate of convergence depends on

X, $\|\cdot\|$, f, and A

and usually has the form

$$d_X(f, A_n) = C_X(f)\left(\frac{1}{n}\right)^{r(X,A)}$$

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

The curse of dimensionality

Usually

$$r(X, A) = \frac{c(X, A)}{d}$$

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

$$n \propto \left(\frac{1}{\epsilon}\right)^{\frac{d}{c(X,A)}}$$

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**

N-widths

Let X be a Banach space of functions, ψ a subset of X, and X_n a n-dimensional subspace of X, 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 X as

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

It is a measure of how well a linear technique can approximate a subset ψ of X.

Example

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

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

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

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.

Approximation error

The approximation error depends mainly on the complexity of class of functions that is approximated, that is from the *properties of the phenomenon being studied*.

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

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

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

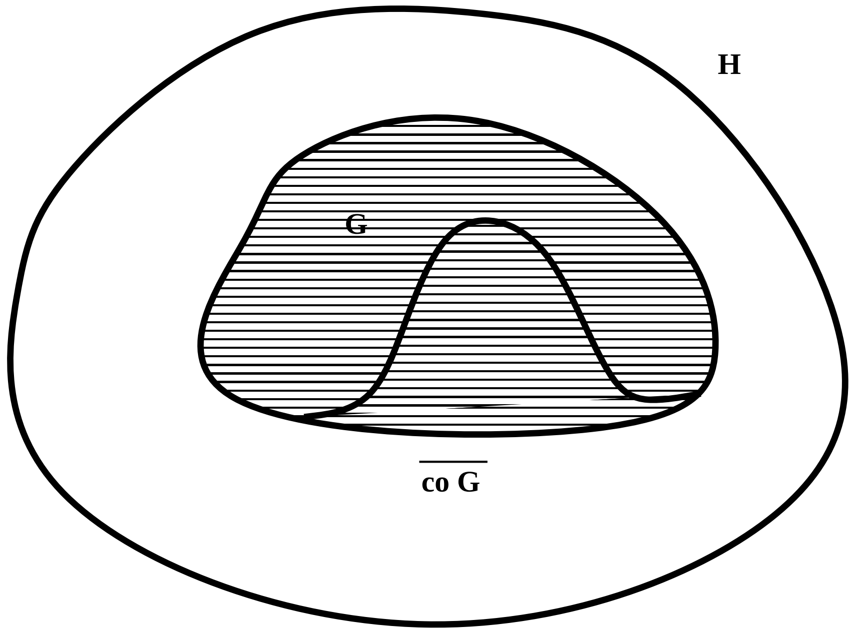
$$\int_{R^d} d\mathbf{s}\ \|\mathbf{s}\||\tilde{f}(\mathbf{s})| < +\infty\ .$$

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 $\|g\| \leq b \quad \forall g \in G$. 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

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

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

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

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

and we have to solve

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

Next step

We define

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

and we have to solve

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

and so on ...

Jones' lemma: (my) instructions for use

"Whenever" we have a function that 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 it can approximated as

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

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

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 m** $>$ **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

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

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

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