Class 1

Introduction to Statistical Learning Theory

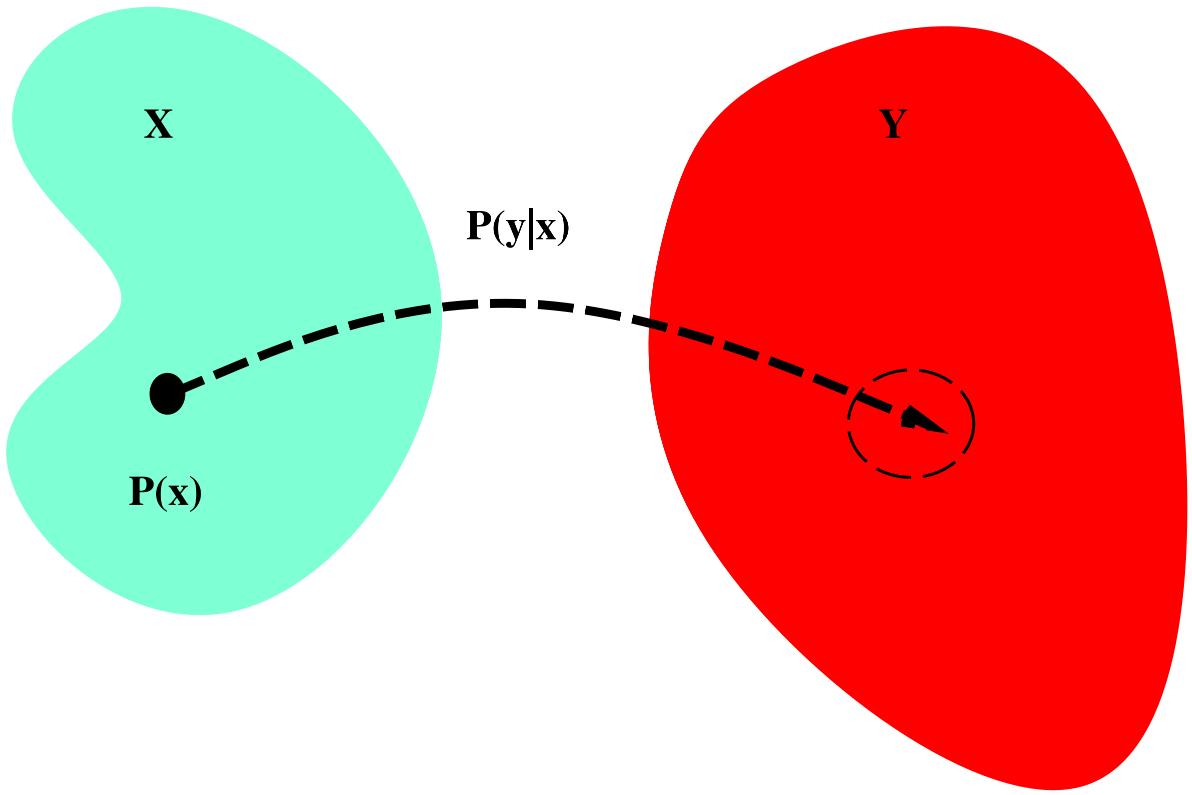
Federico Girosi

Readings for Class 1

Estimation of Dependences Based on Empirical Data. V. Vapnik. Springer-Verlag, 1982. **Chapter 1, 2**.

The Nature of Statistical Learning Theory. V. Vapnik. Springer, 1995. **Chapter 1**.

Learning from Examples



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

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

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

There are different reasons for which the same input $\mathbf{x}$ can generate different outputs y:

- the underlying process is deterministic, but there is noise in the measurement of y;
- the underlying process is not deterministic;
- the underlying process is deterministic, but incomplete information is available.

The underlying process is deterministic but there is noise in the measurement of y

Suppose $h(\mathbf{x})$ is the temperature of the air at the location $\mathbf{x}$ on the surface of the earth. Under the assumption of gaussian noise the quantity that we measure is

$$y = h(\mathbf{x}) + \epsilon$$

where ϵ has a Gaussian distribution with variance σ^2. Then

$$P(y|\mathbf{x}) \propto e^{-\frac{(y-h(\mathbf{x}))^2}{\sigma^2}}$$

The underlying process is deterministic, but incomplete information is available

Suppose y is uniquely determined by x_1 and x_2, so that

$$y = f(x_1, x_2)$$

Then

$$P(y, x_1, x_2) = \delta(y - f(x_1, x_2))P(x_1, x_2)$$

Suppose that we only observe x_1. The joint probability distribution of y and x_1 is

$$P(y, x_1) = \int dx_2 \delta(y - f(x_1, x_2))P(x_1, x_2)$$

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*

The expected risk

Other measures of error could be used:

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

where V is some other monotonically increasing function (for example $V(x) = |x|$). This leads to the following expected risk:

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

Example

In the simple case in which

$$P(y|\mathbf{x}) = \delta(y - h(\mathbf{x}))$$

and $V(x) = x^2$ the expected risk is

$$I[f] = \int_X (h(\mathbf{x}) - f(\mathbf{x}))^2 P(\mathbf{x}) d\mathbf{x} \equiv \|h - f\|^2_{L_2(P)}$$

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

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

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

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;

From now on we will always consider classes of hypothesis spaces of increasing complexity, that we will denote by H_n:

$$H_{n_1} \subset H_{n_2} \subset \ldots \subset H_{n_k} \quad if \quad n_1 \leq n_2 \leq \ldots \leq n_k$$

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

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

- $$H_n = \{f \in L_2 \mid f' \in L_2, \quad \|f'\|^2 \leq n\}$$

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

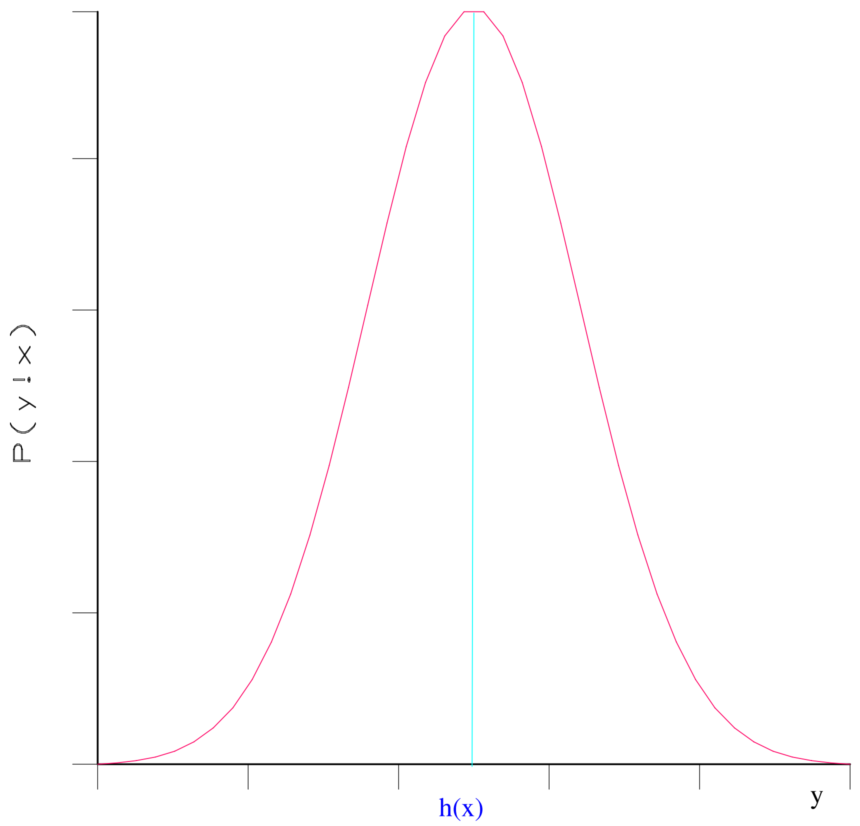
$$f_0(\mathbf{x}) = \int_Y y P(y|\mathbf{x}) dy$$

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

Example

If $P(y|\mathbf{x}) = G(y - h(\mathbf{x}))$ and G is an even function, then the regression function is:

$$f_0(x) = \int_Y y\ G(y - h(\mathbf{x}))dy = h(\mathbf{x})\ .$$



The regression function for a deterministic process and incomplete information

Suppose y is uniquely determined by x_1 and x_2:

$$y = f(x_1, x_2) \ \rightarrow P(y|x_1, x_2) = \delta(y - f(x_1, x_2))$$

but we only observe x_1. Assuming for simplicity that $P(x_1, x_2) = P(x_1)P(x_2)$,

$$f_0(x_1) = \int yP(y|x_1)\, dy = \int f(x_1, x_2)P(x_2)dx_2$$

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.

Examples of target spaces $\mathcal{T}$

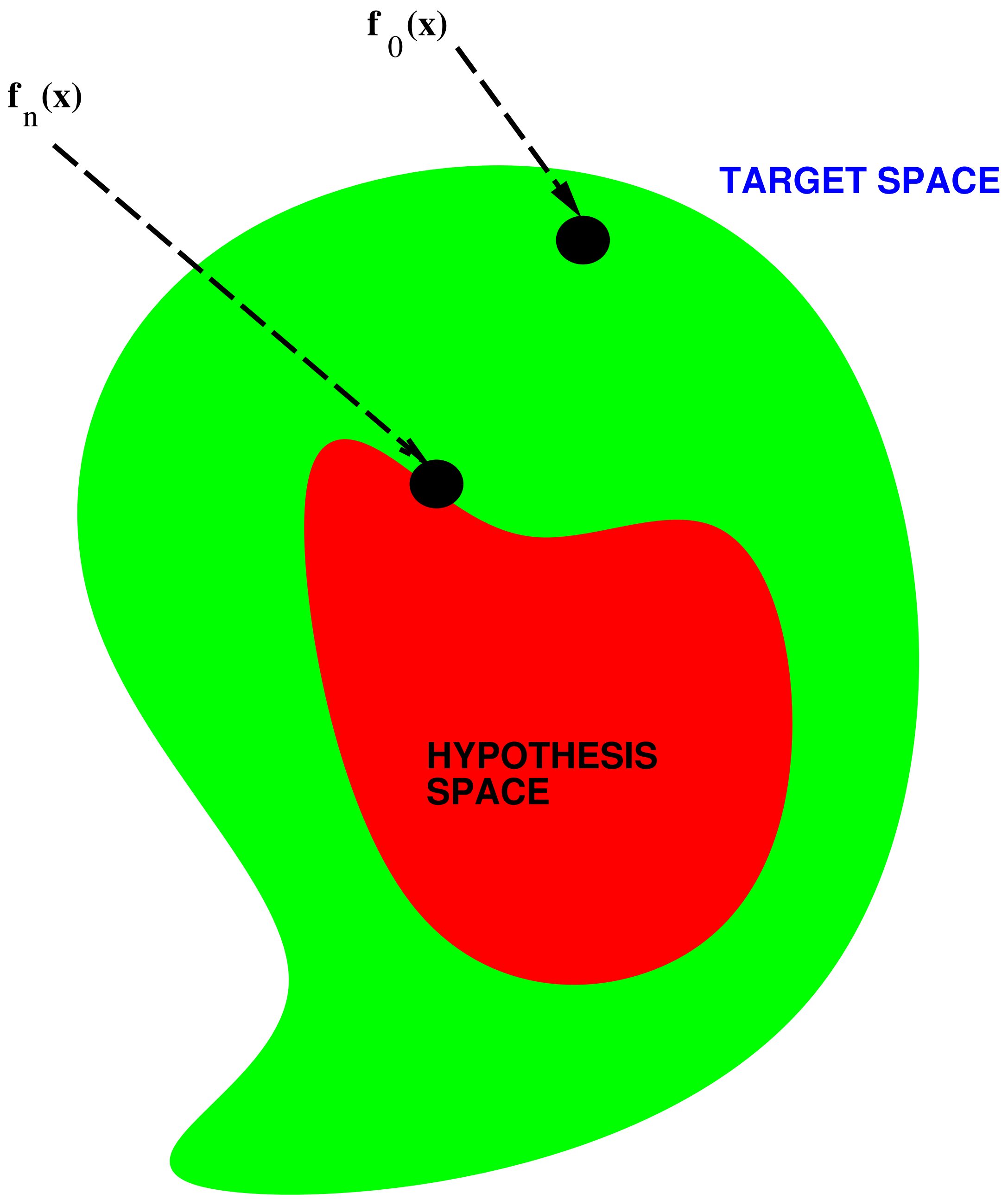
- $\mathcal{T}$ = continuous function with s continuous derivatives in d variables;

- $\mathcal{T}$ = 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;



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

Empirical Risk Minimization

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

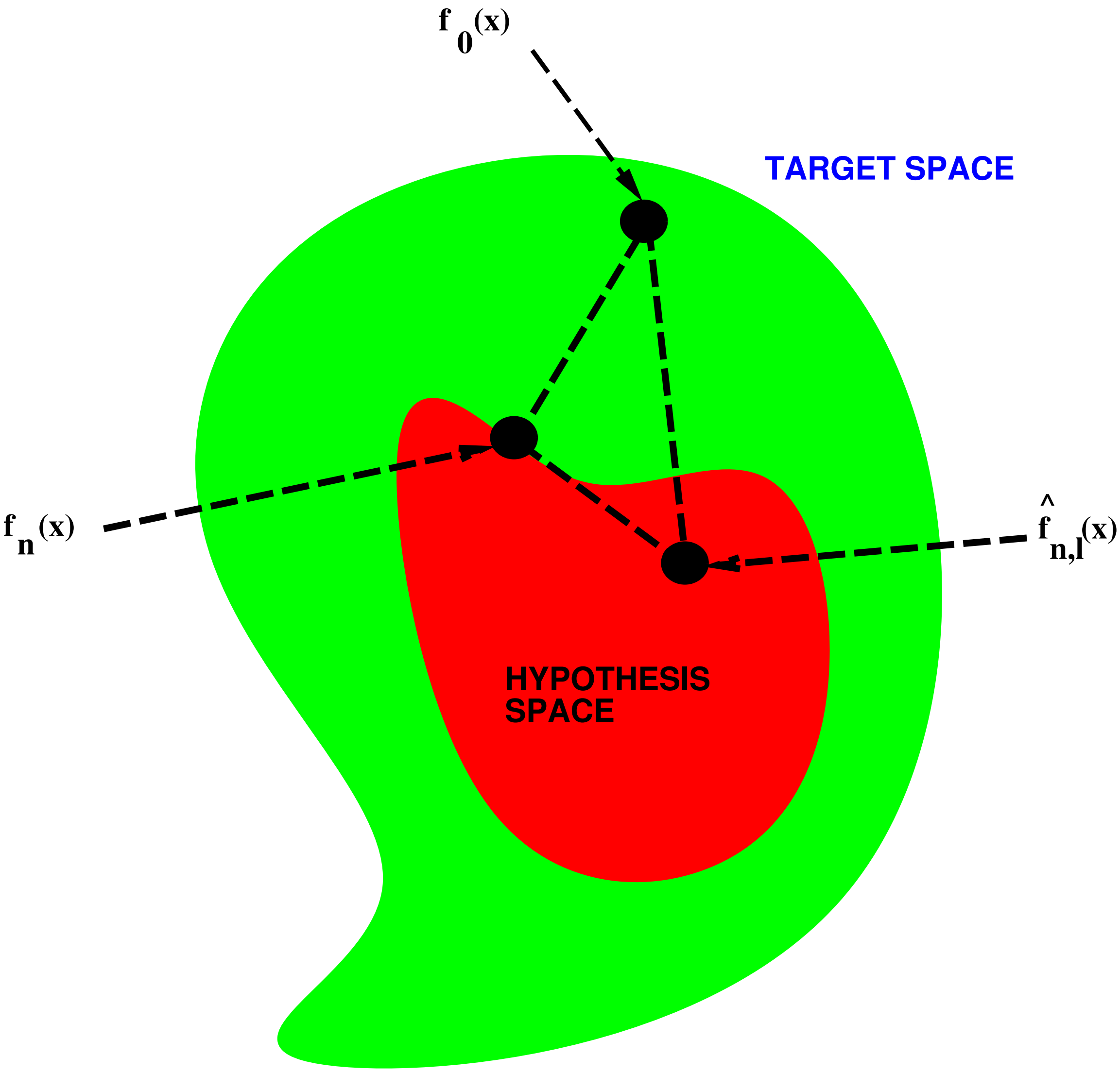
However we can approximate $I[f]$ by the *empirical risk*

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

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*.



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

Empirical risk minimization

We know that, in probability, the empirical risk converges to the expected risk:

$$\lim_{l \to \infty} I_{\text{emp}}[f] = I[f]$$

However, it is not obvious that

$$\lim_{l \to \infty} \min_{f \in H_n} I_{\text{emp}}[f] = \min_{f \in H_n} I[f]$$

This will happen only if H_n is “small” enough (in a sense to be specified later).

Example: Radial Basis Functions

Consider the following hypothesis space:

$$H_n = \{g | g(\mathbf{x}) = \sum_{j=1}^{l} c_j G(\|\mathbf{x} - \mathbf{x}_j\|)\}$$

where G is a Gaussian function.

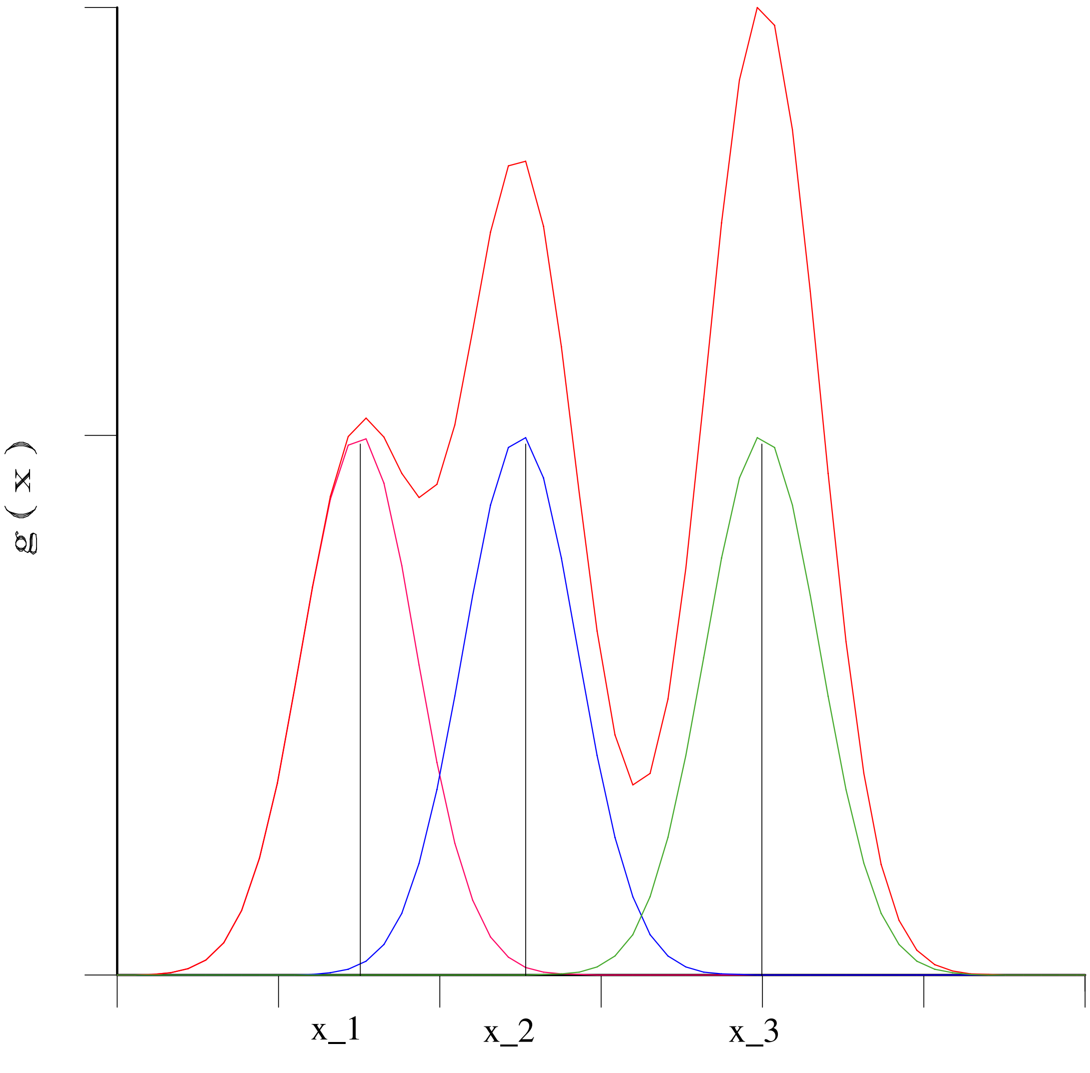
$$\widehat{f}_{n,l}(\mathbf{x}) = \arg\min_{f \in H_n} I_{\text{emp}}[f] = \sum_{j=1}^{l} c_j^* G(\|\mathbf{x} - \mathbf{x}_j\|)$$

$$\Downarrow$$

$$c_1^*, \ldots, c_l^* = \arg\min_{c_1,\ldots,c_l} \sum_{i=1}^{l} \left(y_i - \sum_{j=1}^{l} c_j G(\|\mathbf{x}_i - \mathbf{x}_j\|) \right)^2$$

This approximation scheme is called **Radial Basis Functions**.

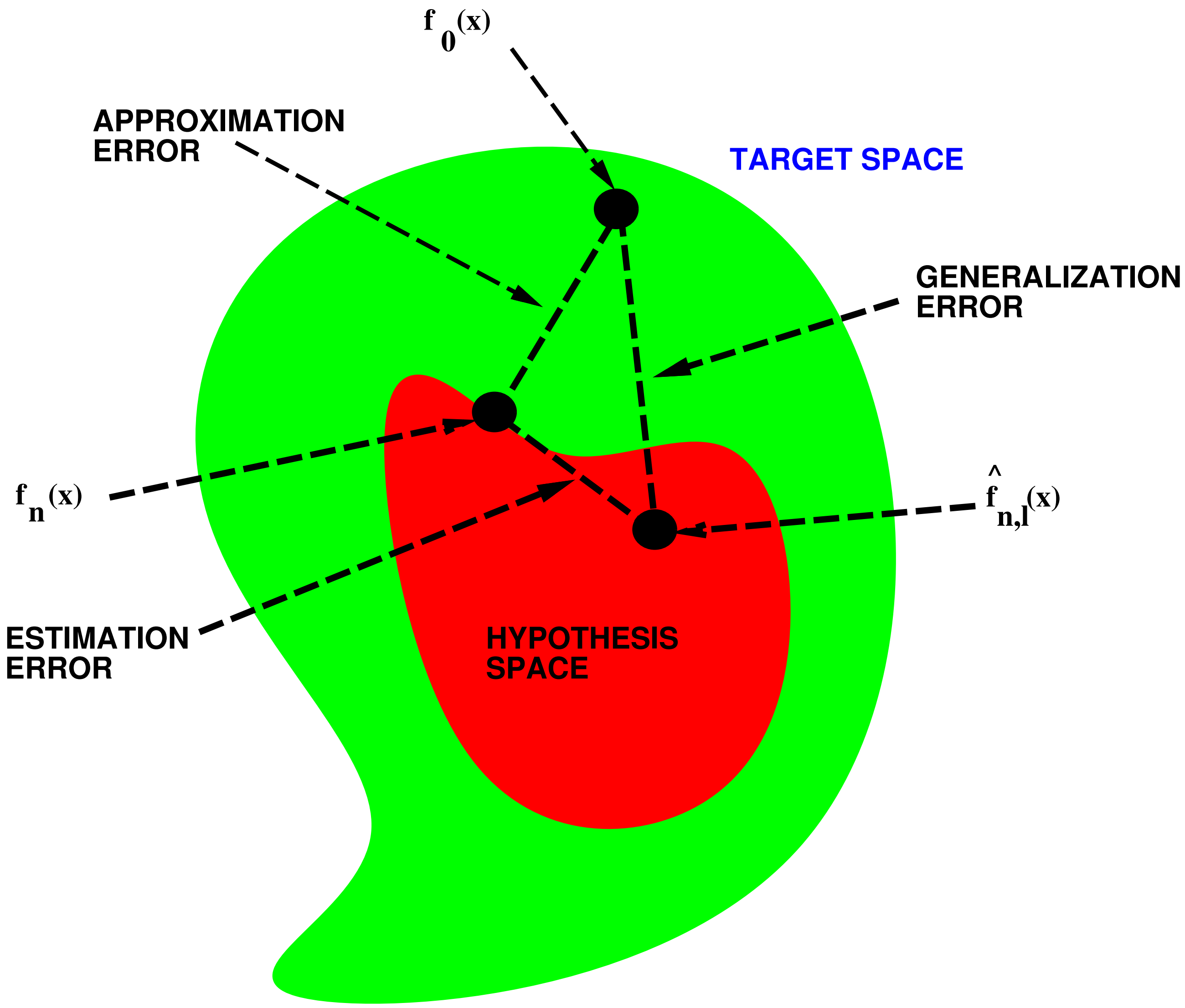
Radial Basis Functions



- 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 $\hat{f}_{n,l}$ (the result of learning)

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

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



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

- **Approximation error:** it depends critically on the target space $\mathcal{T}$, and less critically on the hypothesis space H_n;

- **estimation error:** it does **not** depend critically on the target space $\mathcal{T}$, but it **does** depend critically on the hypothesis space H_n.