Class 2

Empirical risk minimization and regularization theory

Federico Girosi

Menu

1. Learning as function approximation;

2. Empirical risk minimization + smoothness constraint = regularization theory;

3. Solution of the approximation problem in the framework of regularization theory ;

4. Bayesian interpretation of regularization theory;

5. Radial Basis Functions;

6. Additive splines;

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:

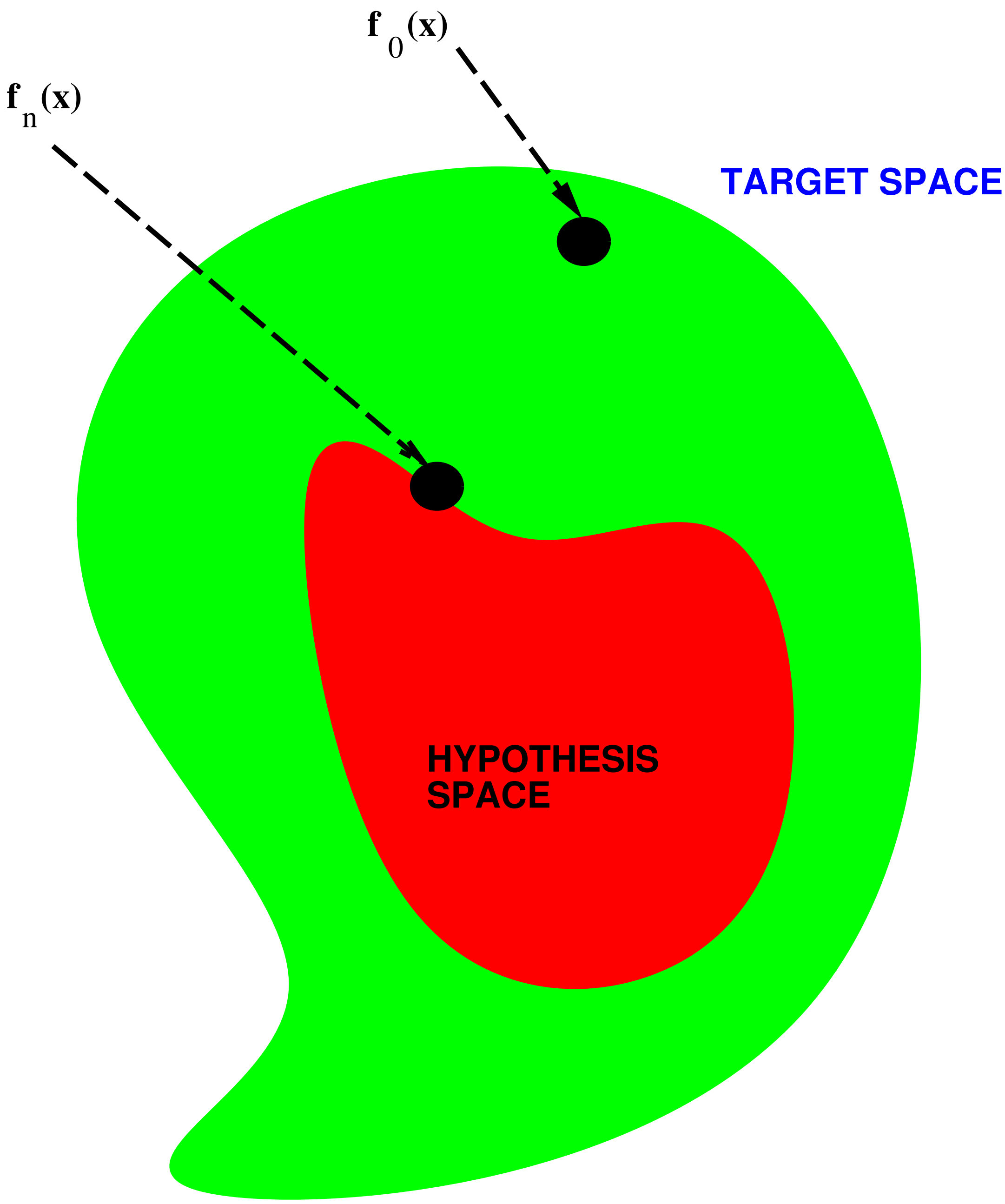
$$f_n = \text{arg} \min_{f \in H_n} I[f]$$

2.

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

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.



f 0(x)
f n(x)
TARGET SPACE
HYPOTHESIS SPACE

Empirical Risk Minimization

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

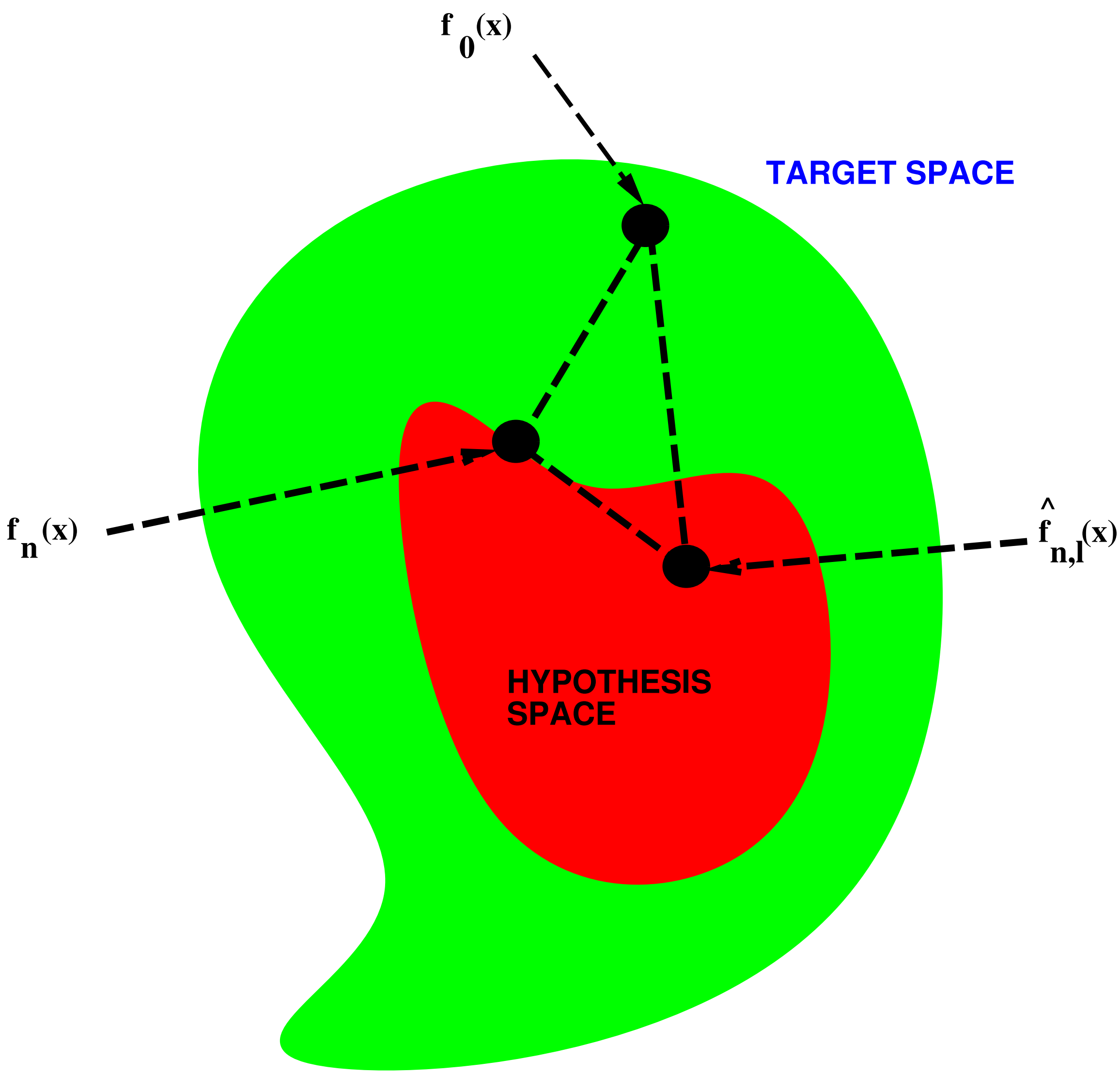
However we can approximate $I[f]$ by the *empirical risk*

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

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

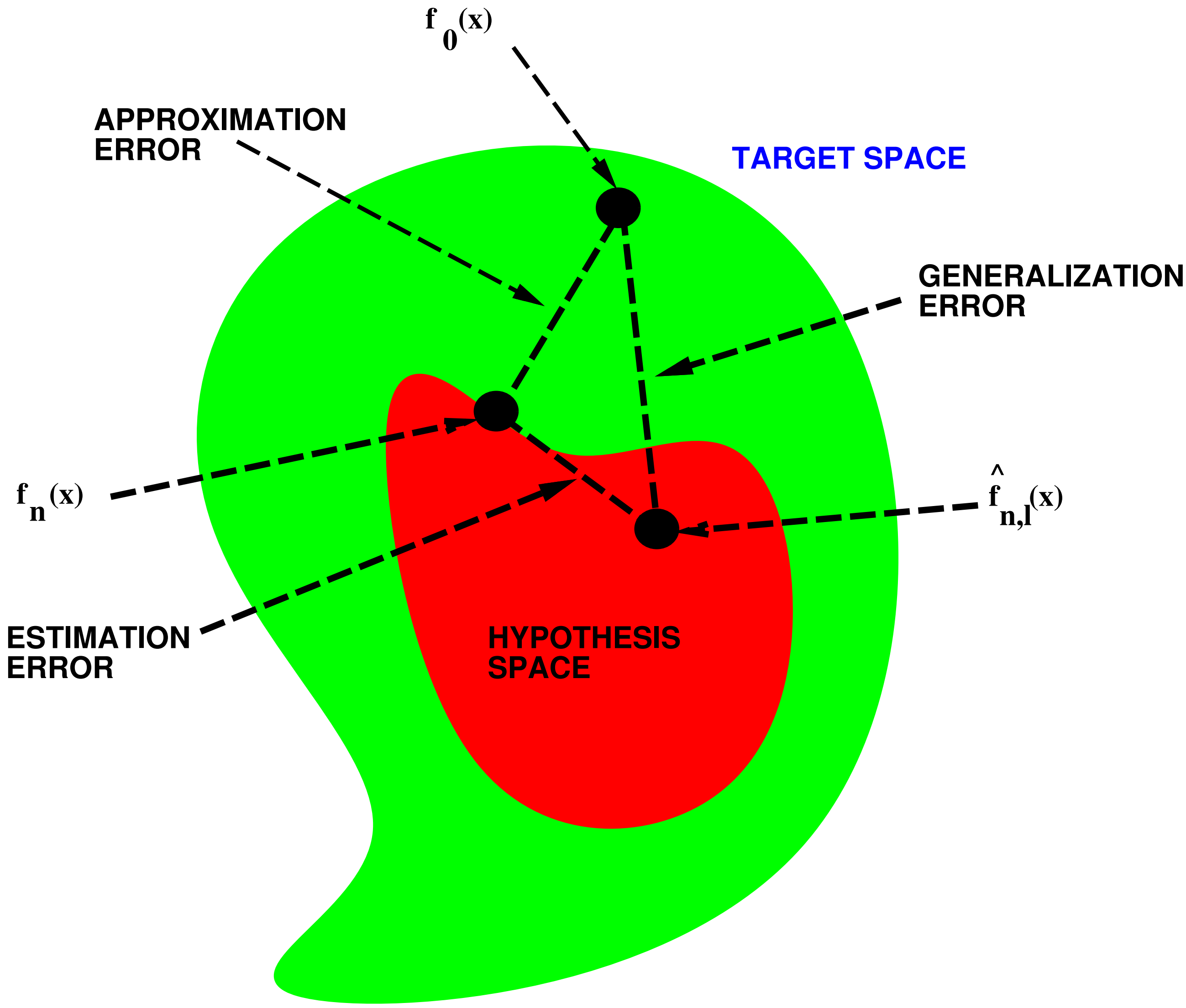
- The best we can possibly do: to find $f_0(\mathbf{x})$

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

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

Example

If $H_n = \{f \mid f(\mathbf{x}) = \sum_{\alpha=1}^{n} c_\alpha \sigma(\mathbf{x} \cdot \mathbf{w}_\alpha + \theta_\alpha)\}$, then we have to solve the following problem

$$\min_{c_\alpha, \mathbf{w}_\alpha, \theta_\alpha} \sum_{i=1}^{l} \left(y_i - \sum_{\alpha=1}^{n} c_\alpha \sigma(\mathbf{x}_i \cdot \mathbf{w}_\alpha + \theta_\alpha) \right)^2$$

(Neural Network Training)

Hypothesis spaces

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

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

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

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

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

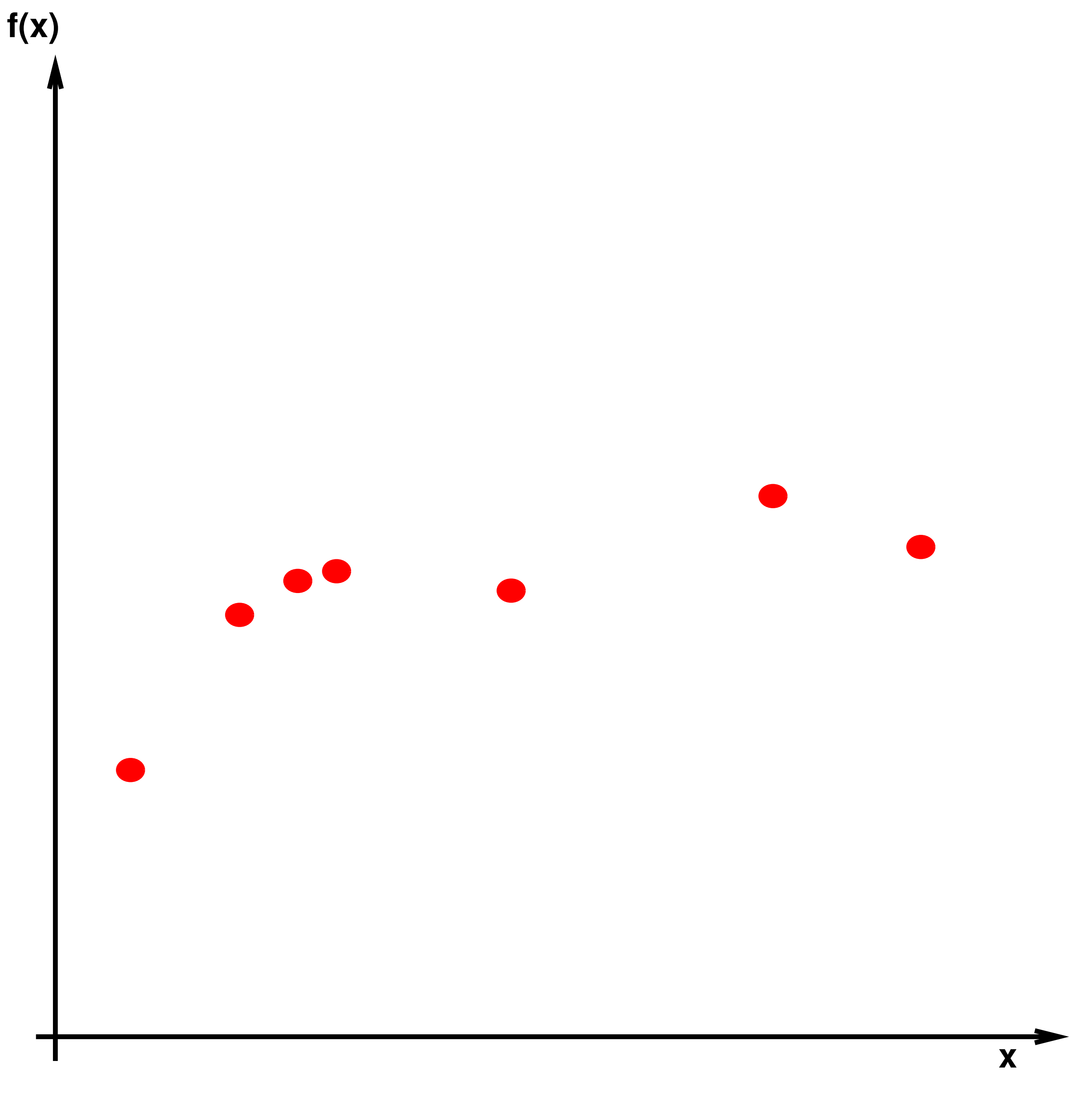
$$f(\mathbf{x}) = \sum_{\alpha=1}^{n} c_\alpha \sigma(\mathbf{x} \cdot \mathbf{w}_\alpha + \theta_\alpha)$$

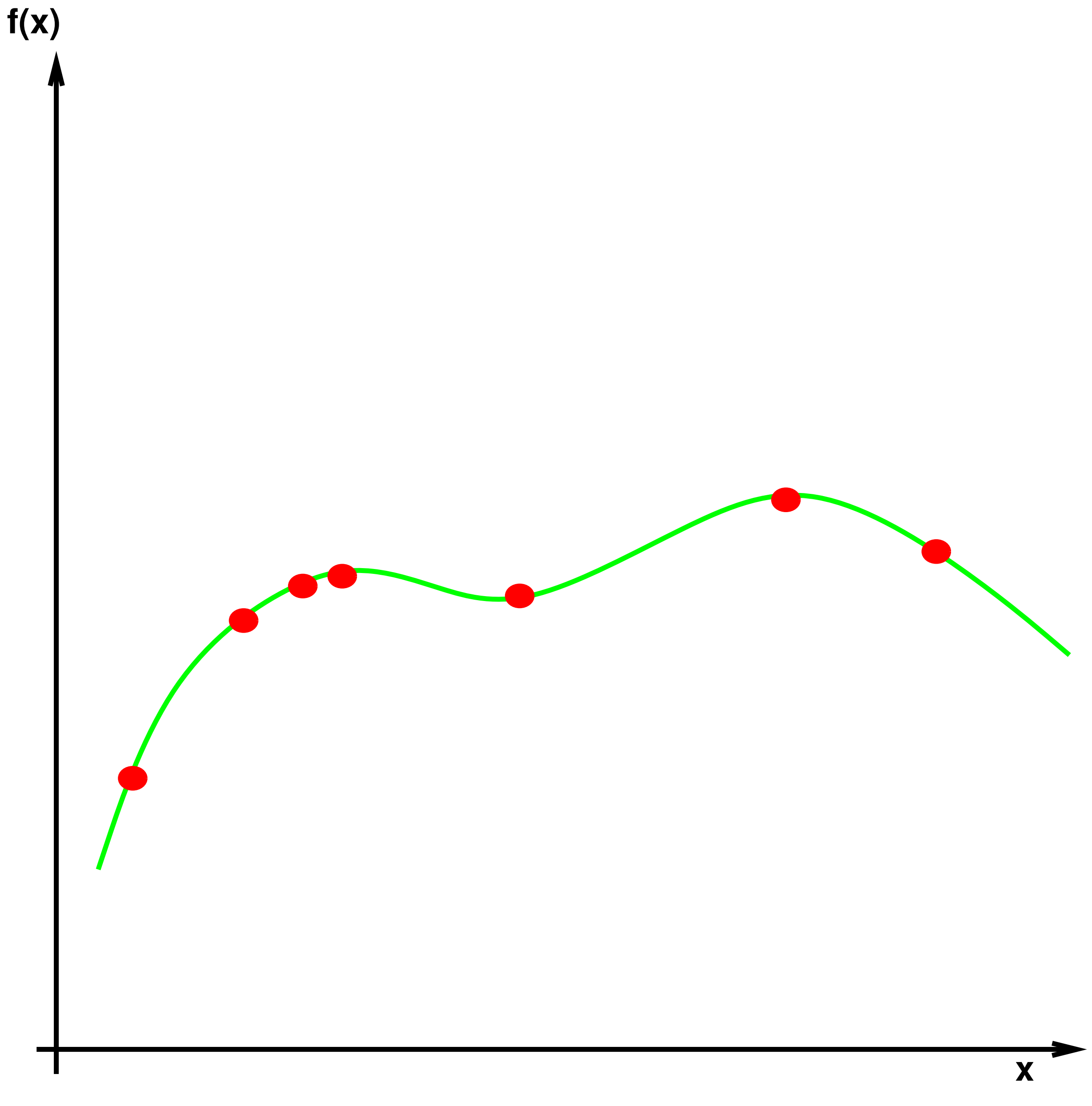
$$f(\mathbf{x}) = \frac{\sum_{\alpha=1}^{n} c_\alpha G(\|\mathbf{x} - \mathbf{t}_\alpha\|)}{\sum_{\alpha=1}^{m} d_\alpha \sigma(\mathbf{x} \cdot \mathbf{w}_\alpha + \theta_\alpha)}$$

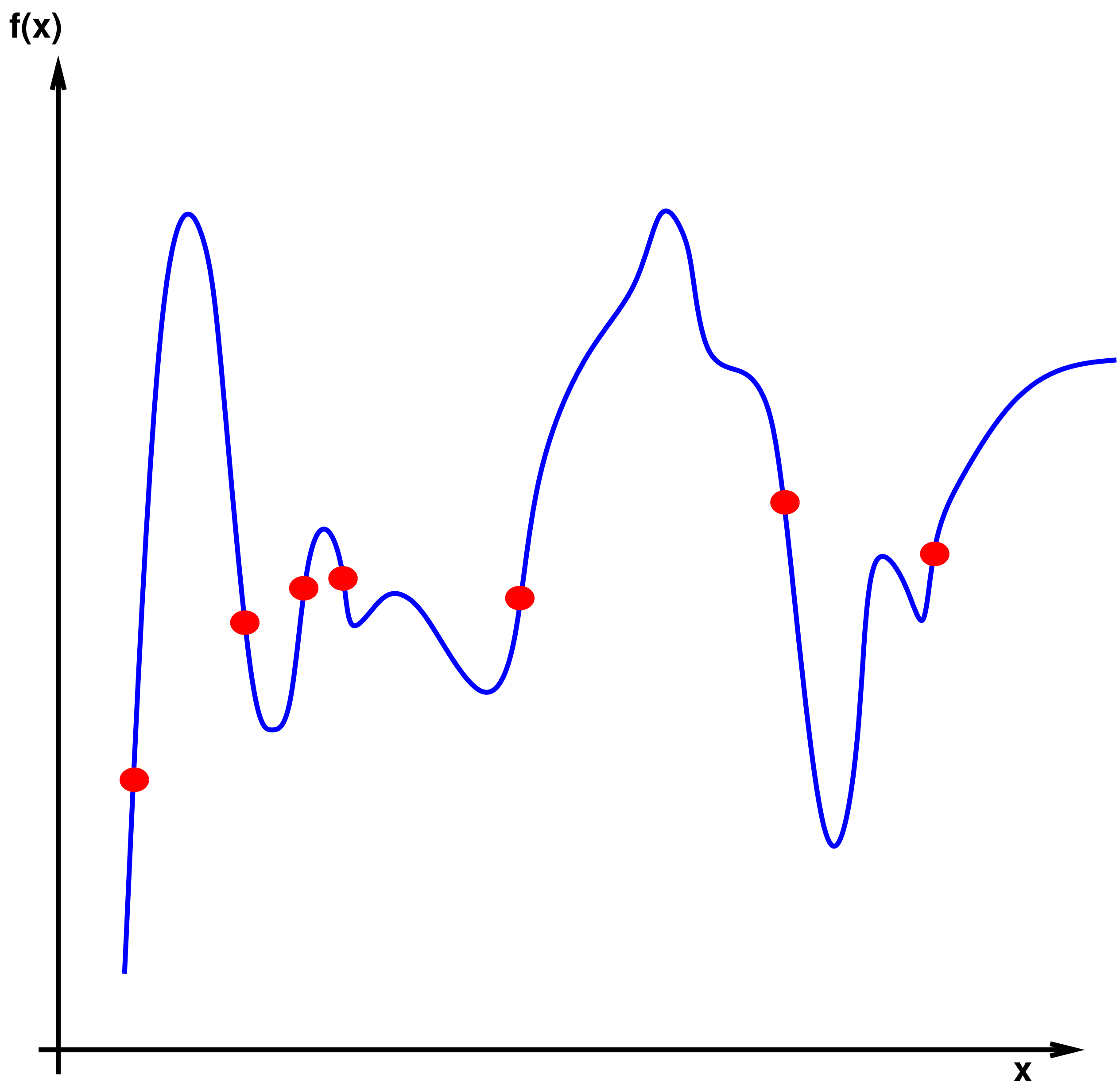
$$\ldots$$

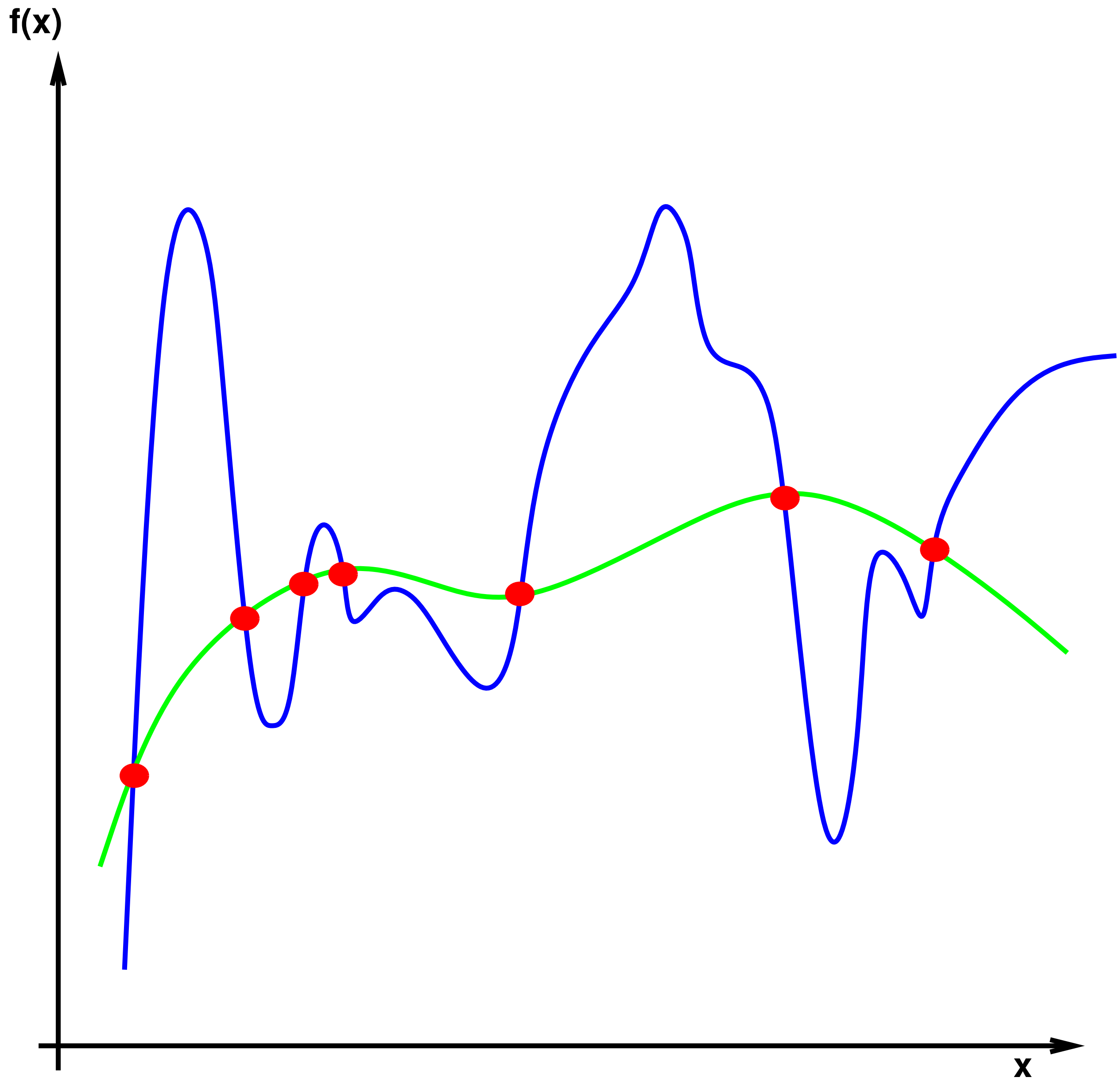
Target spaces

- $\mathcal{T}$ = 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;


f(x)
x


f(x)
x


f(x)
x


f(x)
x

The problem of approximating a function from sparse data has an infinite number of solutions, unless an "appropriate" hypothesis space $\mathcal{H}$ is chosen.

The kind of hypothesis space to use should be chosen according to some *a priori knowledge* on the problem.

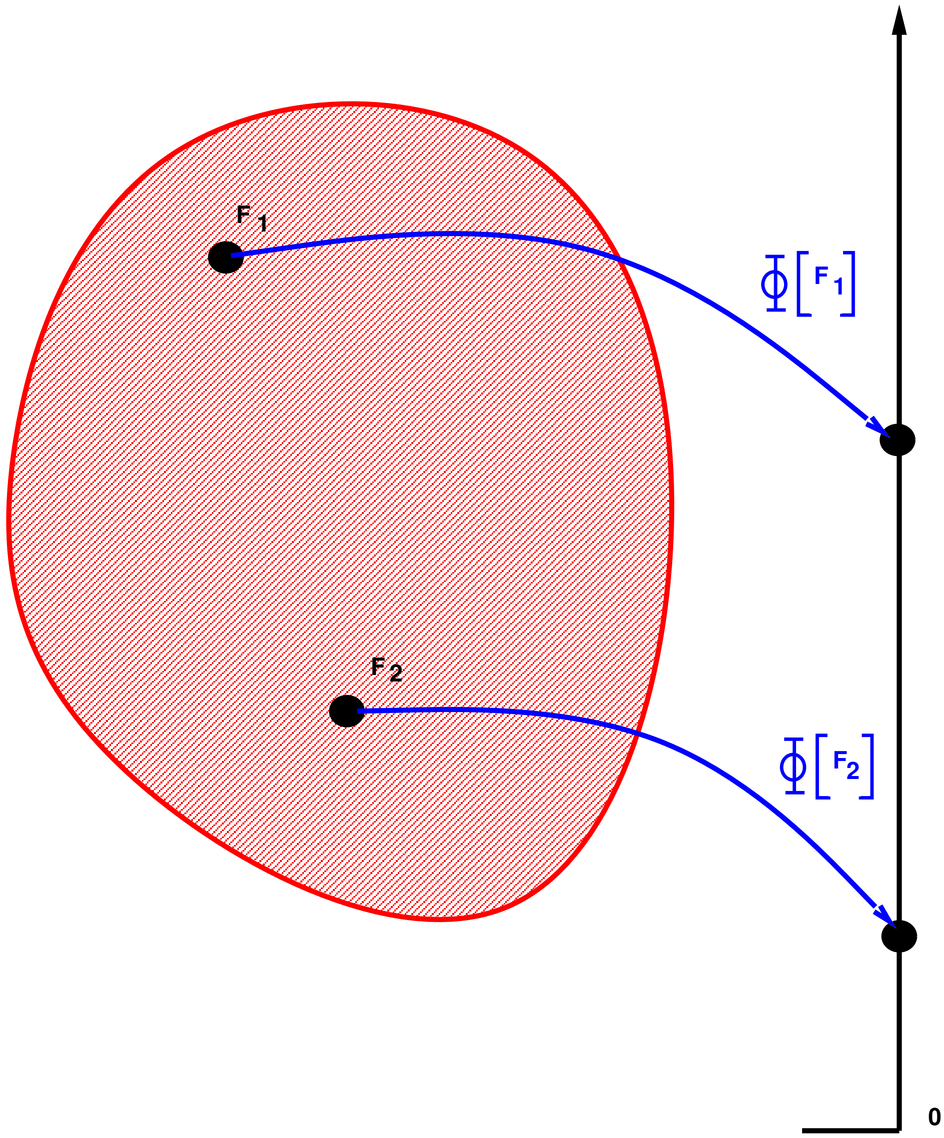
Different choices of $\mathcal{H}$ will correspond to different approximation techniques.

The simplest form of a priori knowledge: *smoothness*

We expect the functions underlying the data to be “nice and smooth”, in the sense that *if two inputs are close, the two corresponding output should be close too*.

smoothness = absence of oscillation

How can we measure smoothness?



Map the functions onto real, positive numbers and compare them!

Smoothness functionals

A map between a set of functions and the set of real numbers is called a *functional*.

We assume that we can define *smoothness* functionals $\phi[f]$, which assume large values for not smooth functions, and small values for functions whith little oscillation.

Smoothness functionals and hypothesis spaces

A smoothness functional $\phi[f]$ naturally defines a *family* of hypothesis spaces $\mathcal{H}_\mu$ as follows:

$$\mathcal{H}_\mu = \{f \mid \phi[f] \leq \mu\}$$

Notice that, if $\mu_1 \leq \mu_2 \leq \ldots \leq \mu_n$ then

$$\mathcal{H}_{\mu_1} \subset \mathcal{H}_{\mu_2} \subset \ldots \subset \mathcal{H}_{\mu_n}$$

Empirical risk minimization and smoothness functionals

If $\mathcal{H}_\mu$ is chosen as hypothesis space the problem of minimizing the empirical risk is the following:

$$\min \sum_{i=1}^{N} (y_i - f(\mathbf{x}_i))^2$$

subject to the constraint

$$\phi[f] \leq \mu$$

Empirical risk minimization and smoothness functionals

The problem above can be solved using the technique of Lagrange multipliers:

$$\min \left[\sum_{i=1}^{N} (y_i - f(\mathbf{x}_i))^2 + \lambda(\phi[f] - \mu) \right]$$

where λ is a positive number (the Lagrange multiplier).

Denoting by f_λ the solution of the minimization problem above, the relationship between λ and μ is given by:

$$\phi[f_\lambda] = \mu$$

Empirical risk minimization and smoothness functionals

- Since μ is a constant, it can be omitted by the minimization problem above, that becomes:

$$\min \left[\sum_{i=1}^{N} (y_i - f(\mathbf{x}_i))^2 + \lambda \phi[f] \right]$$

- since μ is a "free" parameter, and μ determines λ, we consider λ as the free parameter, and forget about μ.

- The hypothesis space will be therefore denoted by $\mathcal{H}_\lambda$.

Empirical risk minimization and smoothness functionals

Therefore, when the hypothesis space is $\mathcal{H}_\lambda$, the principle of empirical risk minimization leads to solve the following minimization problem:

$$\min_{f \in \Phi} H[f]$$

where Φ is the set of functions for which $\phi[f]$ is well defined and $H[f]$ is the following functional:

$$H[f] = \sum_{i=1}^{N} (y_i - f(\mathbf{x}_i))^2 + \lambda \phi[f]$$

Empirical risk minimization and Regularization Theory

This approach has been proposed in the 70's in the different framework of *regularization theory*.

Regularization theory (Tikhonov, 1963; Tikhonov and Arsenin, 1977) is a set of mathematical techniques developed to deal with **ill-posed problems.**

Ill-posed problems

An *ill-posed* problem is a problem that is not *well-posed*

A problem is given by a *datum* g and a *solution* u, and it is well-posed (in the sense of Hadamard) when

1. For each datum g in a class of functions Y there exists a solution u in a prescribed class X (**existence**);

2. The solution u is unique in X (**uniqueness**).

3. The dependence of u upon g is continuous (**continuity**).

Ill-posed Problems

An *ill-posed* problem is a problem that is not *well-posed*

A problem is given by a *datum* g and a *solution* u, and it is well-posed in the sense of  **when**

- **For each datum g in a class of functions Y there exists a solution u in a prescribed class X (*existence*);**
- **The solution u is unique in X (*uniqueness*).**
- **The dependence of u upon g is continuous (*continuity*).**

Regularization theory

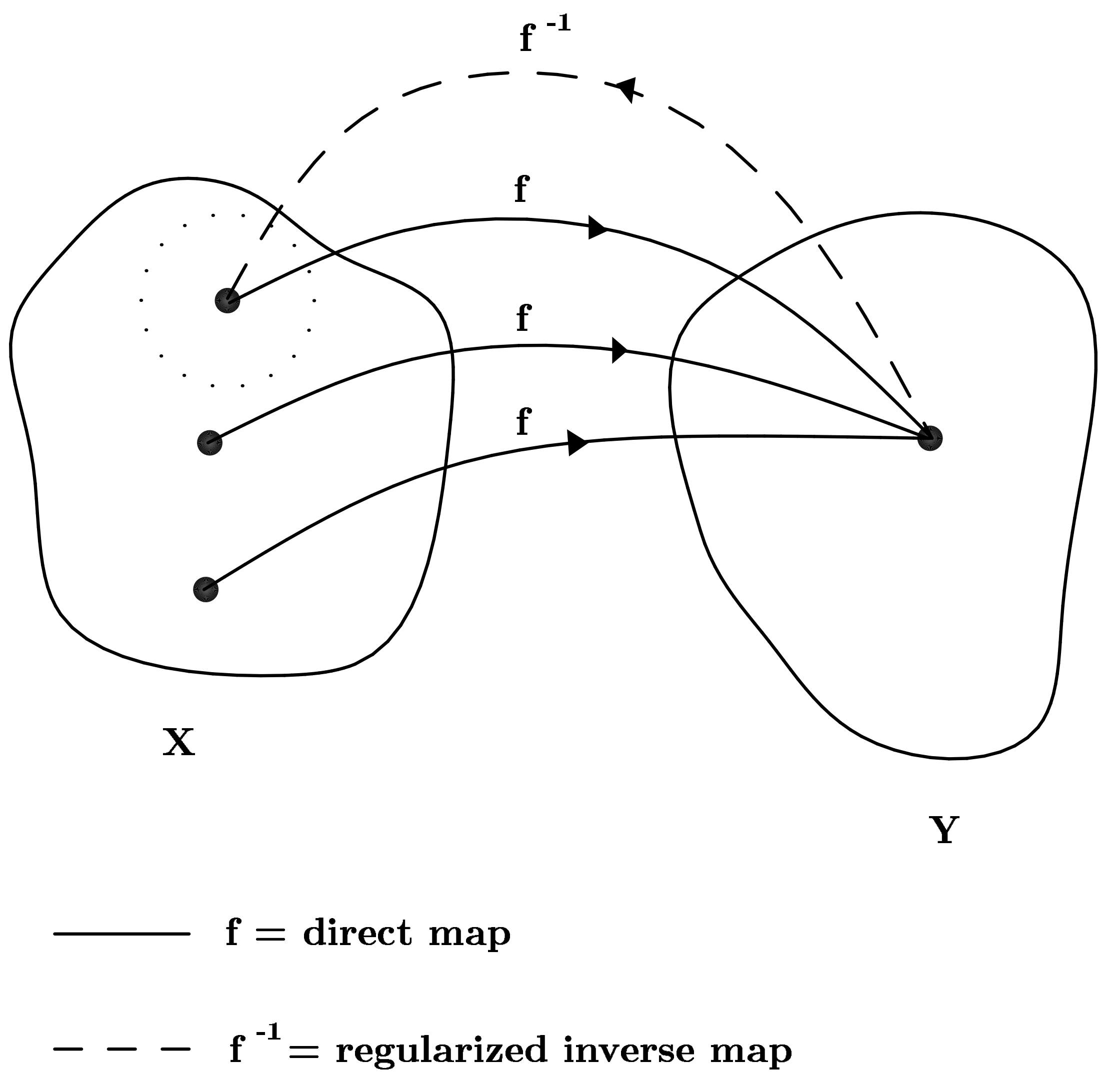
Classical mathematical physics problems are well-posed (Dirichlet problem for elliptic equations, forward problem for the heat equation, direct problem in scattering ...)

Inverse problems are usually ill-posed

Regularization theory (Tikhonov, 1963; Tikhonov and Arsenin, 1977; Morozov, 1984) transforms an ill-posed problem to a well-posed one, using *a priori* knowledge on the nature of the solution.

Inverse problems

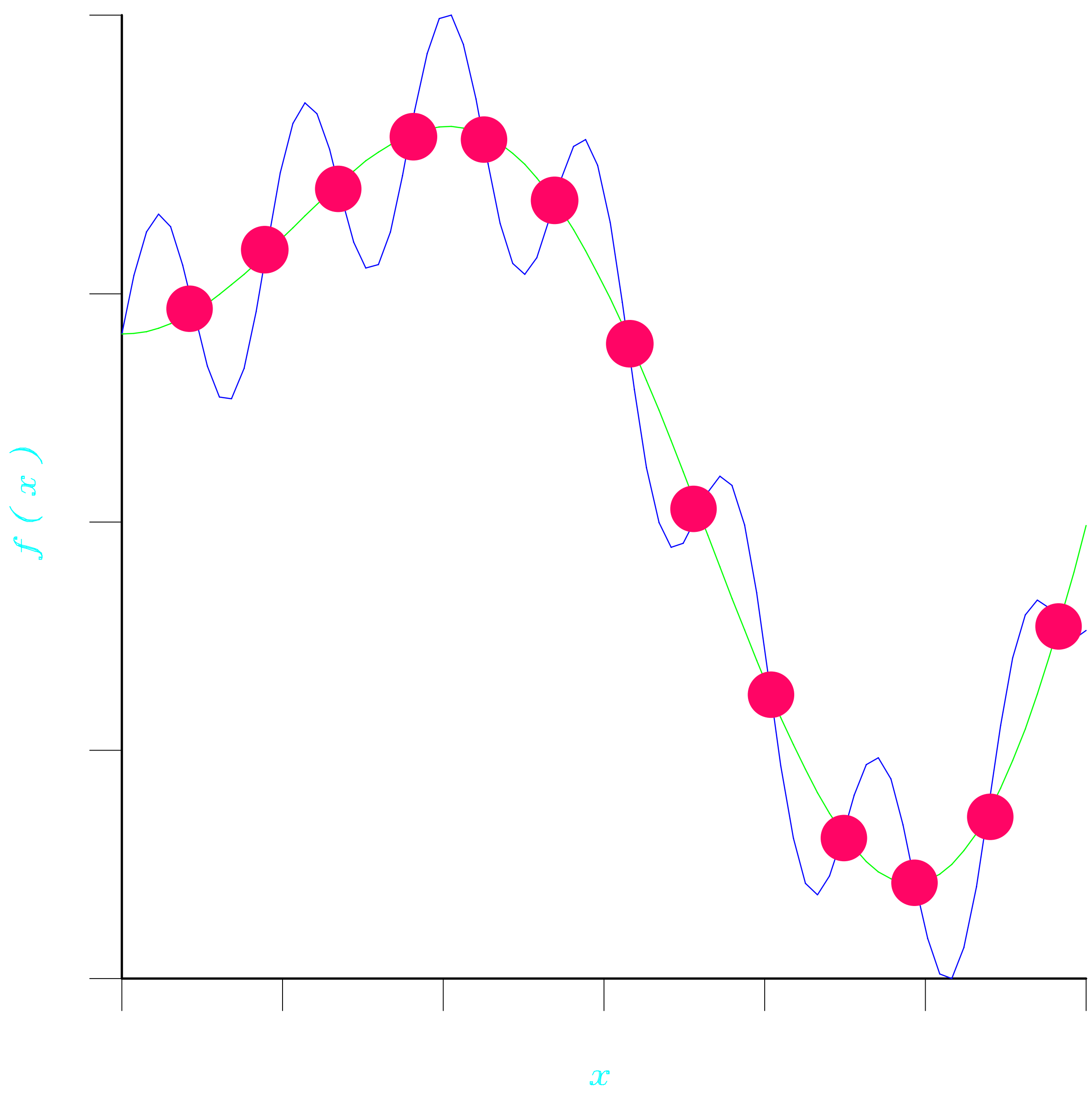
When the solution of the problem is not unique, it is possible to obtain a unique solution *constraining* the space of solutions in appropriate way. The shape of the constraint depends on apriori knowledge on the solution.



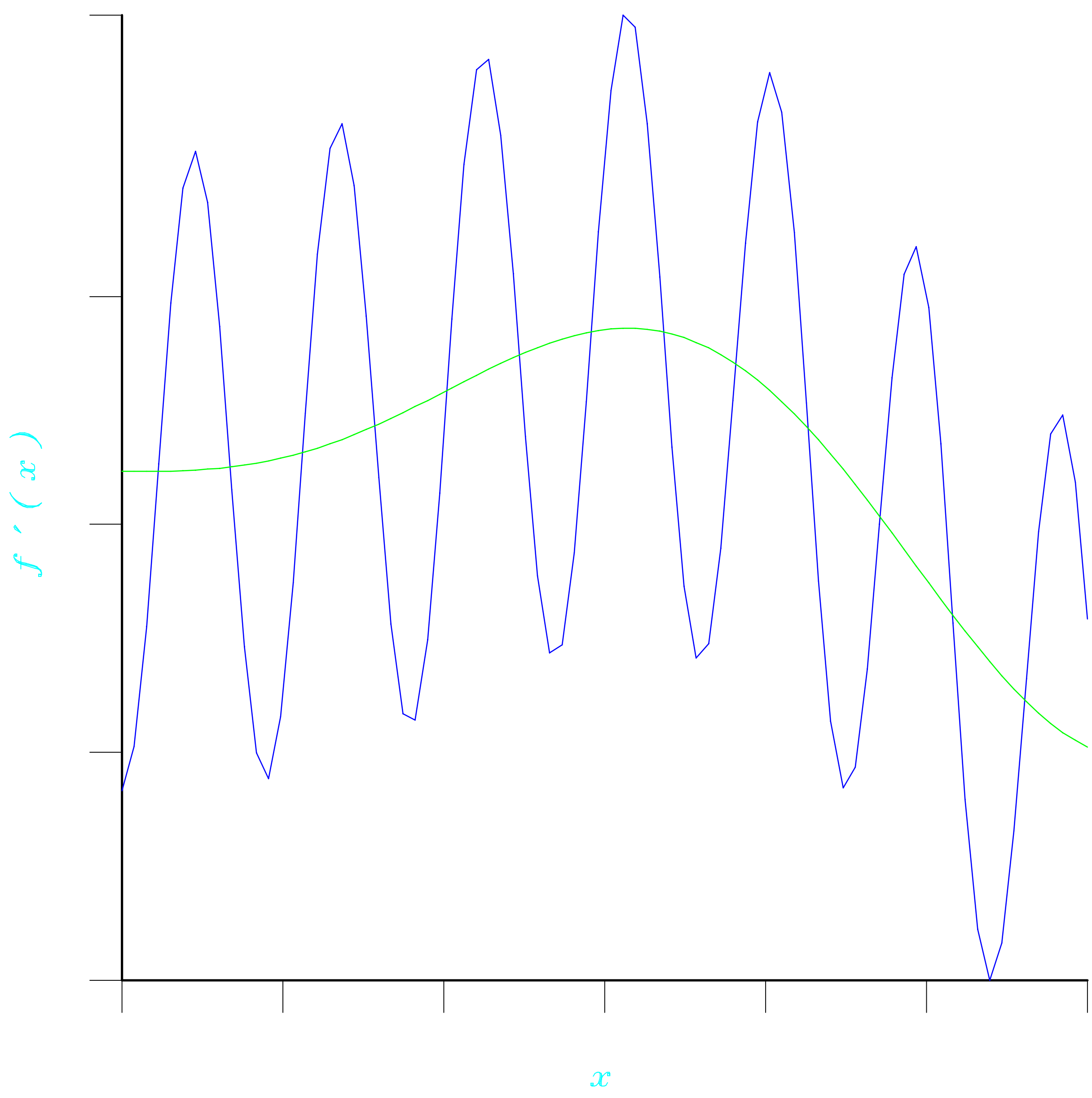
The simplest form of a priori knowledge: *smoothness*

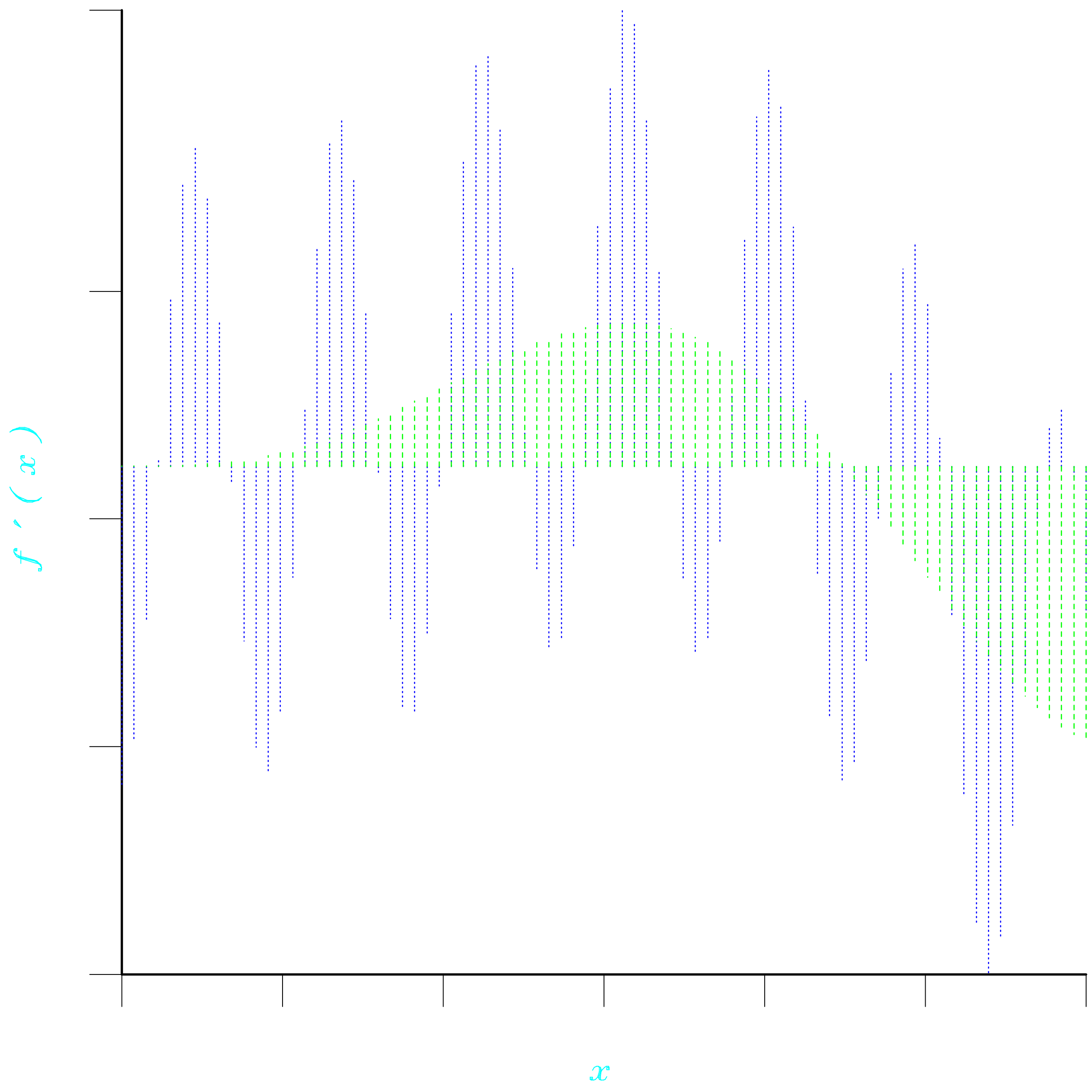
We expect the functions underlying the data to be “nice and smooth”, in the sense that *if two inputs are close, the two corresponding output should be close too*.

smoothness = absence of oscillation

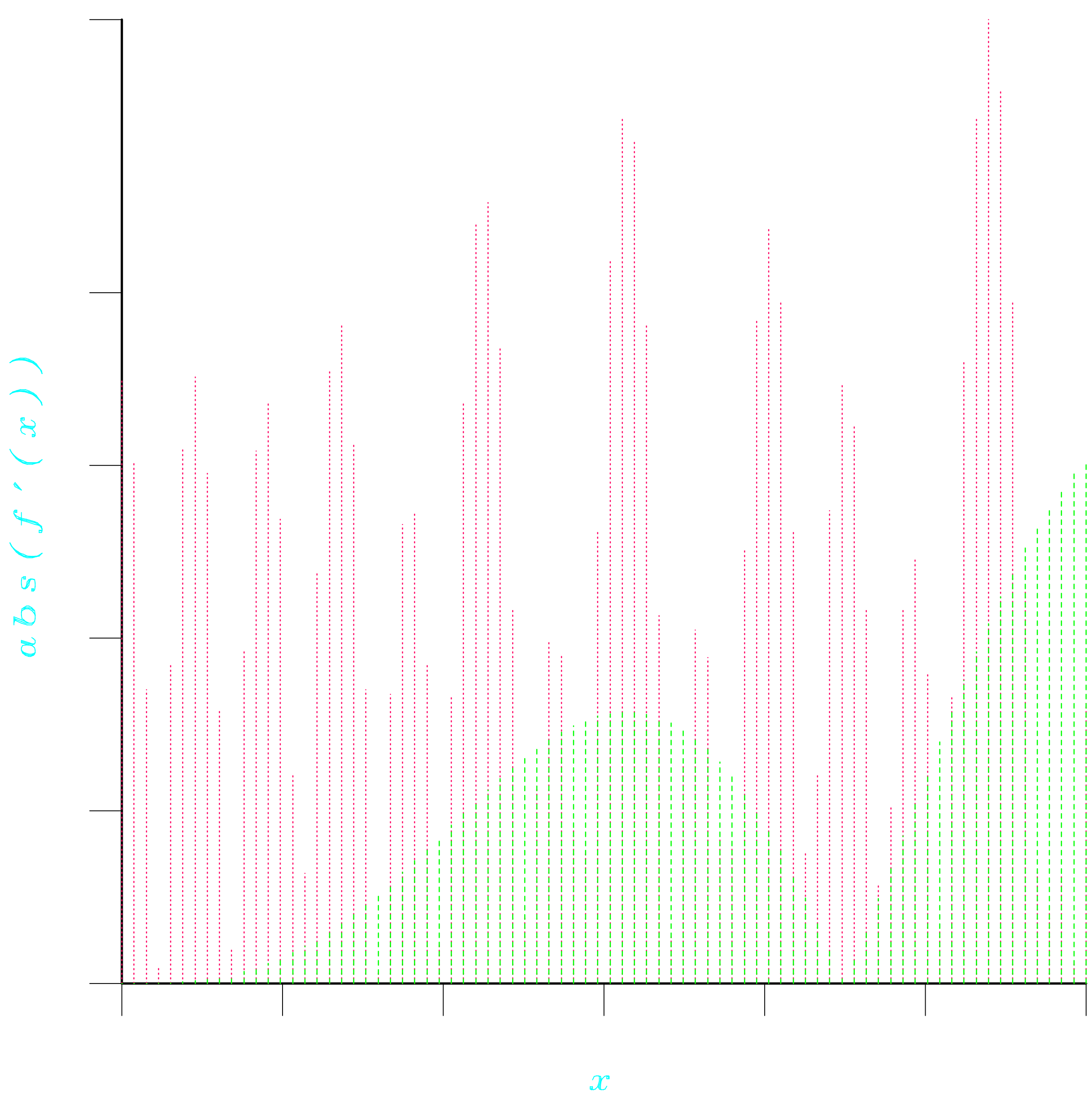


$f(x)$
x

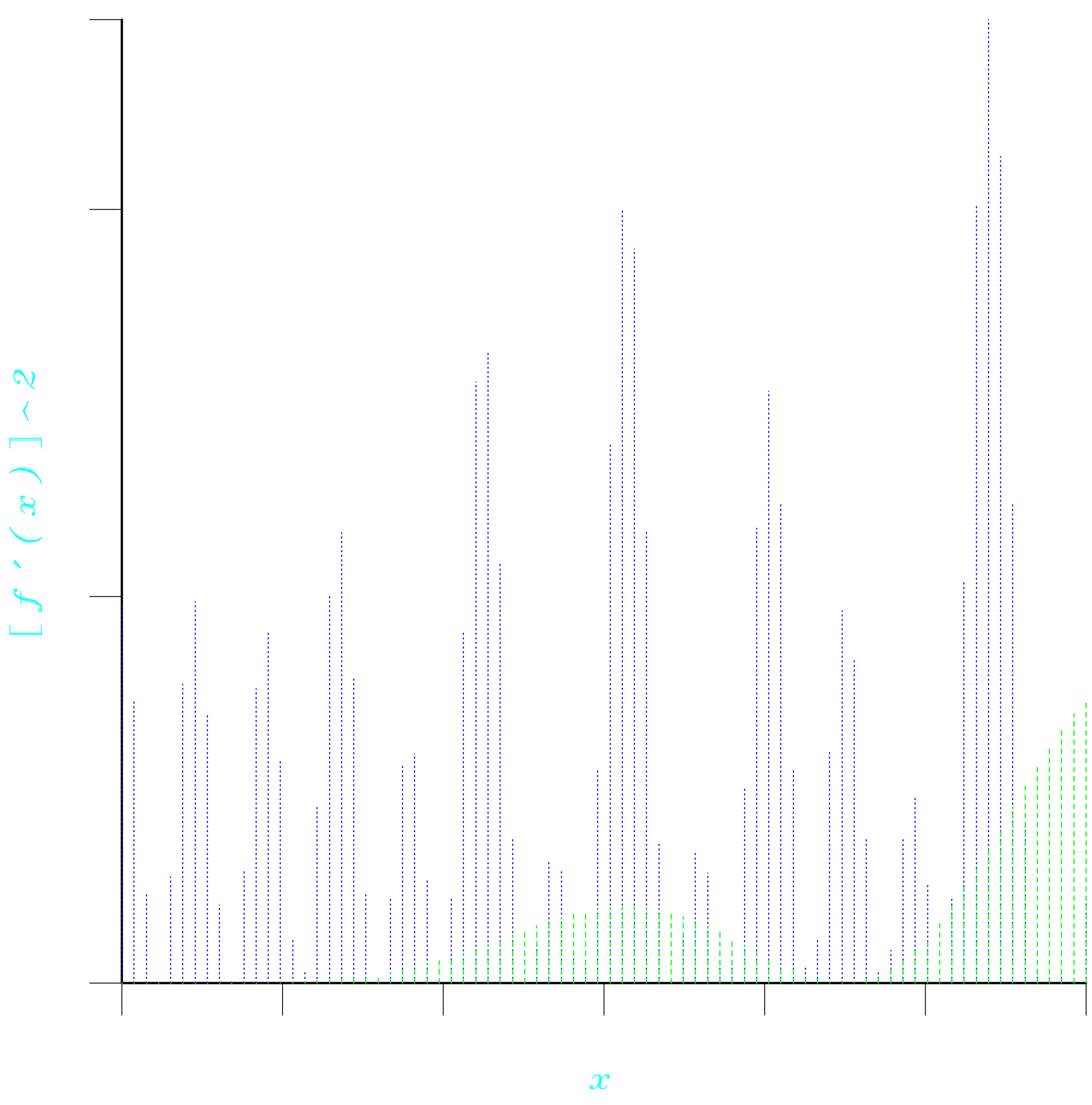

f ′ (x)
x



$f\ '\ (\ x\ )$
x



abs(f′(x))
x


[f ' (x)] ^ 2
x

Smoothness functionals

$$\phi_1[f] = \int_R \; dx \; (f'(x))^2 = \int_R ds \; s^2 |\tilde{f}(s)|^2$$

$$\phi_2[f] = \int_R \; dx \; (f''(x))^2 = \int_R ds \; s^4 |\tilde{f}(s)|^2$$

$$\phi_3[f] = \phi_1[f] + \phi_2[f] = \int_R ds \; (s^2 + s^4) |\tilde{f}(s)|^2$$

More generally

$$\phi[f] = \int_{R^d} \; d\mathbf{s} \; \frac{|\tilde{f}(\mathbf{s})|^2}{\tilde{G}(\mathbf{s})}$$

for any positive, symmetric function $\tilde{G}$ decreasing to zero at infinity.

Regularization approach

A smooth function that approximates the data set $D = \{(\mathbf{x}_i, y_i) \in R^d \times R\}_{i=1}^{l}$ can be found minimizing the functional:

$$H[f] = \sum_{i=1}^{l} (y_i - f(\mathbf{x}_i))^2 + \lambda \int_{R^d} d\mathbf{s} \, \frac{|\tilde{f}(\mathbf{s})|^2}{\tilde{G}(\mathbf{s})}$$

where $\tilde{G}$ is a positive, symmetric function, decreasing to zero at infinity, and λ a small, positive number.

Examples

$\tilde{G}(\mathbf{s}) = \frac{1}{\|\mathbf{s}\|^{2m}}$

$\tilde{G}(\mathbf{s}) = e^{-\|\mathbf{s}\|^2}$

More general: $\tilde{G}$ is the Fourier transform of a *conditionally positive definite function* of order k.

Conditionally positive definite functions

Definition 1 *A continuous function $f(t)$, defined on $[0, \infty)$, is said to be conditionally (strictly) positive definite of order k on R^n if for any distinct points $\mathbf{x}_1, \ldots, \mathbf{x}_N \in R^n$ and scalars $c_1, \ldots, c_N$ such that $\sum_{i=1}^{l} c_i p(\mathbf{x}_i) = 0$ for all $p \in \pi_{k-1}(R^n)$, the quadratic form*

$$\sum_{i=1}^{l} \sum_{j=1}^{l} c_i c_j f(\|\mathbf{x}_i - \mathbf{x}_j\|)$$

is (positive) nonnegative.

Positive definite functions

Definition 2 *A continuous function $f(t)$, defined on $[0, \infty)$, is said to be positive definite on R^n if for any distinct points $\mathbf{x}_1, \ldots, \mathbf{x}_N \in R^n$ and scalars $c_1, \ldots, c_N$ the quadratic form*

$$\sum_{i=1}^{l} \sum_{j=1}^{l} c_i c_j f(\|\mathbf{x}_i - \mathbf{x}_j\|)$$

is (positive) nonnegative.

Bochner's theorem

A function is positive definite if and only if it is the Fourier transform of a positive, integrable function.

Examples

$$\tilde{G}(\mathbf{s}) = \frac{1}{||\mathbf{s}||^{2m}}$$

$$G(\mathbf{x}) = \begin{cases} ||\mathbf{x}||^{2m-n} \ln ||\mathbf{x}|| & \text{if } 2m > n \text{ and } n \text{ is even} \\ ||\mathbf{x}||^{2m-n} & \text{otherwise.} \end{cases}$$

Conditionally positive definite of order $m-1$

===================

$$\tilde{G}(\mathbf{s}) = e^{-||\mathbf{s}||^2}$$

$$G(\mathbf{x}) = e^{-||\mathbf{x}||^2}$$

Positive definite

If G is a conditionally positive definite function of order m, then

$$\phi[f] = \int_{R^d} d\mathbf{s} \, \frac{|\tilde{f}(\mathbf{s})|^2}{\tilde{G}(\mathbf{s})}$$

is a *seminorm* whose null space is the set of polynomials of degree $m - 1$.

If G is a positive definite function then ϕ is a norm.

(Madych and Nelson, 1990)

Main result

(Duchon, 1977; Meinguet, 1979; Wahba, 1977; Madych and Nelson, 1990; Poggio and Girosi, 1989; Girosi, 1992)

The function that minimizes the functional

$$H[f] = \sum_{i=1}^{l} (y_i - f(\mathbf{x}_i))^2 + \lambda \int_{R^d} d\mathbf{s} \, \frac{|\tilde{f}(\mathbf{s})|^2}{\tilde{G}(\mathbf{s})}$$

has the form:

$$f(\mathbf{x}) = \sum_{i=1}^{l} c_i G(\mathbf{x} - \mathbf{x}_i) + \sum_{\alpha=1}^{k} d_\alpha \gamma_\alpha(\mathbf{x})$$

where G is conditionally positive definite of order m, $\tilde{G}$ is the Fourier transform of G and $\{\gamma_\alpha\}_{\alpha=1}^{k}$ is a basis in the space of polynomials of degree $m-1$.

Computation of the coefficients

The coefficients are found by solving the linear system:

$$(G + \lambda I)\mathbf{c} + \Gamma^T \mathbf{d} = \mathbf{y}$$

$$\Gamma \mathbf{c} = 0$$

where I is the identity matrix, and we have defined

$$(\mathbf{y})_i = y_i \ , \quad (\mathbf{c})_i = c_i \ , \quad (\mathbf{d})_i = d_i$$

$$(G)_{ij} = G(\mathbf{x}_i - \mathbf{x}_j) \ , \quad (\Gamma)_{\alpha i} = \gamma_\alpha(\mathbf{x}_i)$$

Numerical considerations

- Extreme care should be taken when solving the linear system, which has a tendency to be *ill-conditioned*;

- Ill-conditioning decreases with increasing values of λ;

- for $\lambda = 0$ numerical difficulties have been reported even for data sets as small as 100 data points;

- Singular Values Decomposition (SVD) or similar techniques are suggested in order to make the solution stable;

Examples

Multivariate splines:

$$\tilde{G}(\mathbf{s}) = \frac{1}{\|\mathbf{s}\|^{2m}}$$

$$G(\mathbf{x}) = \begin{cases} \|\mathbf{x}\|^{2m-n} \ln \|\mathbf{x}\| & \text{if } 2m > n \text{ and } n \text{ is even} \\ \|\mathbf{x}\|^{2m-n} & \text{otherwise.} \end{cases}$$

It requires a polynomial of degree $< m$.

==================

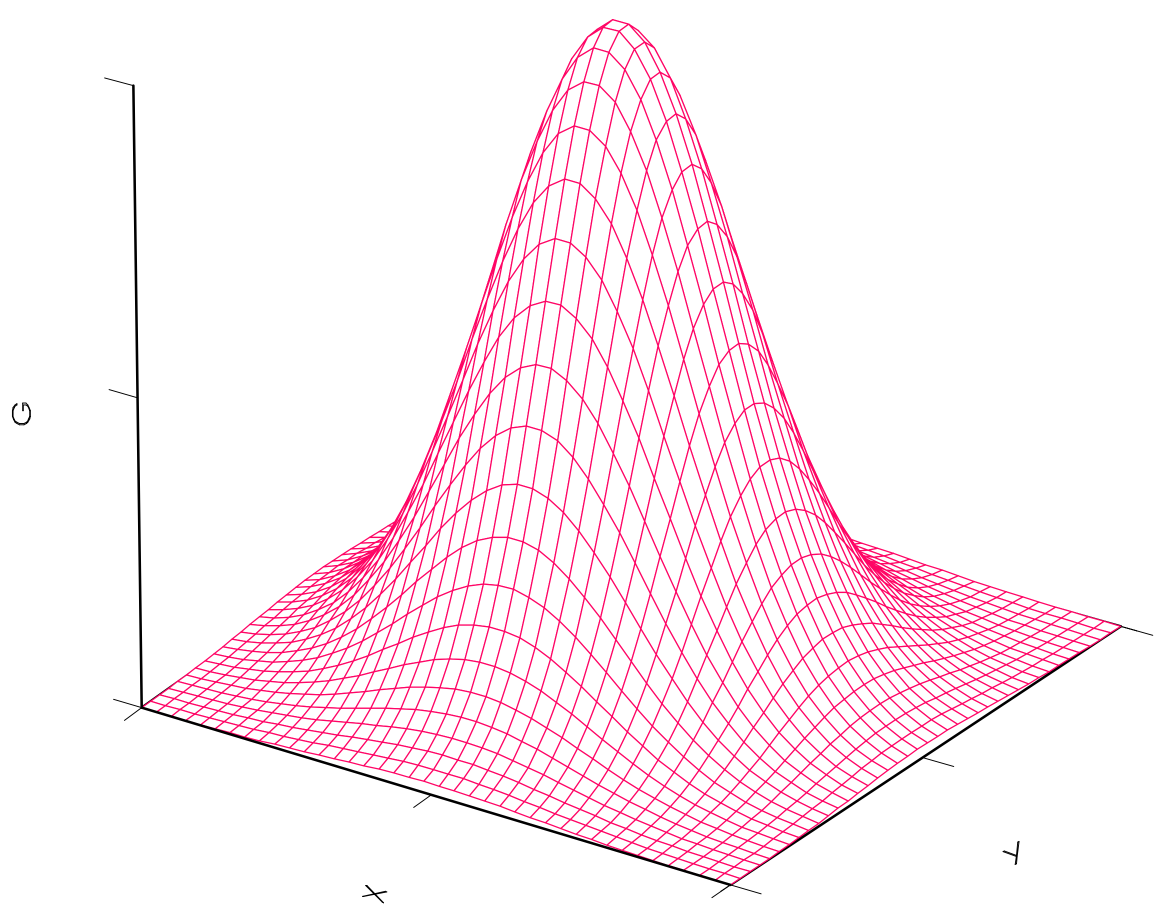
The gaussian

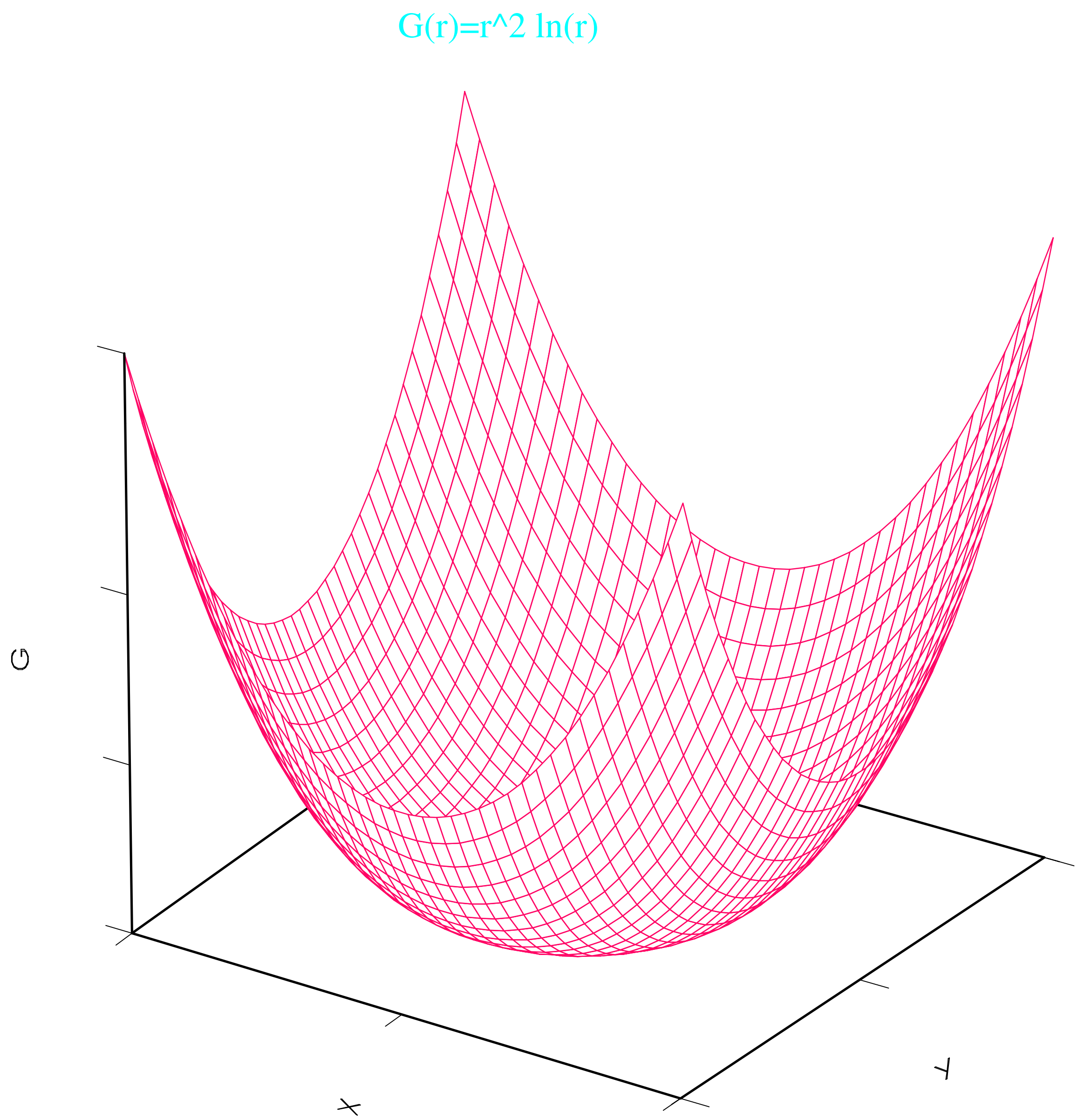
$$\tilde{G}(\mathbf{s}) = e^{-\|\mathbf{s}\|^2}$$

$$G(\mathbf{x}) = e^{-\|\mathbf{x}\|^2}$$

It does not require any polynomial.

G(r)=exp(-r^2)




G(r)=r^2 ln(r)
G
X
Y

Example: the One Dimensional Linear Splines

$$\phi[f] = \int (f'(x))^2 \; dx = \int s^2 |\tilde{f}(s)|^2 \; ds$$

$$\tilde{G}(s) = \frac{1}{s^2}$$

$$G(x-y) = |x-y|$$

$$f(x) = \sum_{i=1}^{l} c_i |x - x_i| + a \; .$$

The solution is a piecewise linear function.

Example: the One Dimensional Cubic Splines

$$\phi[f] = \int (f''(x))^2 \ dx = \int s^4 |\tilde{f}(s)|^2 \ ds$$

$$\tilde{G}(s) = \frac{1}{s^4}$$

$$G(x-y) = |x-y|^3$$

$$f(x) = \sum_{i=1}^{l} c_i |x - x_i|^3 + ax + b \ .$$

The solution is a piecewise cubic polynomial.

Example: Thin PLate Splines in 2 Dimensions

$$\phi[f] = \int_{R^2} dxdy \ \left[\left(\frac{\partial^2 f}{\partial x^2} \right)^2 + 2 \left(\frac{\partial^2 f}{\partial x \partial y} \right)^2 + \left(\frac{\partial^2 f}{\partial y^2} \right)^2 \right]$$

$$\phi[f] = \int_{R^2} \ d\mathbf{s} \ \|\mathbf{s}\|^4 |f(\mathbf{s})|^2$$

$$\tilde{G}(\mathbf{s}) = \frac{1}{\|\mathbf{s}\|^4}$$

$$G(\mathbf{x}) = \|\mathbf{x}\|^2 \ln \|\mathbf{x}\|$$

$$f(\mathbf{x}) = \sum_{i=1}^{l} c_i \|\mathbf{x} - \mathbf{x}_i\|^2 \ln \|\mathbf{x} - \mathbf{x}_i\| + \mathbf{a} \cdot \mathbf{x} + b$$

Example: Gaussian Radial Basis Functions

$$\phi[f] = \int_{R^d} d\mathbf{s}\ e^{\|\mathbf{s}\|^2} |f(\mathbf{s})|^2$$

$$\tilde{G}(\mathbf{s}) = e^{-\|\mathbf{s}\|^2}$$

$$G(\mathbf{x}) = e^{-\|\mathbf{x}\|^2}$$

$$f(\mathbf{x}) = \sum_{i=1}^{l} c_i e^{-\|\mathbf{x}-\mathbf{x}_i\|^2}$$

Radial Basis Functions

Define $r = \|\mathbf{x}\|$

$$G(\mathbf{x}) = e^{-r^2} \qquad \text{gaussian}$$

$$G(\mathbf{x}) = \sqrt{r^2 + c^2} \qquad \text{multiquadric}$$

$$G(\mathbf{x}) = \frac{1}{\sqrt{c^2+r^2}} \qquad \text{inverse multiquadric}$$

$$G(\mathbf{x}) = r^{2n+1} \qquad \text{multivariate splines}$$

$$G(\mathbf{x}) = r^{2n} \ln r \qquad \text{multivariate splines}$$

The role of λ

The solution of the problem

$$\min_{f \in \Phi} \left[\sum_{i=1}^{l} (y_i - f(\mathbf{x}_i))^2 + \lambda \phi[f] \right]$$

depends on λ, and it will be denoted by f_λ. There are two important cases for λ:

1. $\lambda \rightarrow 0$

2. $\lambda \rightarrow +\infty$

Case 1: $\lambda \to 0$

It is easy to show that

$$\lim_{\lambda \to 0} f_\lambda(\mathbf{x}_i) = y_i \quad i = 1, \ldots, l$$

Therefore

$$\lim_{\lambda \to 0} f_\lambda = \text{the smoothest interpolant}$$

Case 2: $\lambda \to +\infty$

When $\lambda \to +\infty$ the functional $H[f]$ is defined only for those functions for which $\phi[f] = 0$.

Therefore

$$\lim_{\lambda \to 0} f_\lambda = \text{the best data fit in the null space of } \phi$$

Intermediate λ

For any value of λ we know that f_λ solves the following minimization problem

$$\begin{cases} \min_{f\in\Phi} \sum_{i=1}^{l} (y_i - f(\mathbf{x}_i))^2 \\ \phi[f] \leq \mu \end{cases}$$

for a value of μ given by

$$\phi[f_\lambda] = \mu.$$

It can be shown that

λ large $\Rightarrow$ μ small

λ small $\Rightarrow$ μ large

Choice of the regularization parameter λ

The parameter λ depends on the signal-to-noise ratio, and it is usually chosen according to *cross-validation* techniques.

The idea under cross-validation is the following:

- for every λ estimate the generalization error;

- choose the λ which minimizes the estimate of generalization error;

Generalization error

If f_0 is the function underlying the data, the generalization error is

$$E_{\text{gen}} = \int_{R^d} d\mathbf{x} \; (f_0(\mathbf{x}) - f_\lambda(\mathbf{x}))^2 P(\mathbf{x})$$

Notice that

$$E_{\text{gen}} \equiv I[f_\lambda]$$

and can be estimated by

$$E_{\text{gen}} \approx \frac{1}{l} \sum_{i=1}^{l} (y_i - f_\lambda(\mathbf{x}_i))^2$$

for very large l.

Cross validation

The generalization error can be estimated in several ways (leave-one-out, n-fold cross-validation ...).

The main idea is not to use all the data set, but to keep a part of it as a test set on which the performances of the approximation for that given value of λ are tested.

A simple experiment

$f(x,y) = e^{-x} * \sin(5 * y) + \cos(4 * x) * e^{-y^2}$

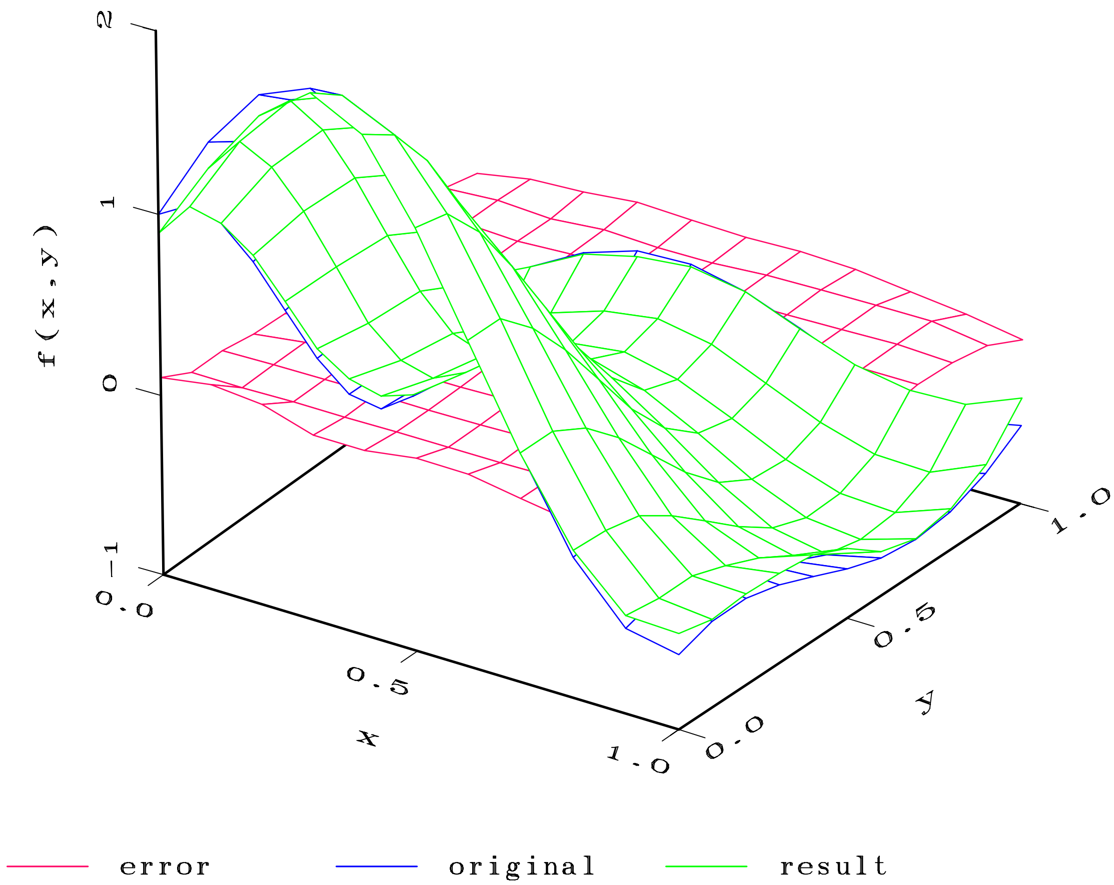
100 data points in $[0,1] \times [0,1]$;

test on a 10×10 grid (121 points);

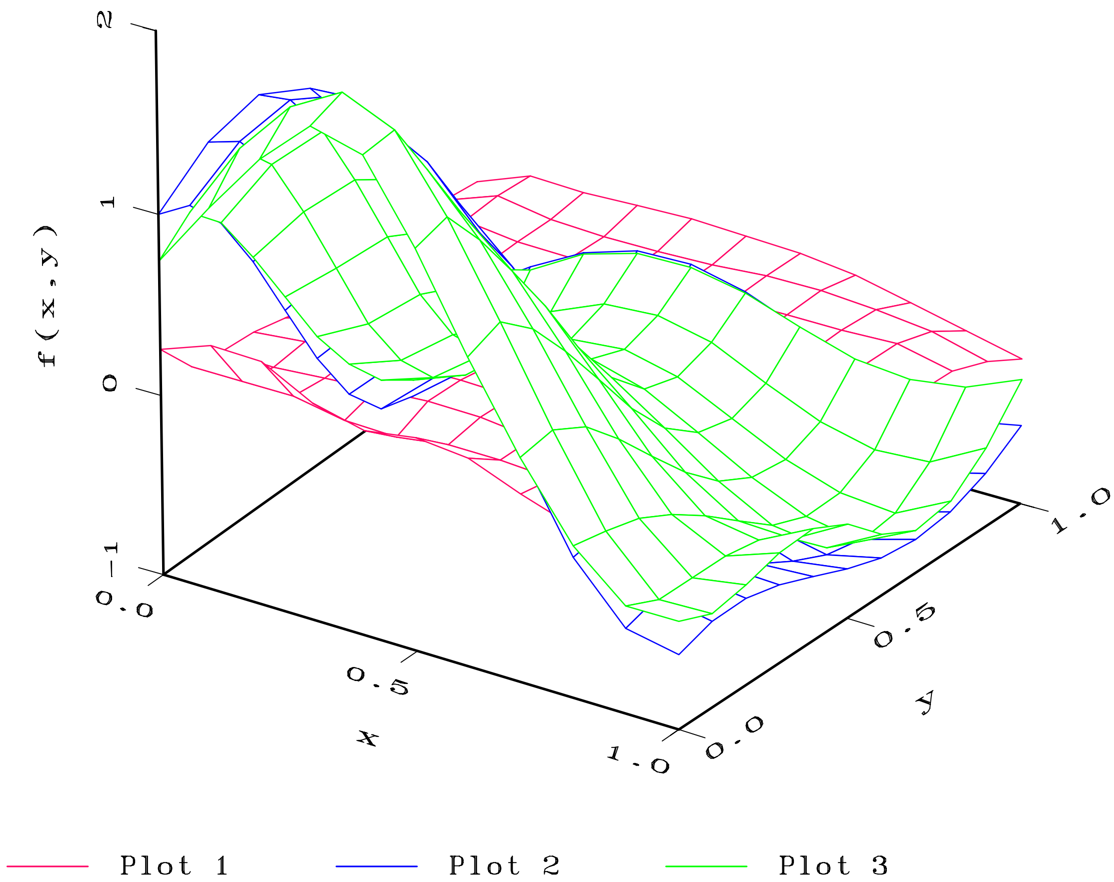
uniform noise added with several intensities (0, 0.01, 0.05, 0.2);

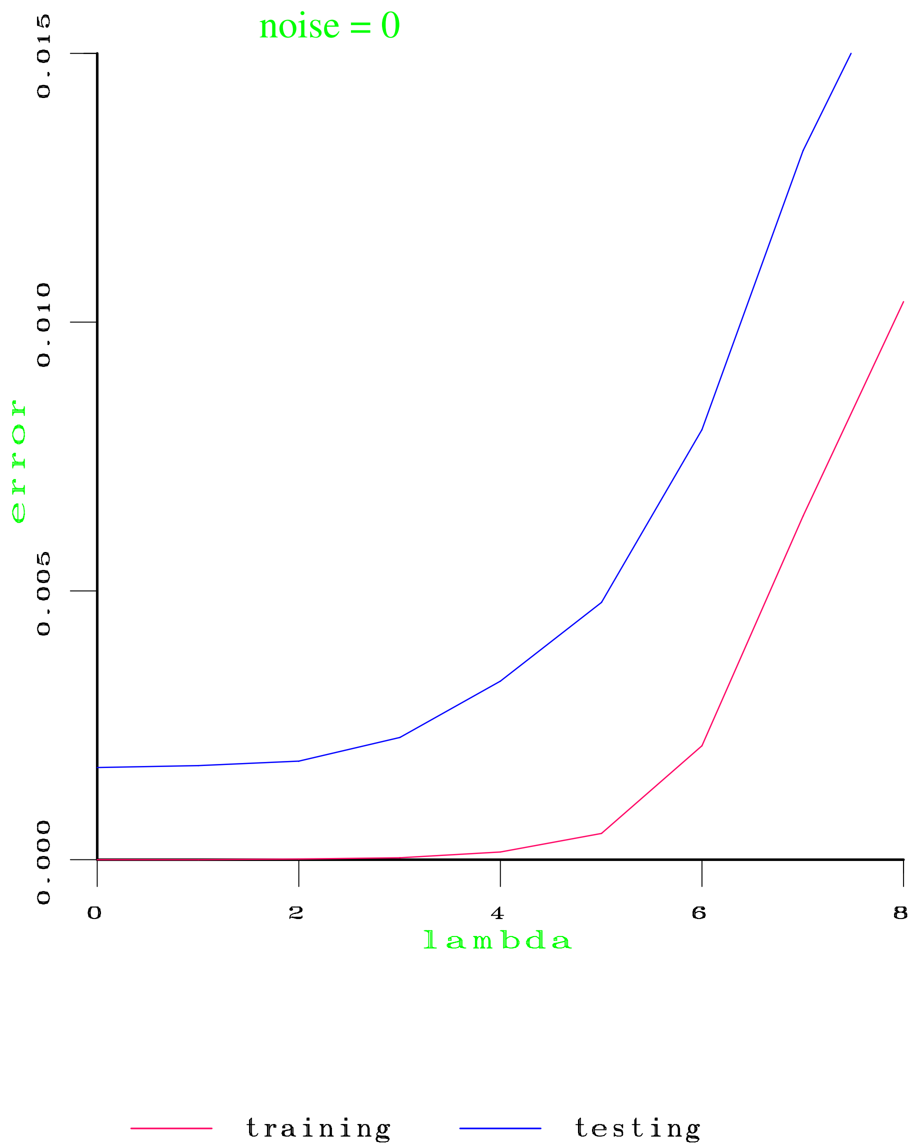
9 values of λ: $(0, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 0.1, 0.5, 1)$

lambda = 0.001

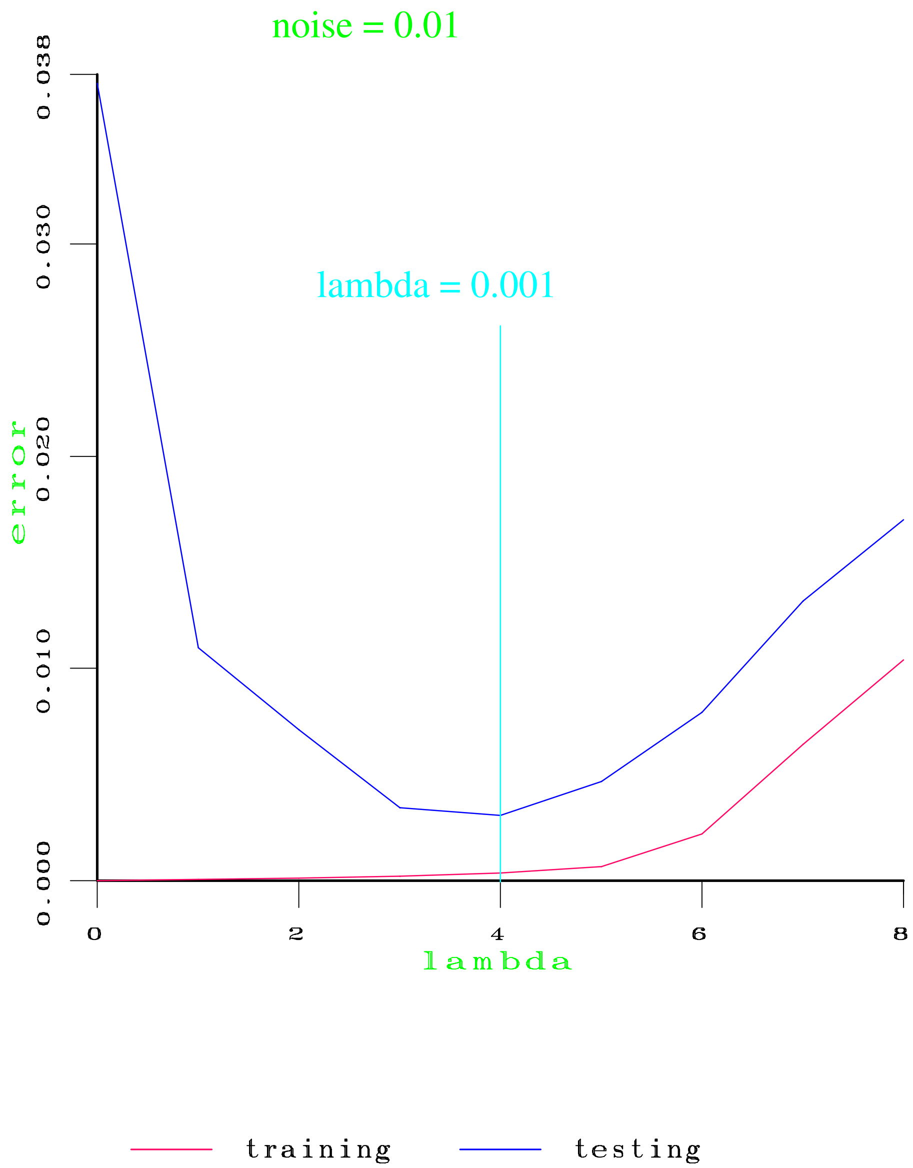


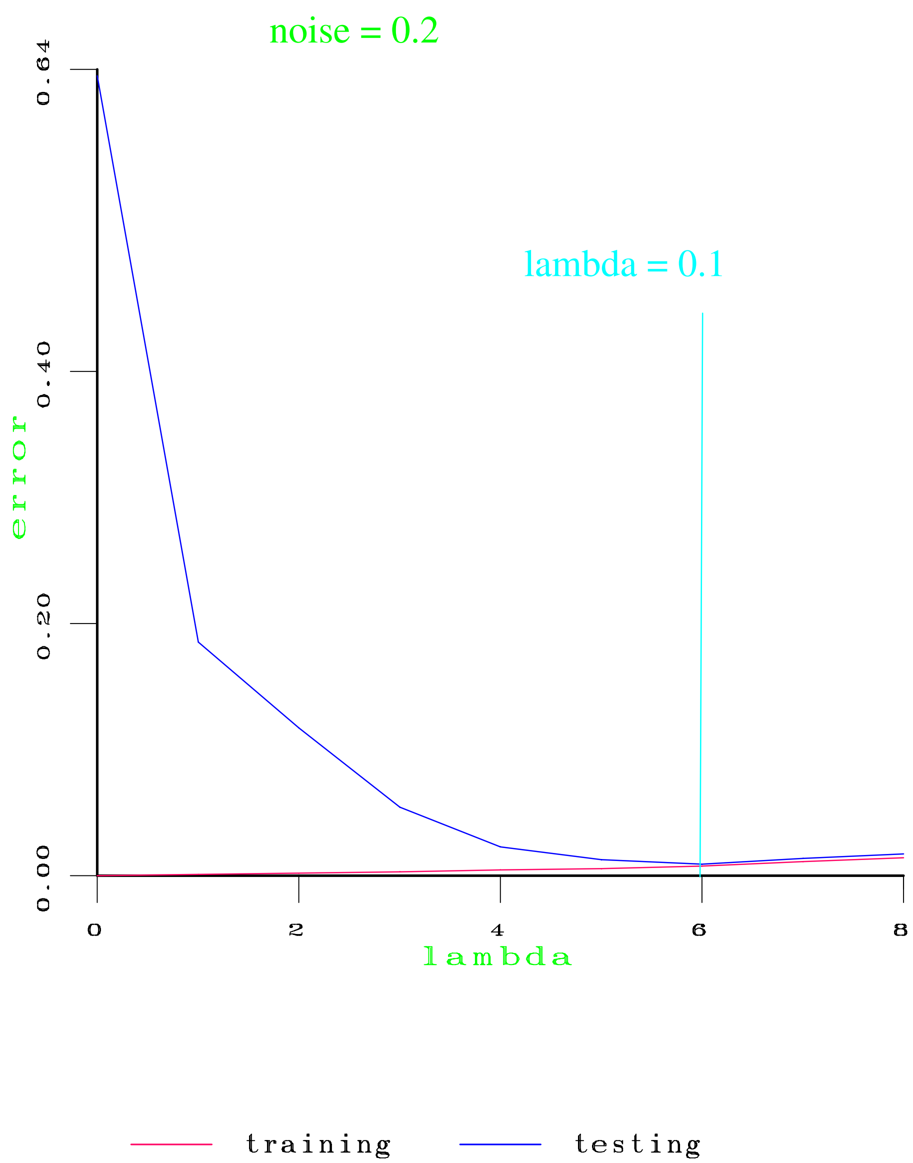
lambda = 0.1



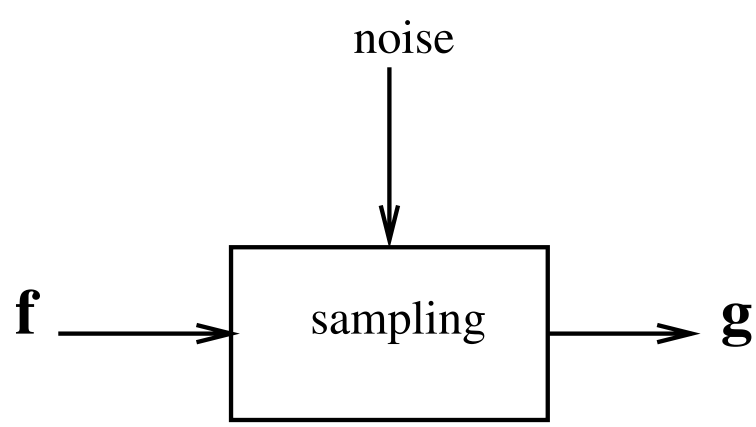


noise = 0
0.000
0.005
0.010
0.015
error
0
2
4
6
8
lambda
training
testing


noise = 0.01
lambda = 0.001
error
lambda
0.000
0.010
0.020
0.030
0.038
0
2
4
6
8
training
testing


noise = 0.2
lambda = 0.1
error
0.64
0.40
0.20
0.00
0
2
4
6
8
lambda
training
testing

A digression: a Bayesian approach to the approximation problem



***f*: original surface**

***g*: data**

The following quantities are defined:

- $\mathcal{P}[f|g]$: the a posteriori probability of the surface f given the data g. It is the quantity of interest.

- $\mathcal{P}[g|f]$: the probability of the data g given the surface f. It is a model of the noise.

- $\mathcal{P}[f]$: the *a priori* probability of the surface f. It depends on the a priori knowledge on the surface f.

Bayes' theorem yields:

$$\mathcal{P}[f|g] \propto \mathcal{P}[g|f] \ \mathcal{P}[f].$$

A probabilistic approach ... (continue)

If the noise is Gaussian the probability $\mathcal{P}[g|f]$ can be written as:

$$\mathcal{P}[g|f] \propto e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{l}(y_i - f(\mathbf{x}_i))^2}$$

where σ is the variance of the noise.

Let $\phi[f]$ be a smoothness functional. We write $\mathcal{P}[f]$ as

$$\mathcal{P}[f] \propto e^{-\lambda\phi[f]}$$

This form of probability distribution gives high probability only to those functions for which the smoothness functional $\phi[f]$ is small, embodying the a priori knowledge that we have on the system.

A probabilistic approach . . . (continue)

Using Bayes theorem the a posteriori probability of f can be written as

$$\mathcal{P}[f|g] \propto e^{-H[f]}$$

where

$$H[f] = \frac{1}{2\sigma^2} \sum_{i=1}^{l} (y_i - f(\mathbf{x}_i))^2 + \lambda\phi[f] \ .$$

The function that maximizes the a posteriori probability $\mathcal{P}[f|g]$ is the one that minimizes the functional $H[f]$.

Regularization solution = Maximum A Posteriori (MAP) estimate

End of digression

The function G does not have to be necessarily radial.

Depending on the choice of G several schemes can be obtained

1. Radial Basis Functions:

$$G(\mathbf{x}) = G(\|\mathbf{x}\|)$$

2. tensor product:

$$G(\mathbf{x}) = G(x^1)G(x^2)\cdots G(x^d)$$

3. additive:

$$G(\mathbf{x}) = G(x^1) + G(x^2) + \cdots + G(x^d)$$

Additive Models

It is also possible to derive from regularization theory, under other smoothness assumptions, models of the form

$$f(\mathbf{x}) = \sum_{\mu=1}^{d} f_\mu(x^\mu)$$

where

$$\mathbf{x} \equiv (x^1, \ldots, x^\mu, \ldots x^d)$$

and each $f_\mu(x^\mu)$ is a one-dimensional Radial Basis Functions.

$$f_\mu(x^\mu) = \sum_{i=1}^{l} c_i^\mu G(x^\mu - x_i^\mu)$$

(Wahba, 1990; Hastie and Tibshirani, 1986, 1990; Girosi, Jones and Poggio, 1993)

A step back: linear approximation

In statistics the most common model is linear:

$$y = \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \ldots + \beta$$

- it provides a simple description of the data;
- it summarizes the contribution of each variable with a single coefficient;

How can it be generalized?

Additive approximation

In an additive approximation scheme we assume that

$$y = \sum_{\mu=1}^{d} f_\mu(x^\mu)$$

Additive models retain the important feature that the effects of the variables can be analyzed separately.

Additive models

Additive models are appealing because a multivariate function is approximated by the sum of one-dimensional functions:

$$f(\mathbf{x}) = \sum_{\mu=1}^{d} f_\mu(x^\mu)$$

The complexity of such models is much less than the one of fully d-dimensional models.

Additive models and regularization theory

What is the smoothness functional corresponding to additive splines?

$$G(\mathbf{x}) = G(x^1) + G(x^2) + \cdots + G(x^d) \quad \Rightarrow$$

$$\Rightarrow \phi[f] = \int_{R^d} d\mathbf{s} \ \frac{|\tilde{f}(\mathbf{s})|^2}{\tilde{G}(\mathbf{s})}$$

$$\tilde{G}(\mathbf{s})?????$$

The Fourier transform of an additive function is a not a nice object ...

Suppose $G = G(x) + G(y)$. Then the Fourier transform is

$$\tilde{G}(s_x, s_y) = \delta(s_y)\tilde{G}(s_x) + \delta(s_x)\tilde{G}(s_y)$$

The reason being that:

$$\int_R \; dxdy \; e^{-i(s_x x + s_y y)} G(x) =$$

$$= \left[\int_R \; dy e^{-i s_y y}\right] \left[\int_R \; dx \; e^{-i s_x x} G(x)\right] = \delta(s_y)\tilde{G}(s_x)$$

Almost **additive models**

If we choose

$$\tilde{G}_\epsilon(s_x, s_y) = \frac{1}{\epsilon}\left(\tilde{g}(s_x)e^{-(\frac{s_y}{\epsilon})^2} + \theta_y \tilde{g}(s_y)e^{-(\frac{s_x}{\epsilon})^2}\right)$$

the corresponding basis function becomes:

$$G_\epsilon(x, y) = g(x)e^{-\epsilon^2 y^2} + g(y)e^{-\epsilon^2 x^2} \ .$$

In the limit of ϵ going to zero G_ϵ becomes an additive basis function.

In the standard approach to additive models one has:

$$f(\mathbf{x}) = \sum_{\mu=1}^{d} f_\mu(x^\mu)$$

$$f_\mu(x^\mu) = \sum_{i=1}^{l} c_i^\mu G(x^\mu - x_i^\mu)$$

with the constraint that

$$c_i^\mu = c_i^\nu \quad \forall \mu, \nu = 1 \ldots d$$

but in practice other choices have been made.

It is common to use the model

$$f(\mathbf{x}) = \sum_{\mu=1}^{d} f_\mu(x^\mu)$$

$$f_\mu(x^\mu) = \sum_{i=1}^{l} c_i^\mu G(x^\mu - x_i^\mu)$$

where the c_i^μ are all independent, but to adopt an iterative training technique called *backfitting*.

Backfitting algorithm

- Initialize: $f_\mu(x^\mu) = 0$;
- cycle: $\mu = 1, \ldots, d, 1, \ldots d, \ldots$
- define

$$r_i^\mu = y_i - \sum_{\nu \neq \mu}^{d} f_\nu(x_i^\nu) = y_i - [f(\mathbf{x}_i) - f_\mu(x_i^\mu)]$$

and find $f_\mu(x^\mu)$ that solves

$$\min_{f_\mu} \sum_{i=1}^{l} (r_i^\mu - f_\mu(x_i^\mu))^2 \ ;$$

- continue until the individual functions do not change anymore.

Step by step ...

- $\mu = 1, \Rightarrow r_i^1 = y_i$

- $f_1(x^1) = \sum_{i=1}^{l} c_i^1 G(x^1 - x_i^1) \qquad \Rightarrow$
  $\Rightarrow$ solve the problem

$$\min_{\mathbf{c}^1} \sum_{j=1}^{l} (y_j - \sum_{i=1}^{l} c_i^1 G(x_j^1 - x_i^1))^2$$

- $\mu = 2, \Rightarrow r_i^2 = y_i - f_1(x_i^1)$

- $f_2(x^2) = \sum_{i=1}^{l} c_i^2 G(x^2 - x_i^2) \qquad \Rightarrow$
  $\Rightarrow$ solve the problem

$$\min_{\mathbf{c}^2} \sum_{j=1}^{l} (y_j - \sum_{i=1}^{l} c_i^2 G(x_j^2 - x_i^2))^2$$

- $\ldots$

Iterative algorithms of this form are appealing because at each step a small minimization problem is solved.

More in class 4 ...