

Menu

1. The problem of approximation as an ill-posed problem;
2. Regularization theory: basic ideas;
3. The variational approach to the approximation problem;
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 $D = \{(\mathbf{x}_i, y_i) \in X \times Y\}_{i=1}^N$ is collected as following:

For $i=1 \rightarrow N$

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 find an hypothesis F^* that minimizes the expected risk:

$$\begin{aligned} F^*(\mathbf{x}) &= \arg \min_{F \in \mathcal{H}} Q[F] = \\ &= \arg \min_{F \in \mathcal{H}} \int_{X \times Y} (y - F(\mathbf{x}))^2 P(\mathbf{x}, y) d\mathbf{x} dy \end{aligned}$$

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

Notice that

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

where

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

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

1. Learning means:

$$\min_{F \in \mathcal{H}} Q[F]$$

2.

$$\min_{F \in \mathcal{H}} Q[F] = \min_{F \in \mathcal{H}} \int_X (y(\mathbf{x}) - F(\mathbf{x}))^2 P(\mathbf{x}) d\mathbf{x}$$

3. Learning \equiv to approximate the regression function $y(\mathbf{x})$, which is an unknown function belonging to some large space of functions, also called the *target space* \mathcal{T} .

From now on we will always assume that the data set $D = \{(\mathbf{x}_i, y_i) \in X \times Y\}_{i=1}^N$ has been obtained sampling some function $h(\mathbf{x})$ in presence of additive noise:

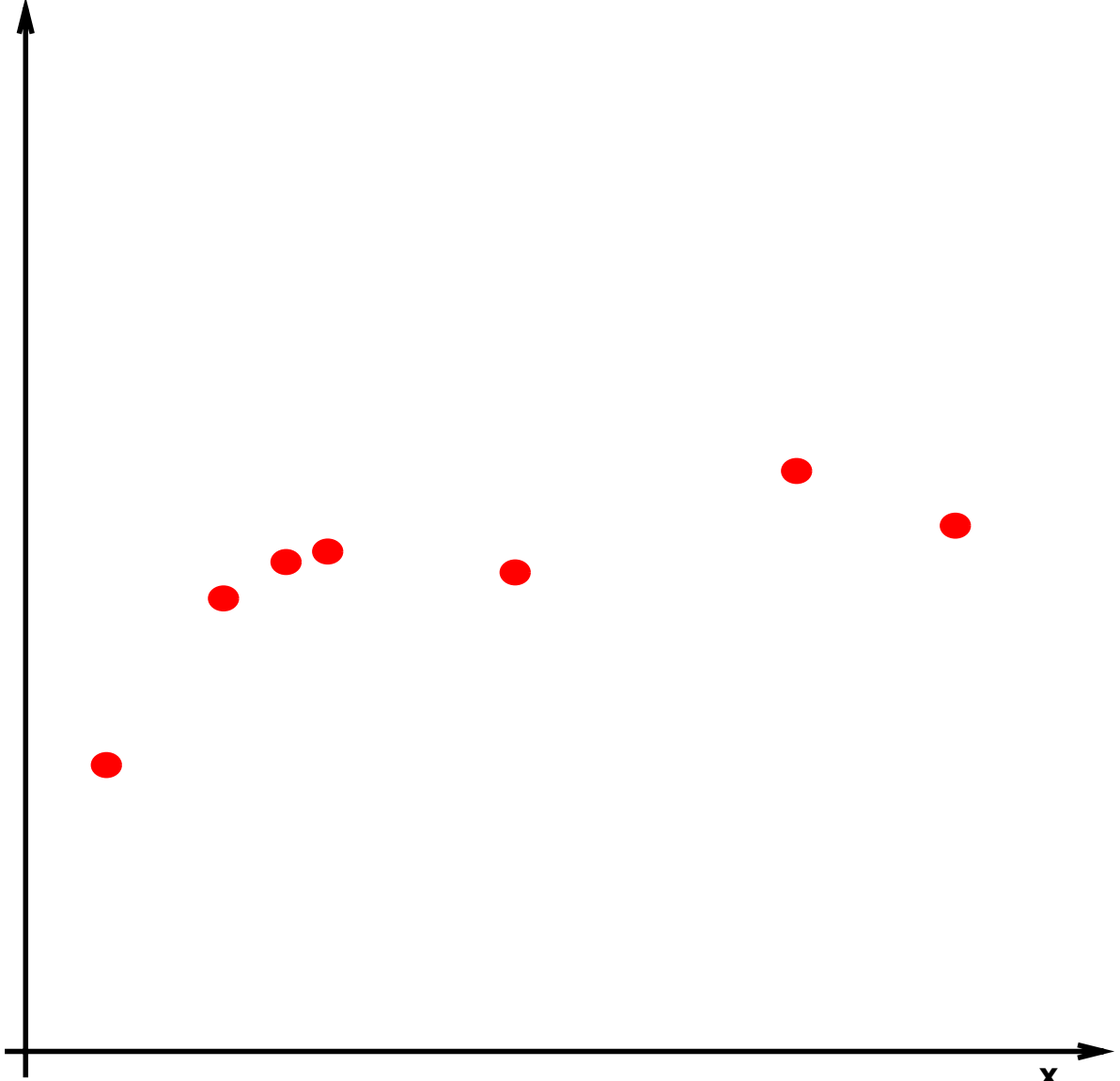
$$y_i = h(\mathbf{x}_i) + \epsilon_i$$

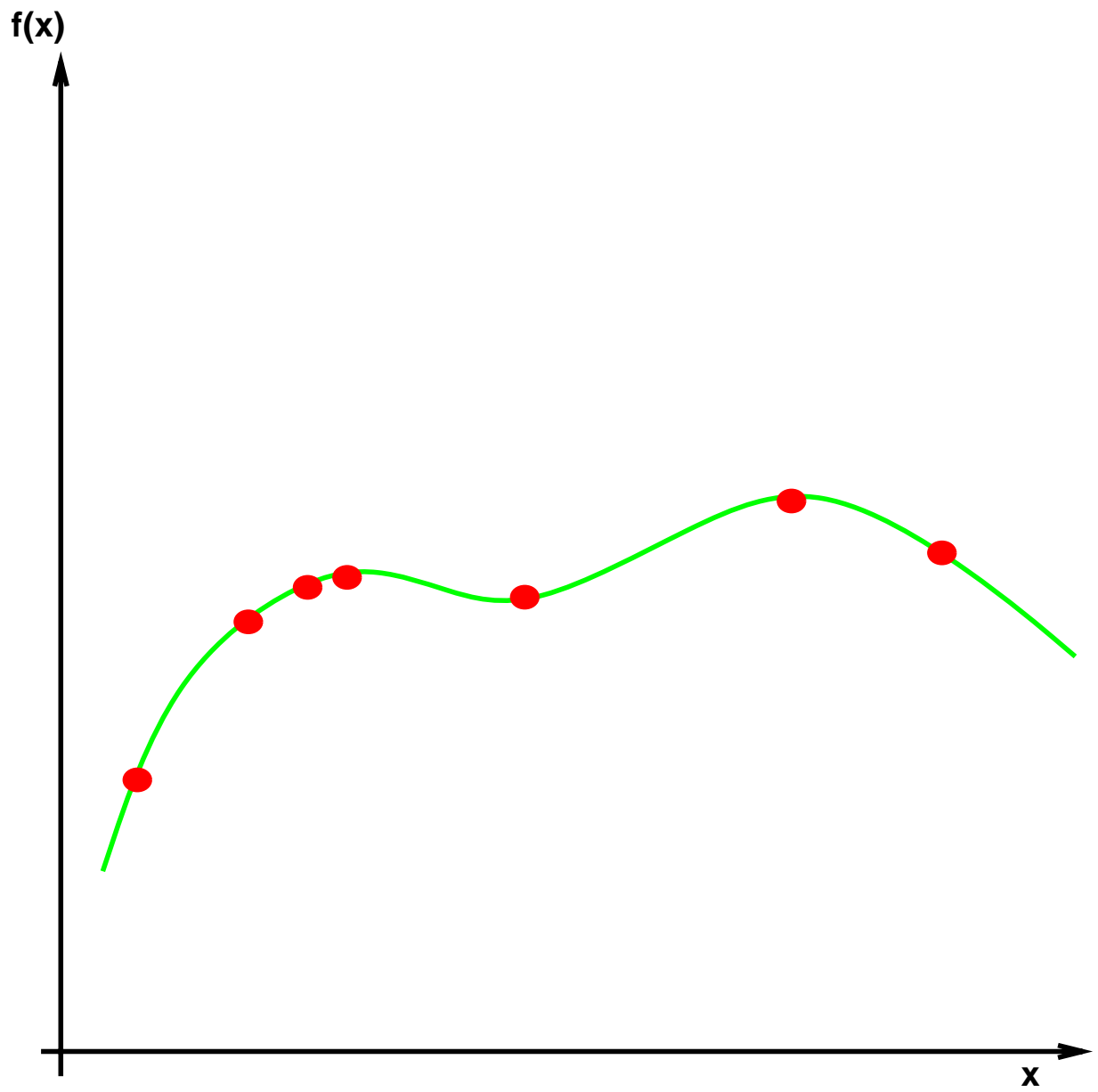
where ϵ_i are i.i.d. random variables with zero mean and probability distribution $\mathcal{P}(\epsilon)$.

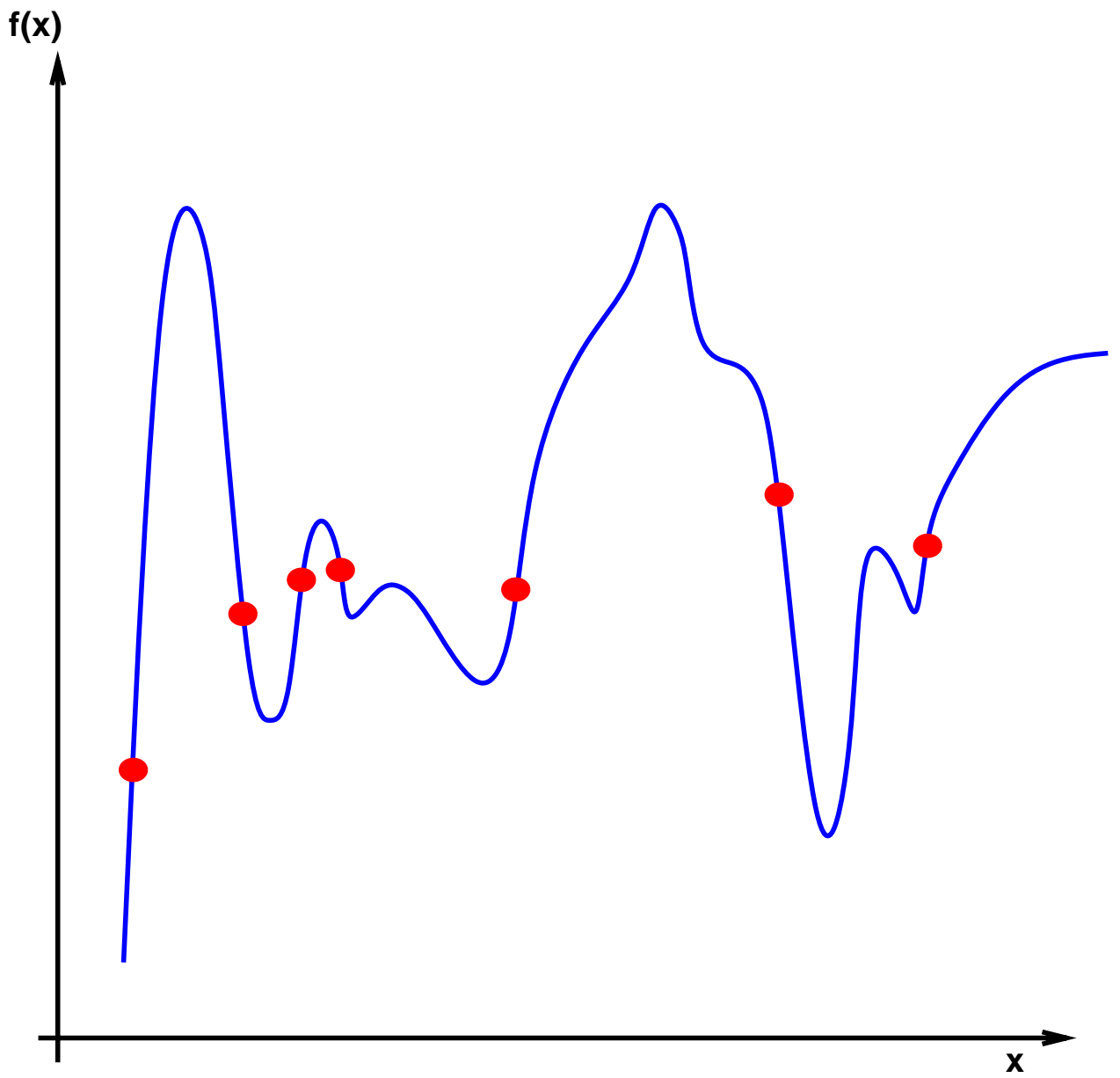
In other words:

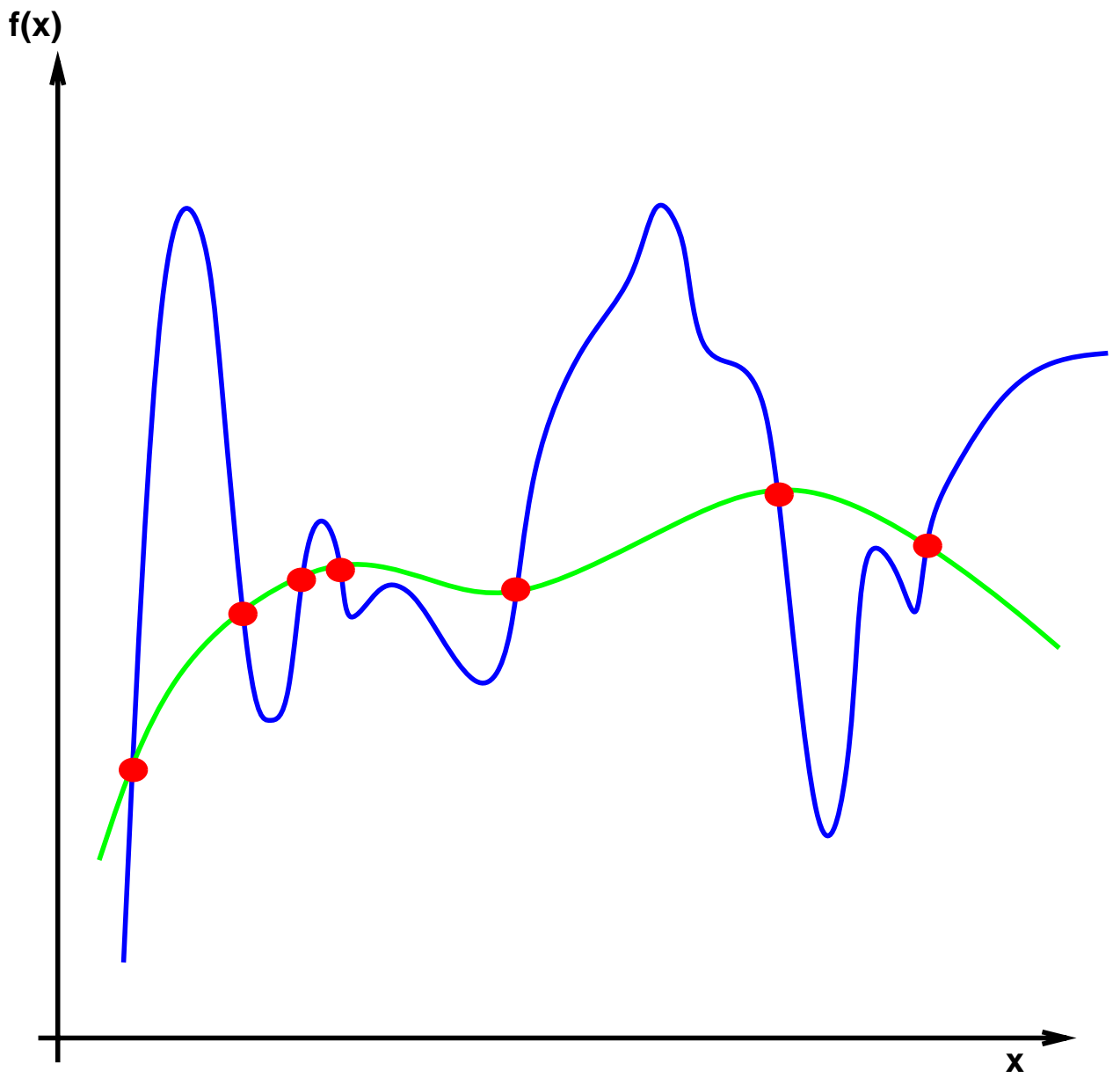
$$P(y|\mathbf{x}) = \mathcal{P}(y - h(\mathbf{x}))$$

$f(x)$









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**).

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.

An ill-posed direct problem: differentiation

$$g(x) = f(x) + \epsilon \sin(\omega x)$$

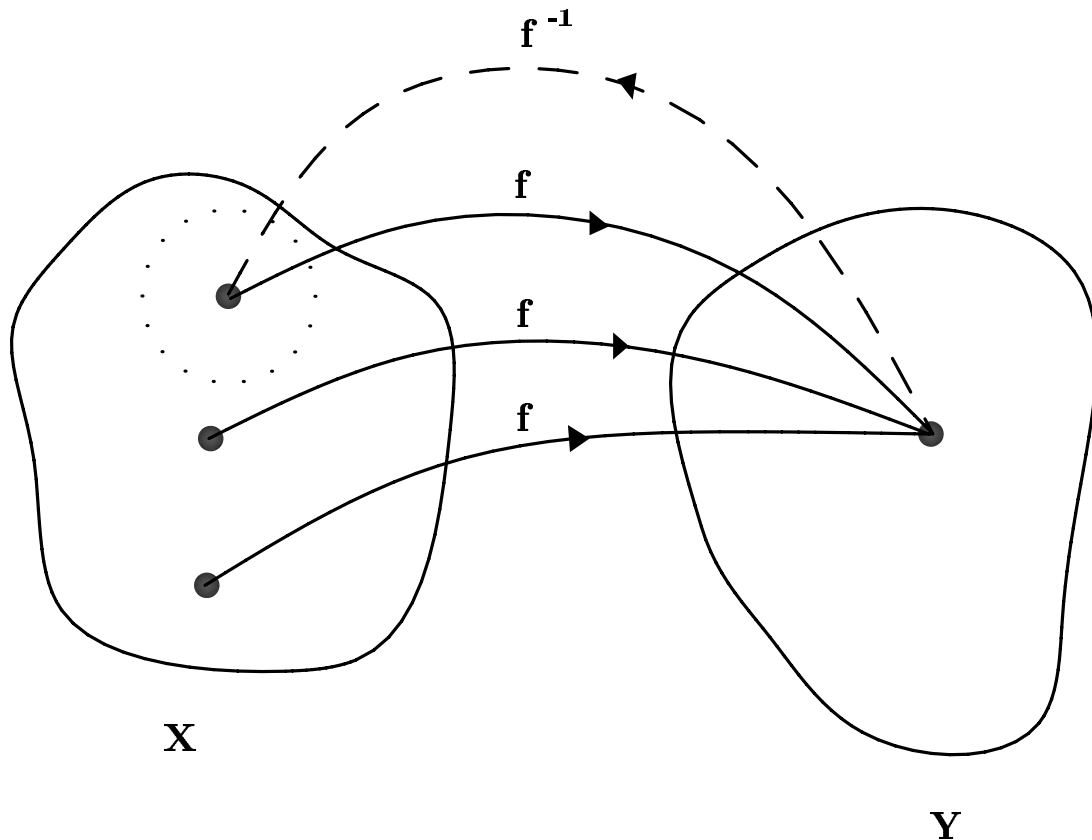
$$g'(x) = f'(x) + \epsilon \omega \cos(\omega x)$$

$$\max_{x \in R} |f(x) - g(x)| = \epsilon$$

$$\max_{x \in R} |f'(x) - g'(x)| = \epsilon \omega$$

Inverse problems

When the solution of the problem is not unique, it is possible to obtain a unique solution *restricting* the space of solutions in appropriate way. This can be done for example imposing some constraint on the possible solution. The shape of the constraint depends on a priori knowledge on the solution.



———— $f =$ direct map

- - - $f^{-1} =$ regularized inverse map

The problem of approximating a function from sparse data has an infinite number of solutions.

Some restrictions has to be imposed on the possible solutions, in order to make the problem well posed.

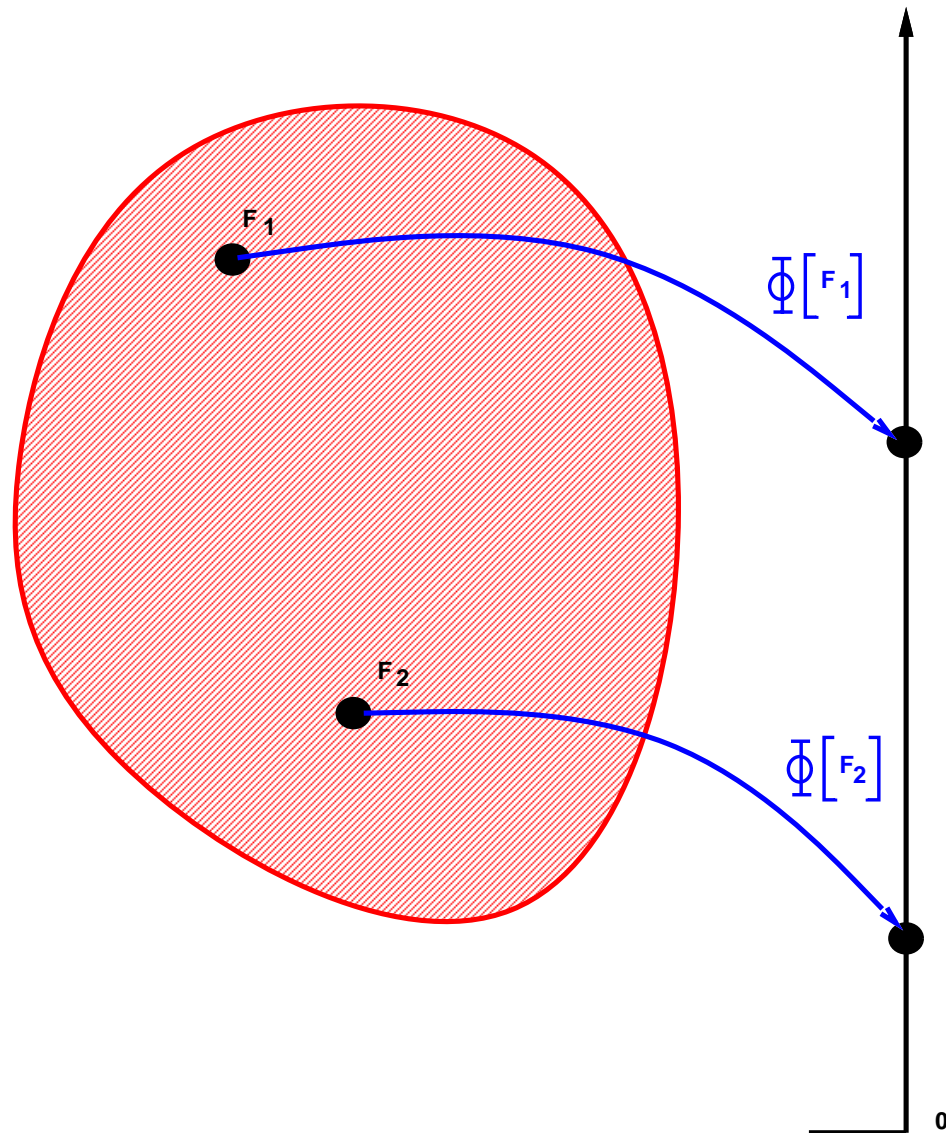
The kind of restrictions depend on the *a priori knowledge* on the problem.

**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 to say whether a function is “nicer” than another function?



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 with little oscillation.

Regularization theory for interpolation

Among all the functions that satisfy

$$\sum_{i=1}^N (y_i - f(\mathbf{x}_i))^2 = 0$$

choose the one that minimizes a smoothness functional $\phi[f]$.

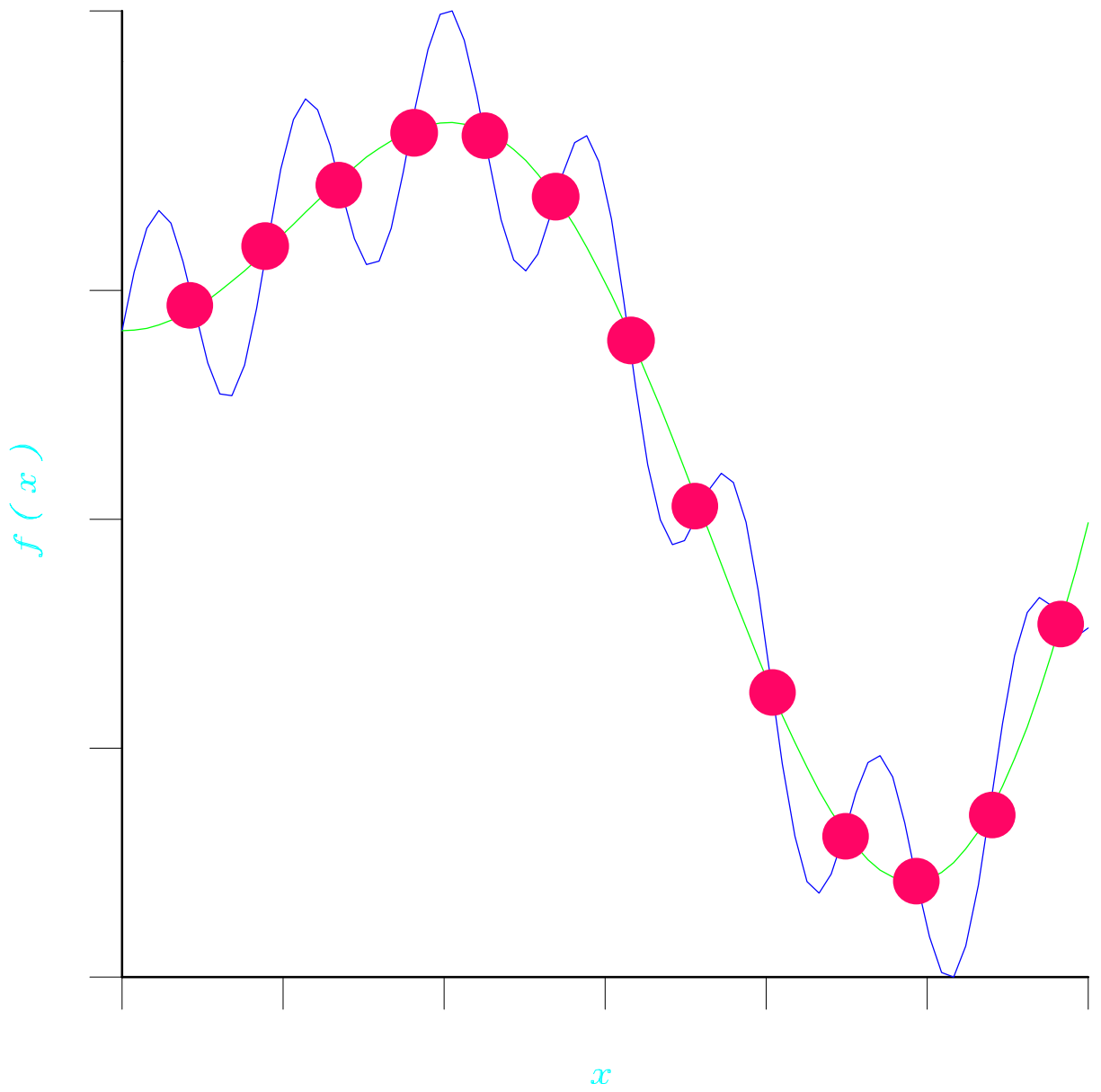
Regularization theory for approximation

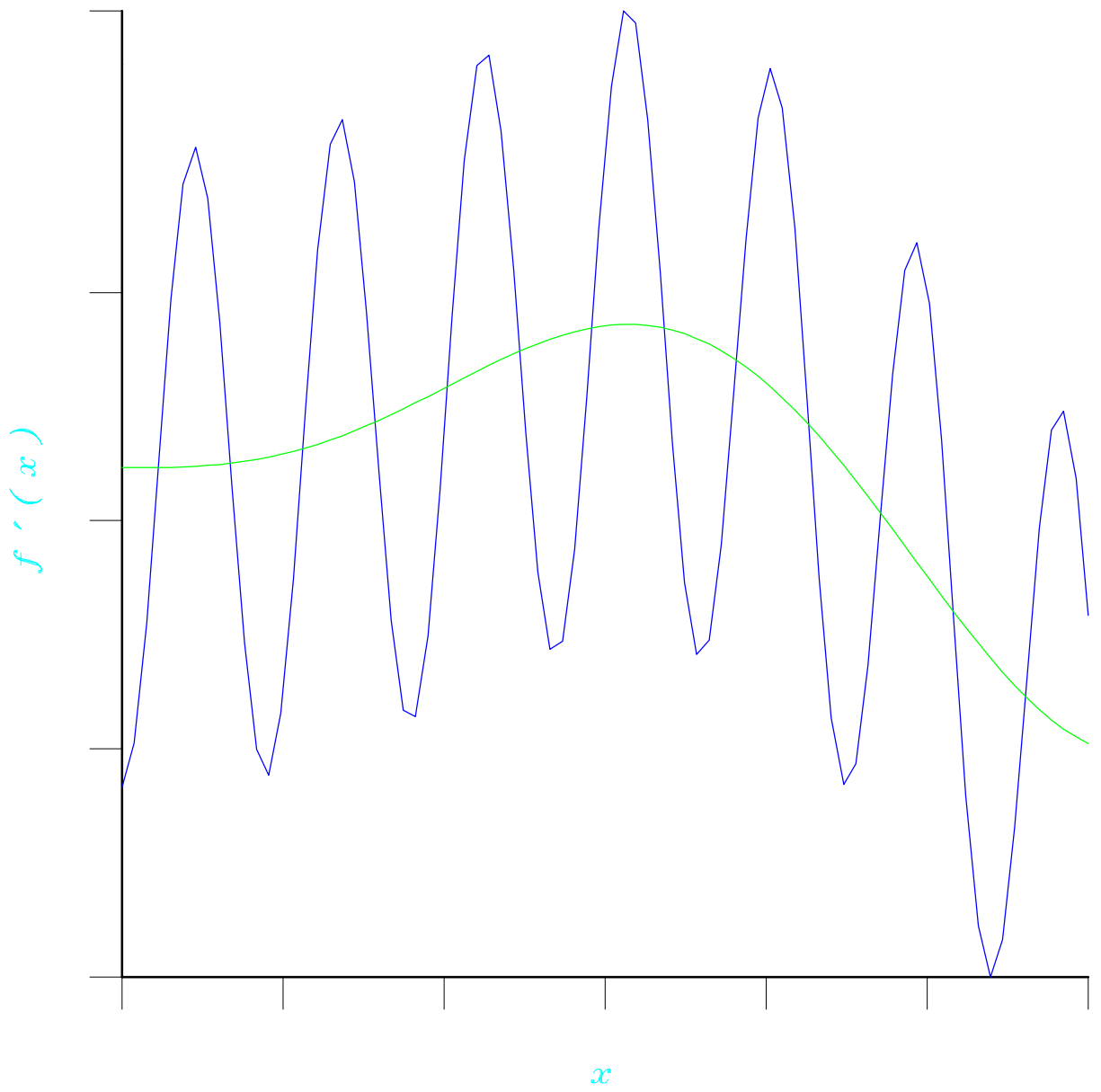
Minimize the functional

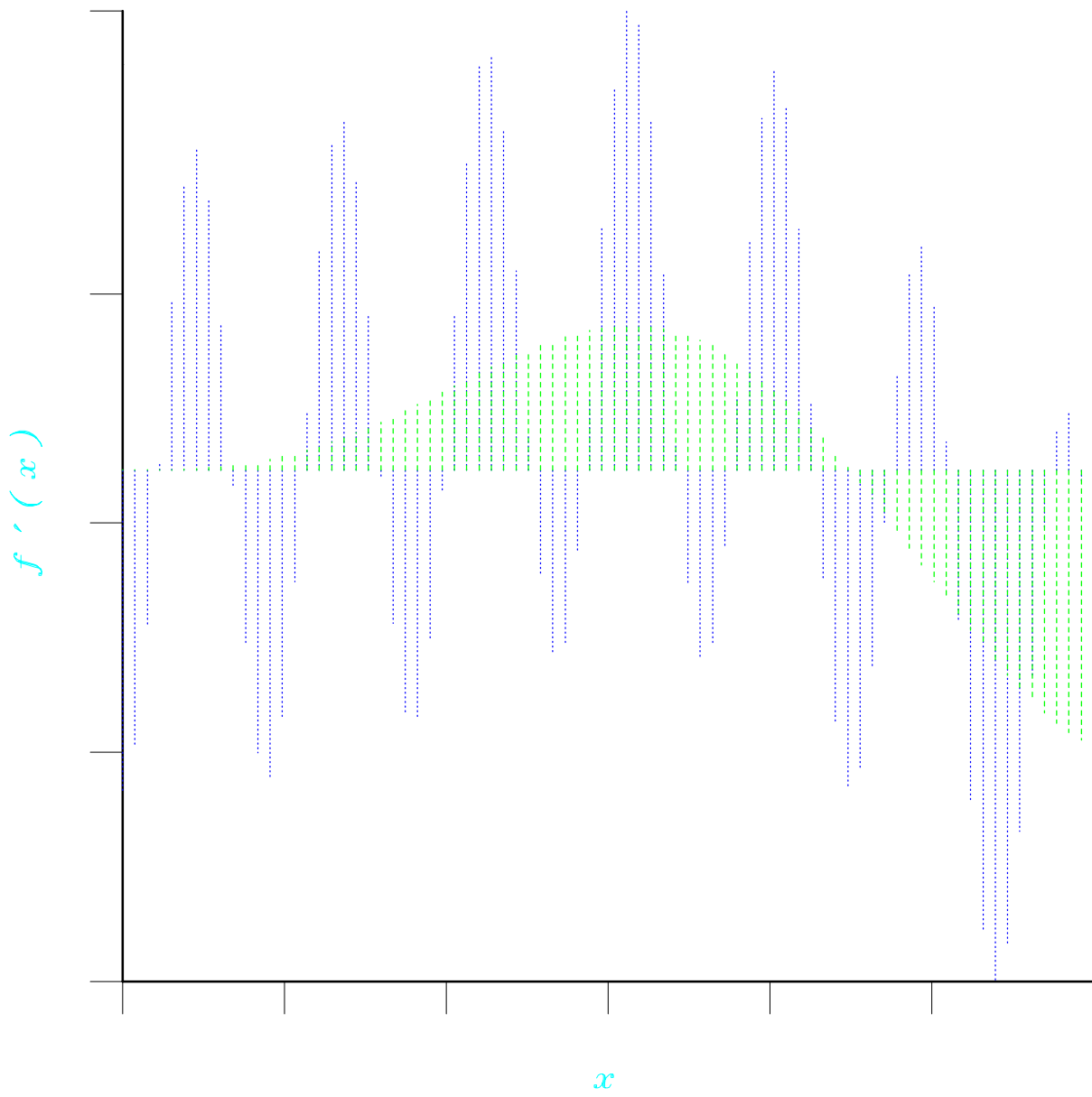
$$H[f] = \sum_{i=1}^N (y_i - f(\mathbf{x}_i))^2 + \lambda \phi[f]$$

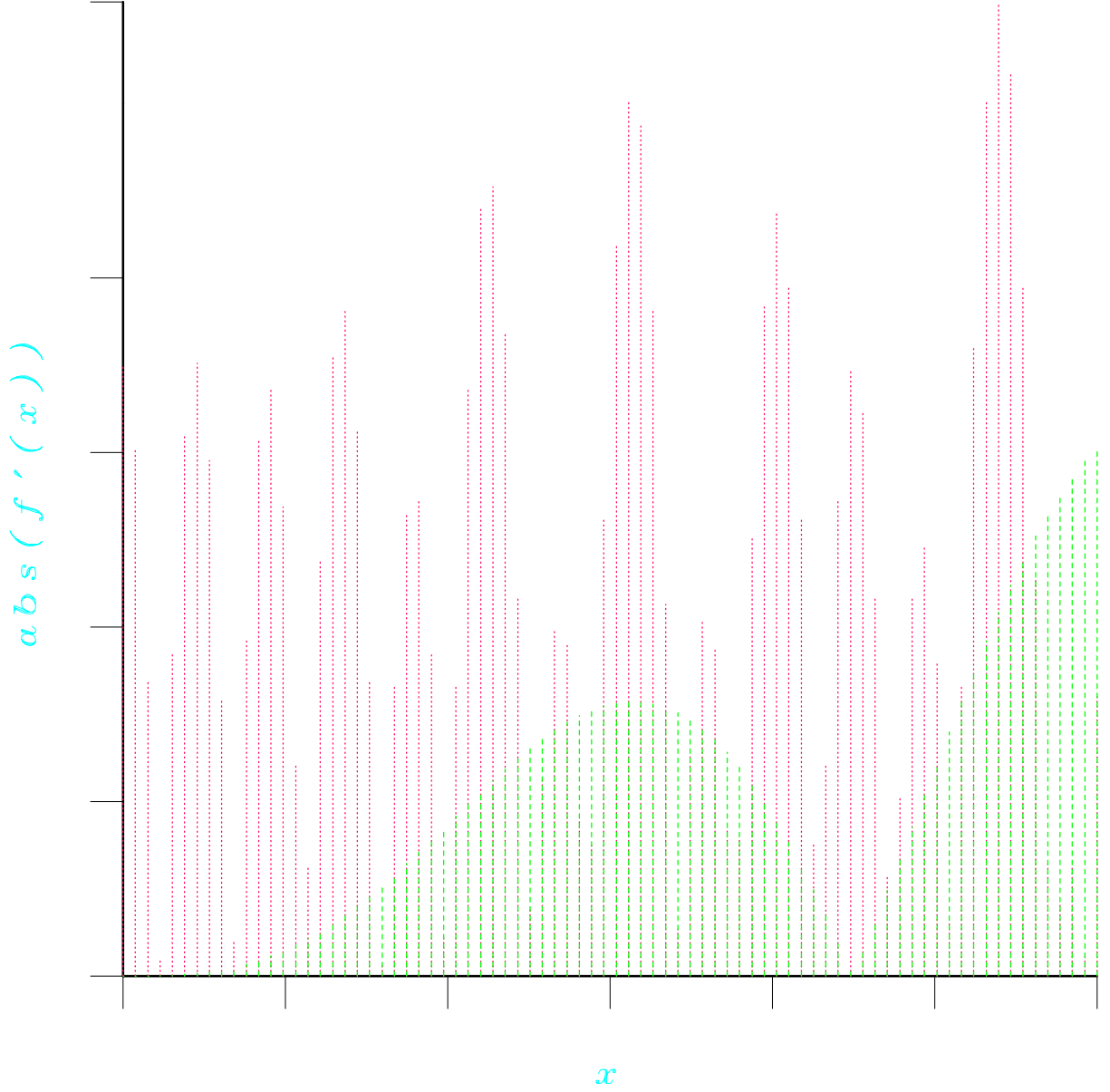
where $\phi[f]$ is a smoothness functional, and λ is a small positive number, called the *regularization parameter*.

λ should be proportional to the amount of noise in the data.

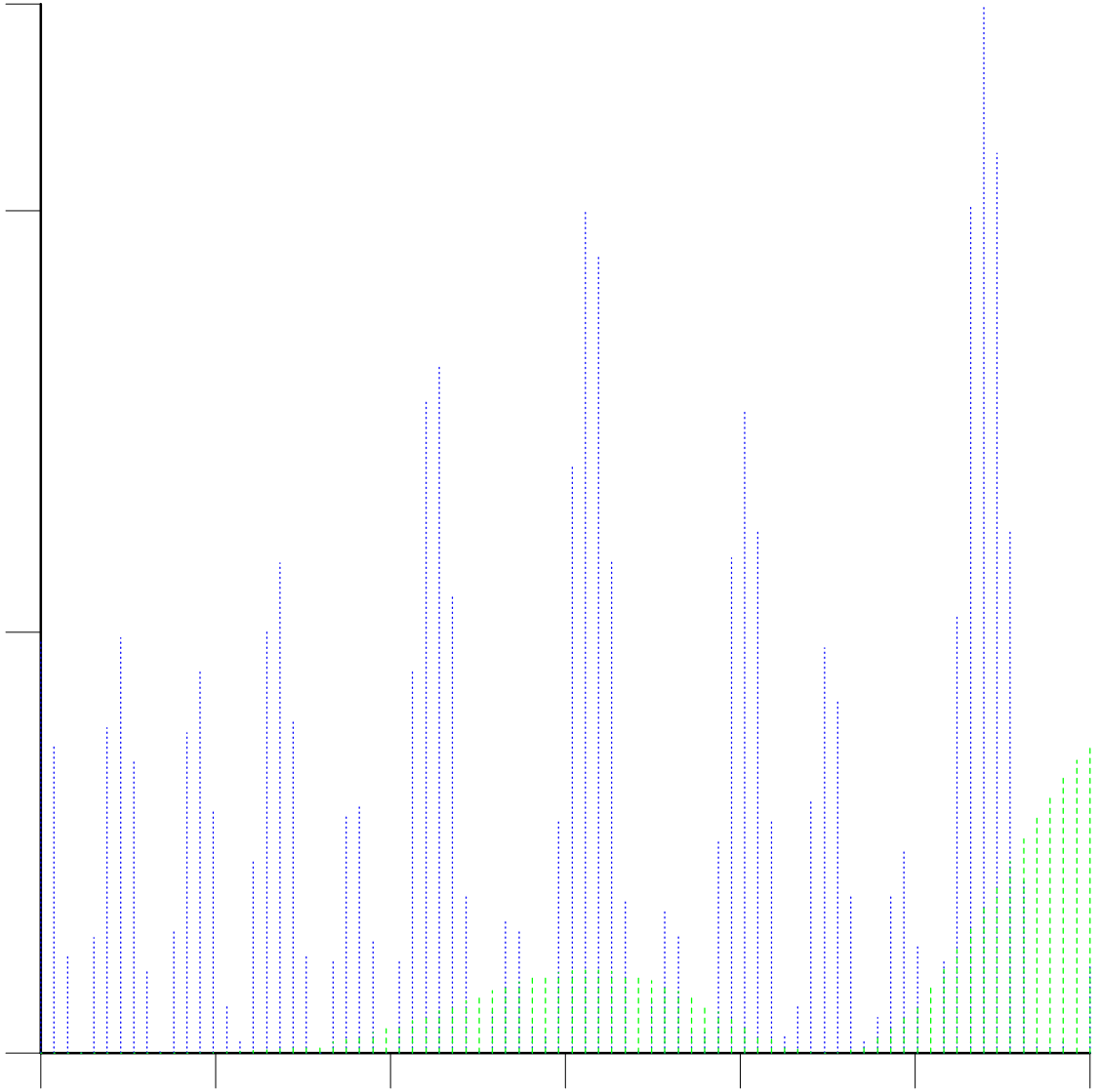








$[f'(x)] \sim 2$



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 $D = \{(\mathbf{x}_i, y_i) \in R^d \times R\}_{i=1}^N$ can be found minimizing the functional:

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

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 $\mathbf{x}_1, \dots, \mathbf{x}_N \in R^n$ and scalars c_1, \dots, c_N such that $\sum_{i=1}^N c_i p(\mathbf{x}_i) = 0$ for all $p \in \pi_{k-1}(R^n)$, the quadratic form

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

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 $\mathbf{x}_1, \dots, \mathbf{x}_N \in R^n$ and scalars c_1, \dots, c_N the quadratic form

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

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

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

has the form:

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

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;