functional methods for learning

alessandro verri

DISI, Università di Genova, Italy

plan

- learning from examples
- connection with inverse problems
- regularization algorithms: tikhonov and iterative methods
- future work

three slides on learning: ingredients

(Vapnik, '98, Girosi and Poggio '90, Cucker and Smale '00)

- the **sample space** $Z = X \times Y$, with X subset of $\mathbb{R}^d$ and Y subset of $\mathbb{R}$
- the **probability measure** $\rho(x, y) = \rho(y|x)\rho_X(x)$ on the sample space Z
- the **training set** $\mathbf{z} = (\mathbf{x}, \mathbf{y}) = \{(x_1, y_1), \cdots, (x_n, y_n)\}$, a sequence of n examples drawn *i.i.d.* according to the probability ρ
- the **hypotheses space** $\mathcal{H}$ is the function space where we look for the solution.

find $f_{\mathbf{z}}$ **such that** $f_{\mathbf{z}}(x_{new}) \sim y_{new}$

three slides on learning: problem formulation

(Cucker and Smale ’00, Györfi et. al. 02)

we want to minimize the expected error

$$\mathcal{E}(f) = \int_{X\times Y} (f(x) - y)^2 d\rho(x, y)$$

the minimizer of the above functional is the regression function

$$f_\rho(x) = \int_Y y\rho(y|x)$$

the problem is approximating f_ρ from $\mathcal{H}$ given the sample $\mathbf{z} \sim \rho$

three slides on learning: consistency

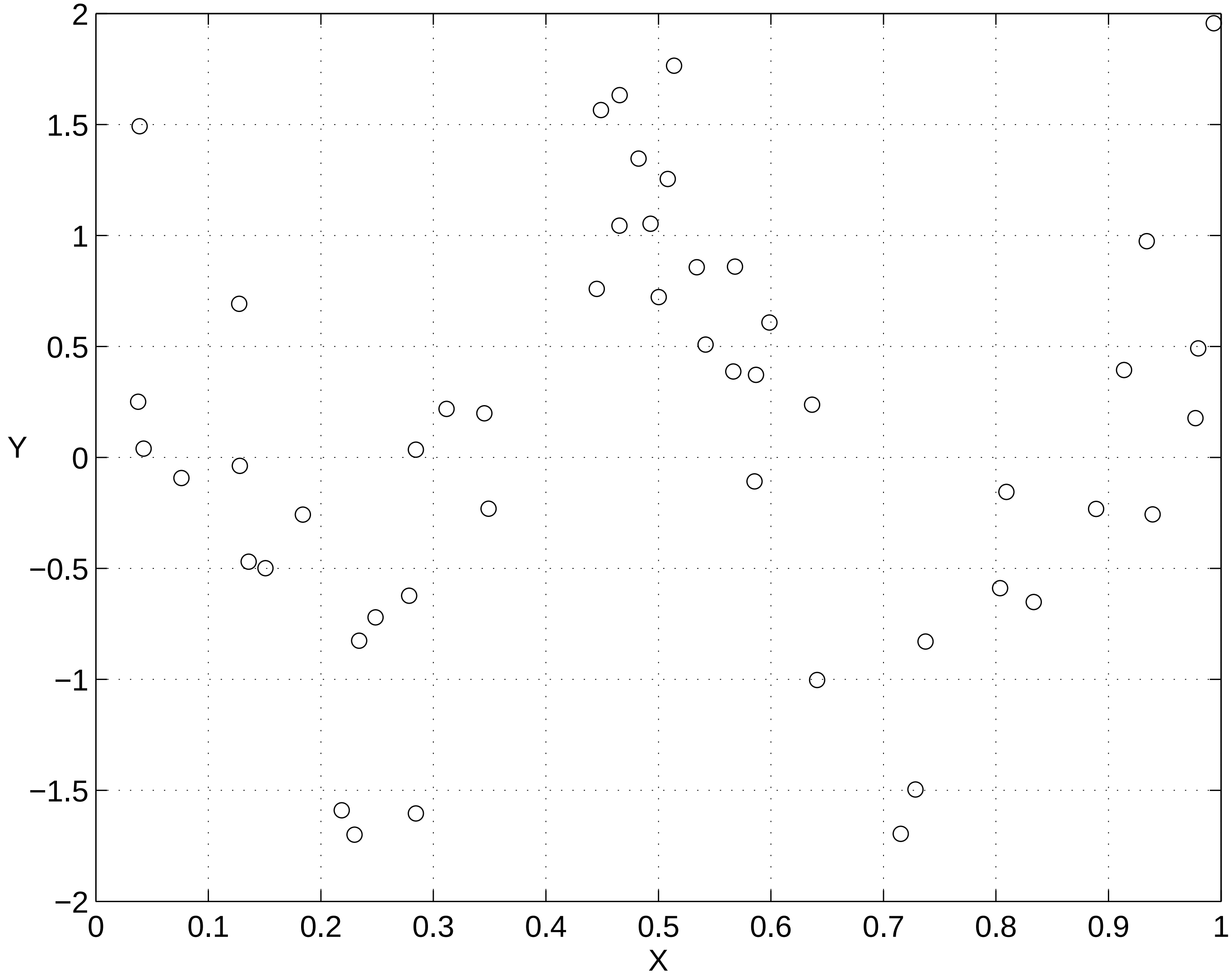
Formally we look for a probabilistic bound for all $\varepsilon > 0$

$$\mathrm{P}\left[\, \mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_\rho) > \varepsilon \right] \leq \eta(\varepsilon, n)$$

and study the rate of the convergence in probability of $\mathcal{E}(f_{\mathbf{z}})$ to $\mathcal{E}(f_\rho)$ as the number of examples increase, namely consistency.

if we want to have convergence rates we need some assumption on the problem, i.e. to restrict the class of possible probability measures.

one dimensional regression



learning as function approximation: few remarks

- noise model: to an input x corresponds a set of outputs distributed according to $\rho(y|x)$, (compare with $y = f_\rho(x) + \xi$)
- the inputs x are not chosen but **sampled** according to ρ_X.
- very few assumption on $\rho(y|x)$ and ρ_X.
- usually dimensionality $= d >> n =$number of data (bioinformatics, image classification, text categorization...).

some references

1. T.Poggio, F. Girosi, **247** Science (1990) 978-982
2. Girosi, M. Jones, T. Poggio, **7** Neural Comp. (1995) 219-269
3. V. Vapnik, *Statistical learning theory*, 1998
4. T. Evgeniou, M. Pontil, T. Poggio, **13** Adv.Comp.Math. (2000) 1-50
5. F. Cucker, S. Smale, Bull. Amer. Math. Soc., **39** (2002) 1-49

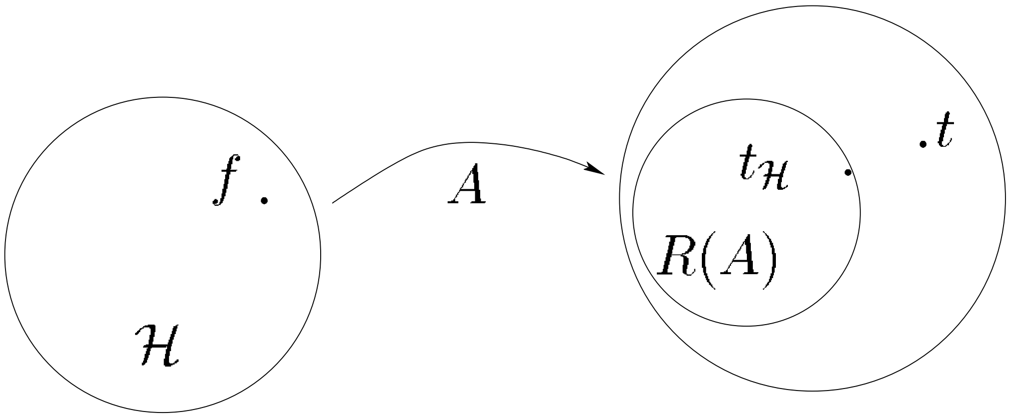
two slides on inverse problems

let $A : \mathcal{H} \to \mathcal{G}$ given $t \in \mathcal{G}$ find f s.t.

$$Af = t$$

the problem can be ill-posed: the solution doesn't exist, is not unique, does not depend continuously on the data. consider the best solution on the hypotheses space that is

$$t_{\mathcal{H}} = \operatorname*{argmin}_{f \in \mathcal{H}} \|Af - t\|^2_{\mathcal{G}}$$



inverse problems (cont.)

the available data are usually affected by noise. in a deterministic model

$$\|t - t_\delta\|_{\mathcal{G}} \leq \delta$$

The generalized solution $t_{\mathcal{H}}$ is not stable w.r.t. noise.

regularization techniques allow to find stable approximation to $t_{\mathcal{H}}$. Tikhonov regularization replaces the least squares problem with

$$\operatorname*{argmin}_{f \in \mathcal{H}} \{\|Af - t_\delta\|_{\mathcal{G}}^2 + \lambda \|f\|_{\mathcal{H}}^2\}$$

more ingredients: hypotheses space

(schwartz '64, aronzajin '50)

we consider reproducing kernel Hilbert spaces (RKHS).

(very roughly) these are Hilbert spaces completely characterized by a (symmetric) positive-definite function $K(x, s)$ namely the kernel.

the following reproducing property holds, if $f \in \mathcal{H}$ and $K_x = K(\cdot, x)$ then

$$\langle f, K_x \rangle_{\mathcal{H}} = f(x)$$

where $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ is the scalar product in $\mathcal{H}$

we assume $\kappa = \sup_{x \in X} \sqrt{K(x, x)} < \infty$

an inverse problems point of view on learning

(De Vito et al. '04)

recall that for $f \in L^2(X, \rho_X)$

$$\mathcal{E}(f) = \|f - f_\rho\|_\rho^2 + \mathcal{E}(f_\rho)$$

consider the inclusion operator $I_K : \mathcal{H} \to L^2(X, \rho_X)$

$$\inf_{f \in \mathcal{H}} \mathcal{E}(f) = \inf_{f \in \mathcal{H}} \|I_K f - f_\rho\|_\rho^2$$

minimizing the expected error is the least square problem associated to the embedding equation

$$I_K f = f_\rho$$

stochastic discretization

(Bertero et al. '85, Girosi and Poggio '89, Smale and Zhou '04 De Vito et al. '04)

given $\mathbf{z} = (\mathbf{x}, \mathbf{y})$, consider the sampling operator $S_{\mathbf{x}} : \mathcal{H} \to \mathbb{R}^d$

$$(S_{\mathbf{x}} f)_i = f(x_i)$$

we have

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - y_i)^2 = \min_{f \in \mathcal{H}} \|S_{\mathbf{x}} f - \mathbf{y}\|_d^2$$

minimizing the empirical error is the least square problem associated to the problem

$$S_{\mathbf{x}} f = \mathbf{y} \iff f(x_i) = y_i \quad i = 1, \ldots, n$$

learning problem revisited...

(De Vito, Caponetto, Odone, De Giovannini, and Rosasco '04)

we are interested to a linear inverse problem and its discretization

$$I_K f = f_\rho \qquad\qquad S_{\mathbf{x}} f = \mathbf{y}$$

since we don't control the discretization we demand the regularization algorithmn to take care of such indetermination

convergence: given a solution $f_{\mathbf{z}} \in \mathcal{H}$ we want the residual to converge to zero (in probability) as the number of samples increases in fact

$$\mathcal{E}(f_{\mathbf{z}}) - \inf_{f \in \mathcal{H}} \mathcal{E}(f) = \|I_K f_{\mathbf{z}} - P f_\rho\|_\rho^2$$

where P is th projection onto the closure of $\mathcal{H}$ in $L^2(X, \rho_X)$

regularization algorithms for learning

tikhonov regularization

$$f_{\mathbf{z}}^{\lambda} = (S_{\mathbf{x}}^* S_{\mathbf{x}} + \lambda I)^{-1} S_{\mathbf{x}}^* \mathbf{y}$$

landweber iteration

$$f_{\mathbf{z}}^{t+1} = f_{\mathbf{z}}^{t} - \gamma (S_{\mathbf{x}}^* S_{\mathbf{x}} f_{\mathbf{z}}^{t} - S_{\mathbf{x}}^* \mathbf{y}), \qquad f_{\mathbf{z}}^{0} = 0,$$

with

$$\gamma = \frac{1}{\kappa^2}$$

the algorithms

both algorithms boil down to find

$$f_{\mathbf{z}}(x) = \sum_{i=1}^{n} \alpha K(x_i, x)$$

where for tikhonov

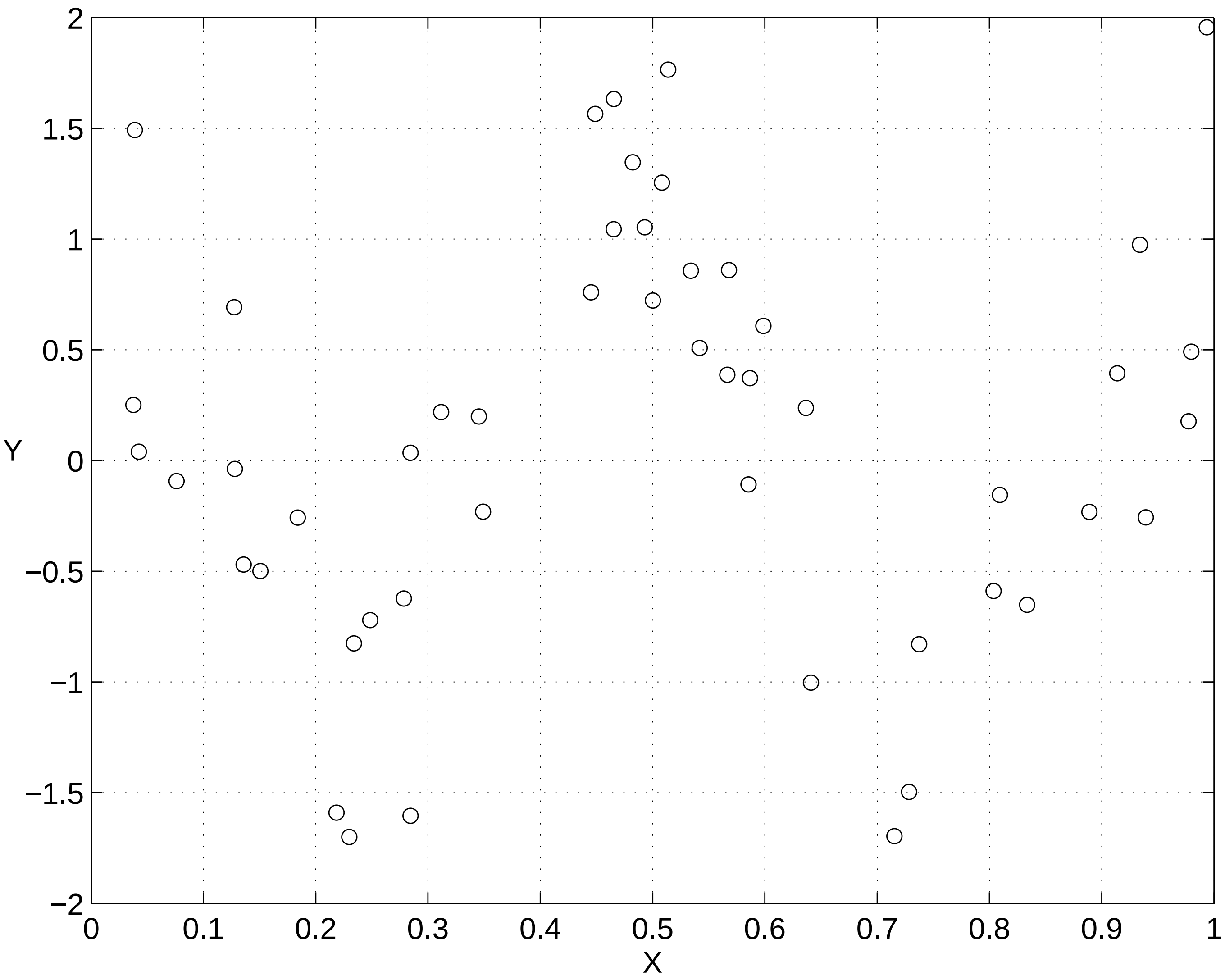
$$\alpha^{\lambda} = (\mathbf{K} + \lambda n I)^{-1}\mathbf{y} \qquad \mathbf{K}_{ij} = K(x_i, x_j)$$

while for landweber

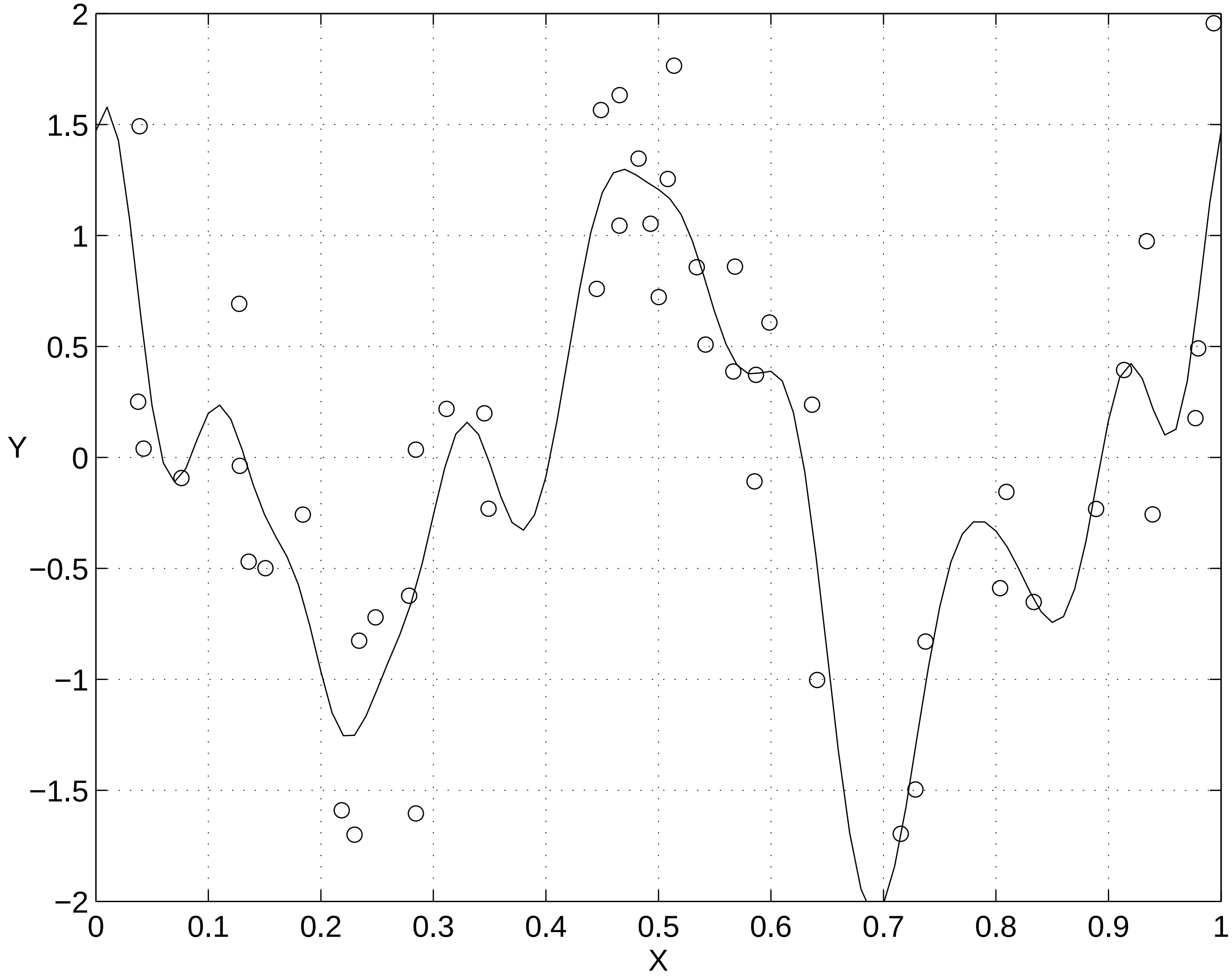
$$\alpha^{t+1} = \alpha^{t} - \frac{\gamma}{n}(\mathbf{K}\alpha^{t} - \mathbf{y}) \qquad \mathbf{K}_{ij} = K(x_i, x_j)$$

we want to know how well each solution approximates f_ρ

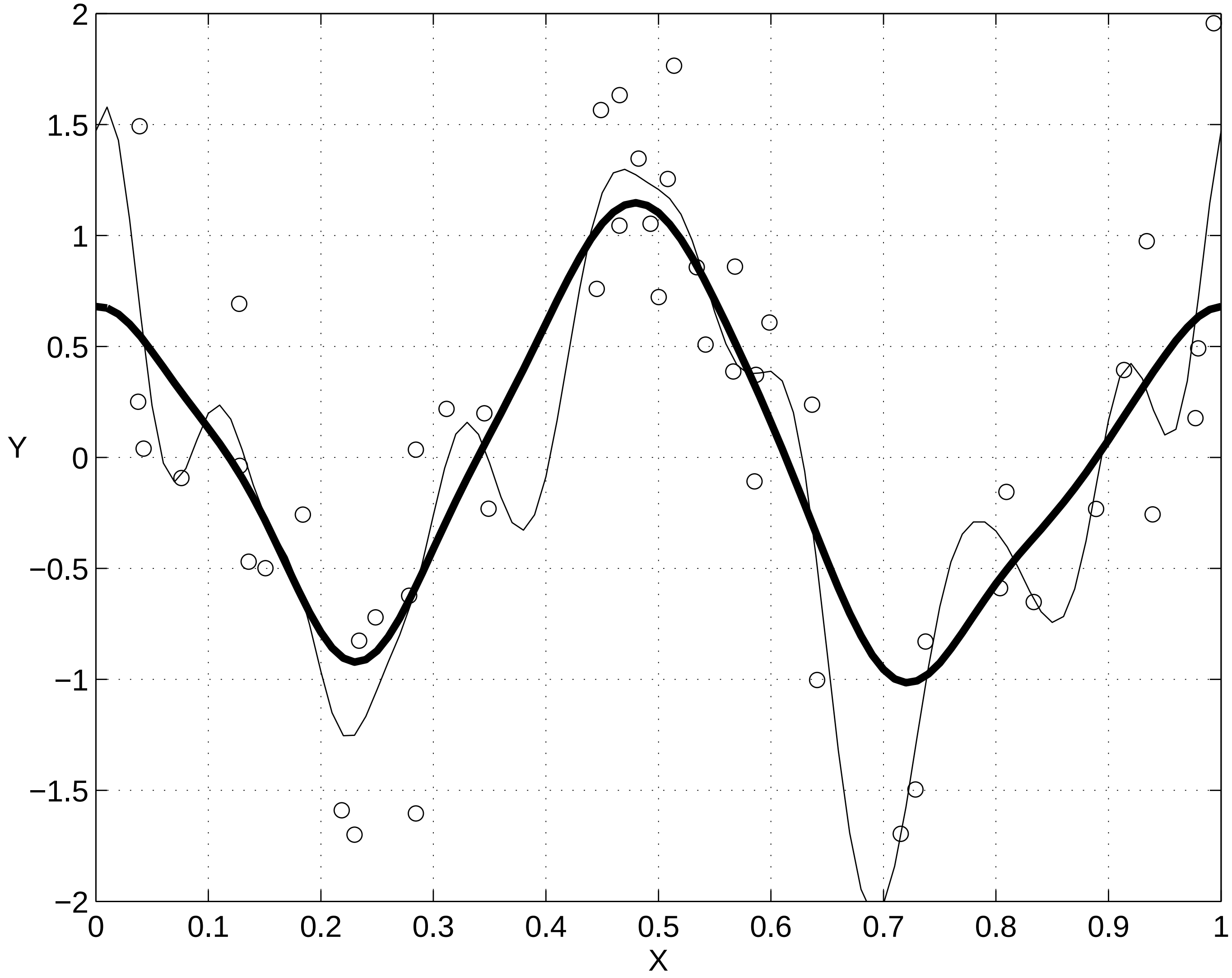
one dimensional regression



one dimensional regression



one dimensional regression



tikhonov case: analytic result

$$\left| \sqrt{\mathcal{E}(f_{\mathbf{z}}^{\lambda}) - \mathcal{E}(f_{\rho})} - \sqrt{\mathcal{E}(f^{\lambda}) - \mathcal{E}(f_{\rho})} \right| \leq$$

$$\frac{1}{\sqrt{\lambda}} \left(\frac{\|S_{\mathbf{x}}^{*} S_{\mathbf{x}} - I_K^{*} I_K\|_{\mathcal{L}(\mathcal{H})}}{\sqrt{\lambda}} + \|S_{\mathbf{x}}^{*}\mathbf{y} - I_K^{*} f_{\rho}\|_{\mathcal{H}} \right)$$

the two terms in the rhs do not depend on λ and are of probabilistic nature: the effect of the regularization procedure is factorized by analytic methods

generalized bennett inequality

- since $\mathcal{H}$ is an rkhs, that is, $f(x) = \langle f, K_x \rangle_{\mathcal{H}}$

$$\begin{array}{rclcrcl} S_{\mathbf{x}}^* \mathbf{y} & = & \frac{1}{n}\sum_{i=1}^n y_i K_{x_i} & \qquad & I_K^* f_\rho & = & \mathbb{E}_{x,y}[y K_x] \\ S_{\mathbf{x}}^* S_{\mathbf{x}} & = & \frac{1}{n}\sum_{i=1}^n \langle \cdot, K_{x_i} \rangle_{\mathcal{H}} K_{x_i} & & I_K^* I_K & = & \mathbb{E}_x[\langle \cdot, K_x \rangle K_x] \end{array}$$

- **theorem** [Smale-Yao ('04)]
let $\xi : X \times Y \to \mathcal{H}$ be a random variable, $\|\xi(x,y)\|_{\mathcal{H}} \le 1$

$$\mathrm{P}\left[\left\| \frac{1}{n} \sum_{i=1}^n \xi(x_i, y_i) - \mathbb{E}_{x,y}(\xi) \right\|_{\mathcal{H}} \ge \varepsilon \right] \le 2 \exp\left[-\frac{n}{2} \varepsilon \log(1+\varepsilon) \right] = \eta$$

probabilistic bound

with probability greater than $1-\eta$

$$\left|\sqrt{\mathcal{E}(f_{\mathbf{z}}^{\lambda})-\mathcal{E}(f_\rho)}-\sqrt{\mathcal{E}(f^{\lambda})-\mathcal{E}(f_\rho)}\right| \leq \left(\frac{C_1}{\sqrt{\lambda^2 n}}+\frac{C_2}{\sqrt{\lambda n}}\right)\sqrt{\log\frac{4}{\eta}}+o\left(\frac{1}{\sqrt{\lambda^2 n}}\right)$$

- the subset of $\mathbf{z}\in(X\times Y)^n$ for which the bound holds depends on n and η, but not on λ
- C_1 and C_2 are numerical (simple) constants
- $o(\frac{1}{\sqrt{\lambda^2 n}})$ depends also on η

consistency and rates

if $f_\rho \in \mathcal{H}$ then

- for tikhonov: $\lambda_n = n^{1/2}$ whp

$$\mathcal{E}(f_{\mathbf{z}}^{\lambda_n}) - \mathcal{E}(f_\rho) \leq C_\eta n^{-\frac{1}{2}}$$

(caponnetto and de vito, 2005; smale and zhou, 2005)

- for landweber: $t_n = n^{1/3}$ whp

$$\mathcal{E}(f_{\mathbf{z}}^{t_n}) - \mathcal{E}(f_\rho) \leq C_\eta n^{-\frac{1}{3}}$$

(yao, rosasco, and caponnetto, 2005)

future work

- semiterative regularization
- a posteriori regularization parameter choices: discrepancy principle
- connection between sparsity, regularization and feature selection