Spectral Filtering for MultiOutput Learning

Lorenzo Rosasco
Center for Biological and Computational Learning, MIT
Universita' di Genova, Italy

Plan

- Learning with kernels
- Multioutput kernel and regularization
- Spectral filtering
- Perspectives

Scalar Case

- function estimation from samples

$f : R^d \to R$ $\qquad (x_i, y_i)_{i=1}^n$

- kernel models

$$f = \sum_j K(x_j, \cdot) c_j$$

Kernels and Regularization

RKHS: Definitions

Hilbert space of functions $\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}}$ such that $\exists\ k : R^d \times R^d \to R$ and

$$k(x, \cdot) \in \mathcal{H}$$

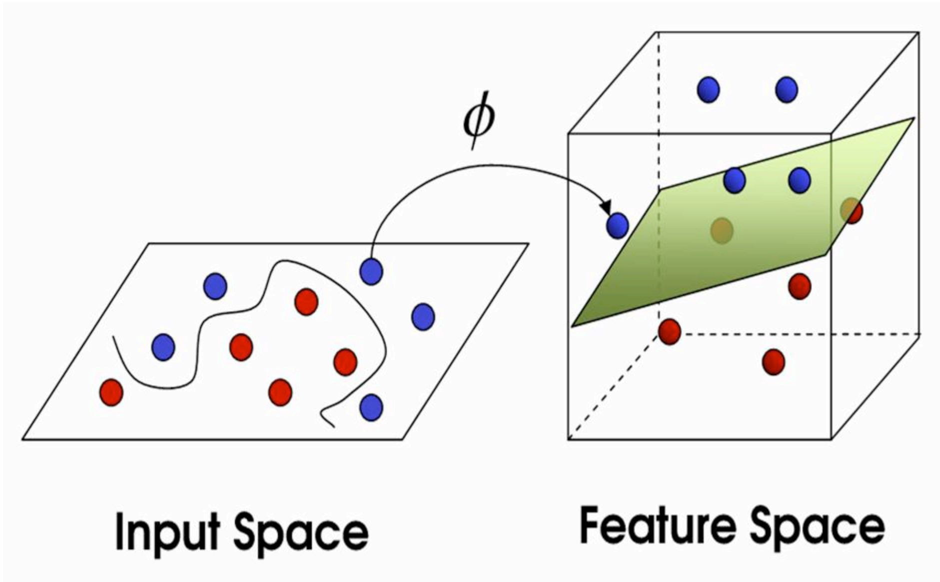
and

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

Tikhonov Regularization

$$\min_{f \in \mathcal{H}} \{ \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \|f\|_{\mathcal{H}}^2 \}.$$

Kernel Design



- feature map

$$K(x, s) = \langle \Phi(x), \Phi(s) \rangle$$

- regularizers

$$J(f) = \|f\|_{\mathcal{H}}^2$$

Multiple Outputs

- vector functions

$$f : R^d \to R^T$$

- samples

$$(x_i^T, y_i^T)_{i=1}^{n_T}$$

- kernel models

$$f = \sum_j K(x_j, \cdot) c_j \qquad c_j \in R^T$$

RKHS

RKHS: Definitions

Hilbert space of functions $\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}}$ such that $\exists\ K : R^d \times R^d \longrightarrow R^{T \times T}$ and for $c \in R^T$

$$K(x, \cdot)c \in \mathcal{H}$$

and

$$f(x) = \langle f, K(\cdot, x)c \rangle_{\mathcal{H}}$$

Tikhonov Regularization

$$\min_{f=(f^1,\ldots,f^T)\in\mathcal{H}} \{\sum_{j=1}^{T} \frac{1}{n_T} \sum_{i=1}^{n} (y_i^j - f^j(x_i^j))^2 + \lambda \|f\|_{\mathcal{H}}^2\}.$$

Which Kernels?

Component wise definition

$$K : (R^d, T) \times (R^d, T) \to R \qquad K((x,t),(x',t'))$$

A general class of kernels

$$K(x,x') = \sum_r k_r(x,x') A_r$$

Kernels and Regularizers

Consider

$$K(x, x') = k(x, x')A$$

Then

$$\|f\|_{\mathcal{H}}^2 = \sum_{j,i} A^{\dagger}_{j,i} \langle f^j, f^i \rangle_k$$

with

$$f = (f^1, \cdots, f^T)$$

Example: Mixed Effect

$$\Gamma_\omega(x, x') = K(x, x')(\omega \mathbf{1} + (1 - \omega)\mathbf{I})$$

$$J(f) = A_\omega \left(B_\omega \sum_{\ell=1}^{T} ||f^\ell||_K^2 + \omega T \sum_{\ell=1}^{T} ||f^\ell - \frac{1}{T} \sum_{q=1}^{T} f^q||_K^2 \right)$$

Example: Clustering Outputs

M specifies the clusters

$$G_{lq} = \epsilon_1 \delta_{lq} + (\epsilon_2 - \epsilon_1) M_{lq}$$

$$K(x, x') = k(x, x') G^{\dagger}$$

$$J(f) = \epsilon_1 \sum_{c=1}^{r} \sum_{l \in I(c)} ||f^l - \overline{f}_c||_K^2 + \epsilon_2 \sum_{c=1}^{r} m_c ||\overline{f}_c||_K^2,$$

Example: Graph

M is an adjacency matrix among the tasks

$$L = D - M,$$

$$K(x, x') = k(x, x') L^{\dagger}$$

$$J(f) = \frac{1}{2} \sum_{\ell, q=1}^{T} ||f^{\ell} - f^{q}||_K^2 M_{\ell q} + \sum_{\ell=1}^{T} ||f^{\ell}||_K^2 M_{\ell \ell}.$$

Inference and Computations

Least Squares and Tikhonov

$$c = (K + \lambda n I)^{-1} Y$$

Kernel Matrix is $(Tn_T) \times (Tn_T)$

c, Y are Tn_T

Computing the solution for N different regularization parameter is expensive

$$O(N(Tn_T)^3)$$

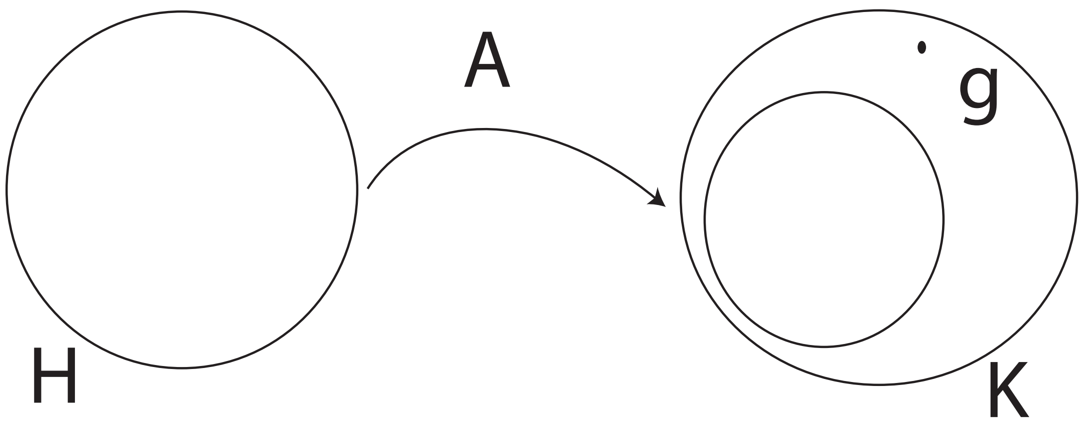
Ill-posed Problems

Well-posedness in the sense of Hadamard

- A solution exists
- The solution is unique
- The solution depends continuously on the data

Problems that are not well-posed are termed *ill-posed*.

$$Af = g$$



Regularization and Filtering

$$c = \sum_i \frac{1}{\sigma_i + \lambda n} \langle u_i, Y \rangle u_i$$

Spectral Filtering

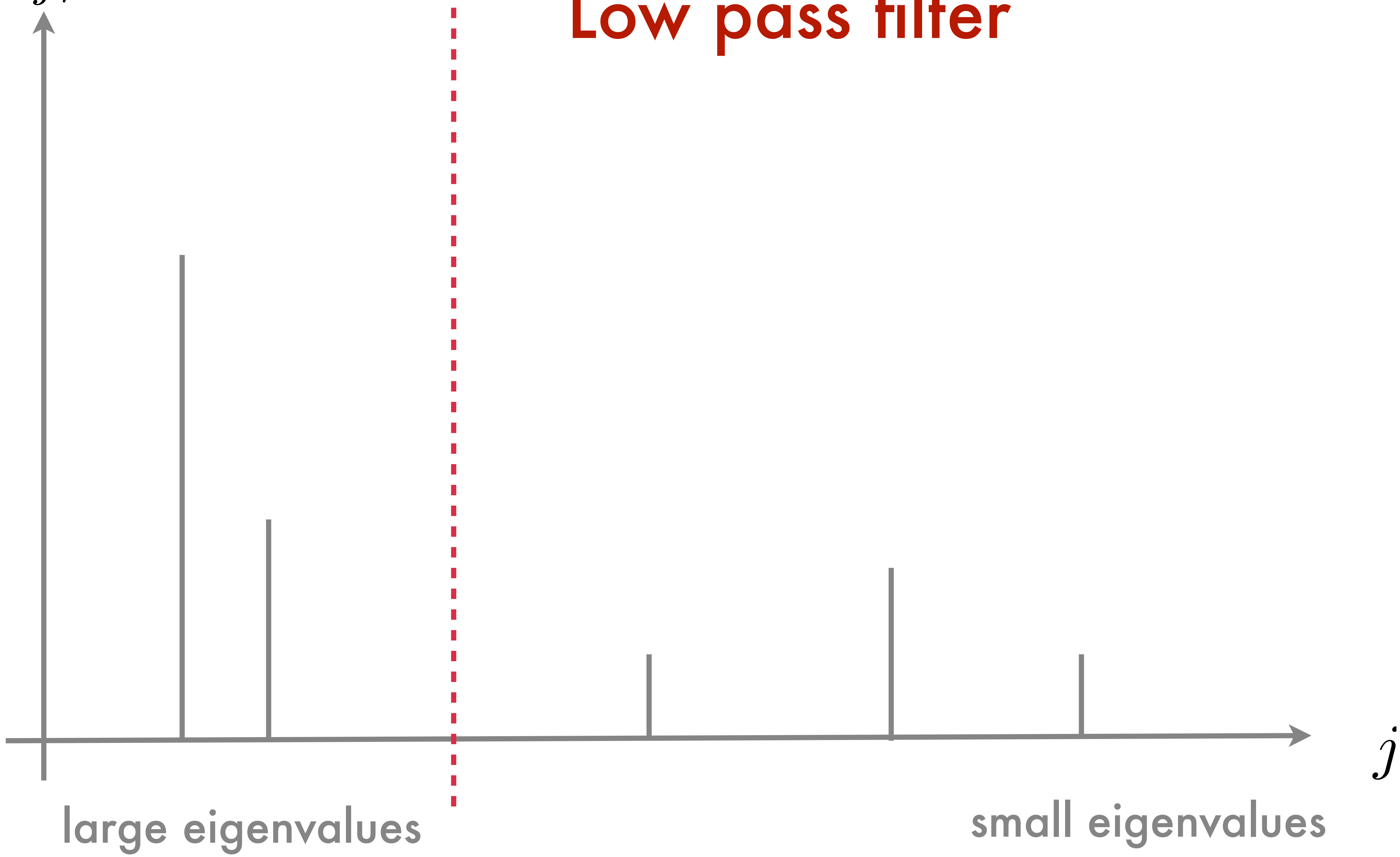
$$c = \sum_i G_\lambda(\sigma_i) \langle u_i, Y \rangle u_i$$

Regularization and Filtering

$$\frac{1}{\sigma_j}\langle Y, u_j \rangle$$

$$G_\lambda(\sigma_j)\langle Y, u_j \rangle$$

Low pass filter



Classical Examples

- Tikhonov Regularization

$$G_\lambda(\sigma) = \frac{1}{\sigma + \lambda}$$

Other Examples

Many other Examples of Filters (only some known in machine learning)

- TSVD (principal component regression)
- Landweber iteration (L_2 boosting)
- ν- method
- iterated Tikhonov

(Engl et al., Rosasco et l. ’05, Lo Gerfo et al. ’08, Bauer et. al. ’05)

Early Stopping

The filter correspond to a truncated expansion of the inverse.

$$G_\lambda(\sigma) = \eta \sum_{j=1}^{t} (1 - \eta\sigma)^j \sim \frac{1}{\sigma}$$

$$A^{-1} \sim \eta \sum_{j=1}^{t} (I - \eta A)^j$$

Implementation

set $\alpha_0 = 0$

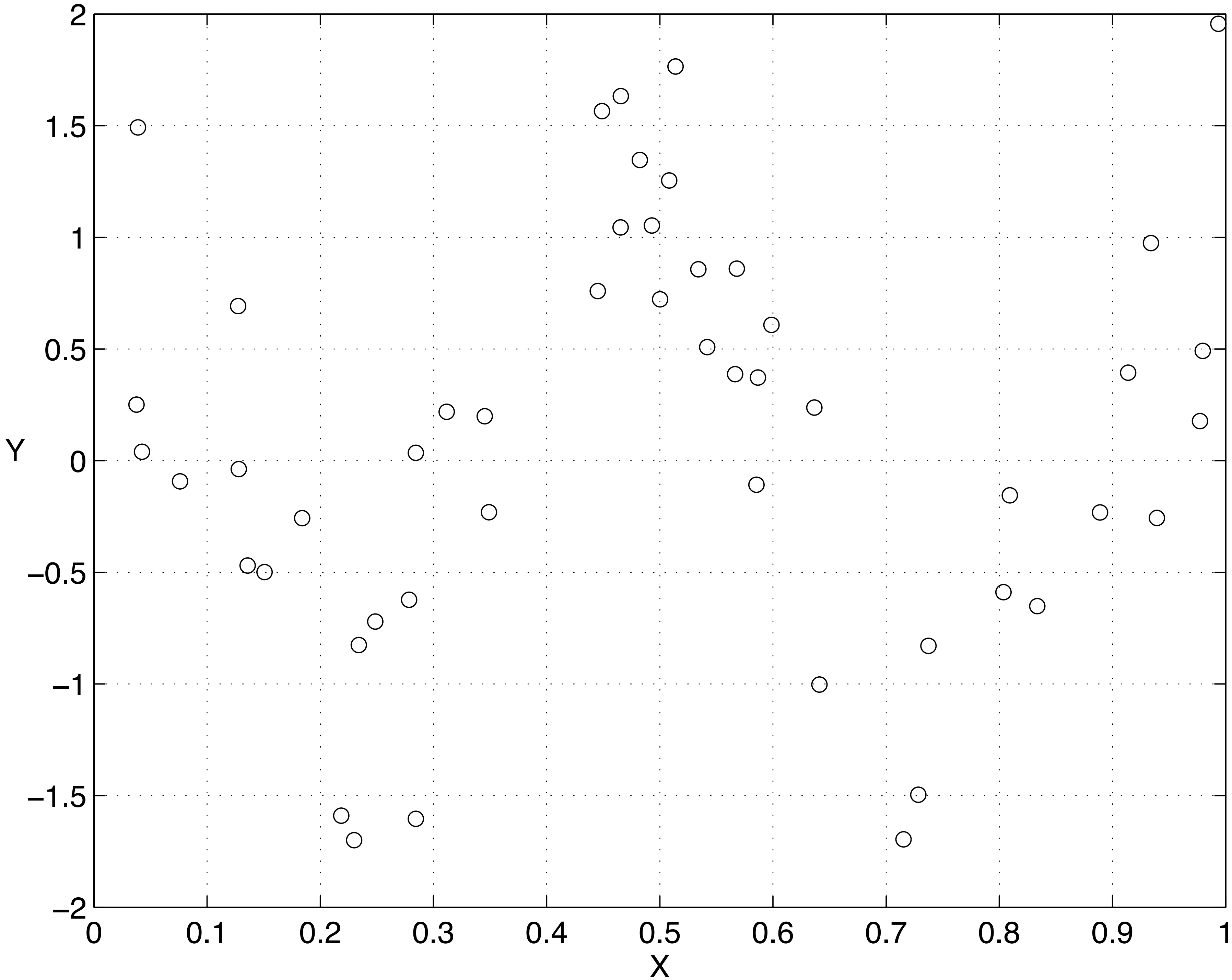
for i $= 1, \dots, t$

$$\alpha_i = \alpha_{i-1} + \eta(Y - \mathbf{K}\alpha_{i-1})$$

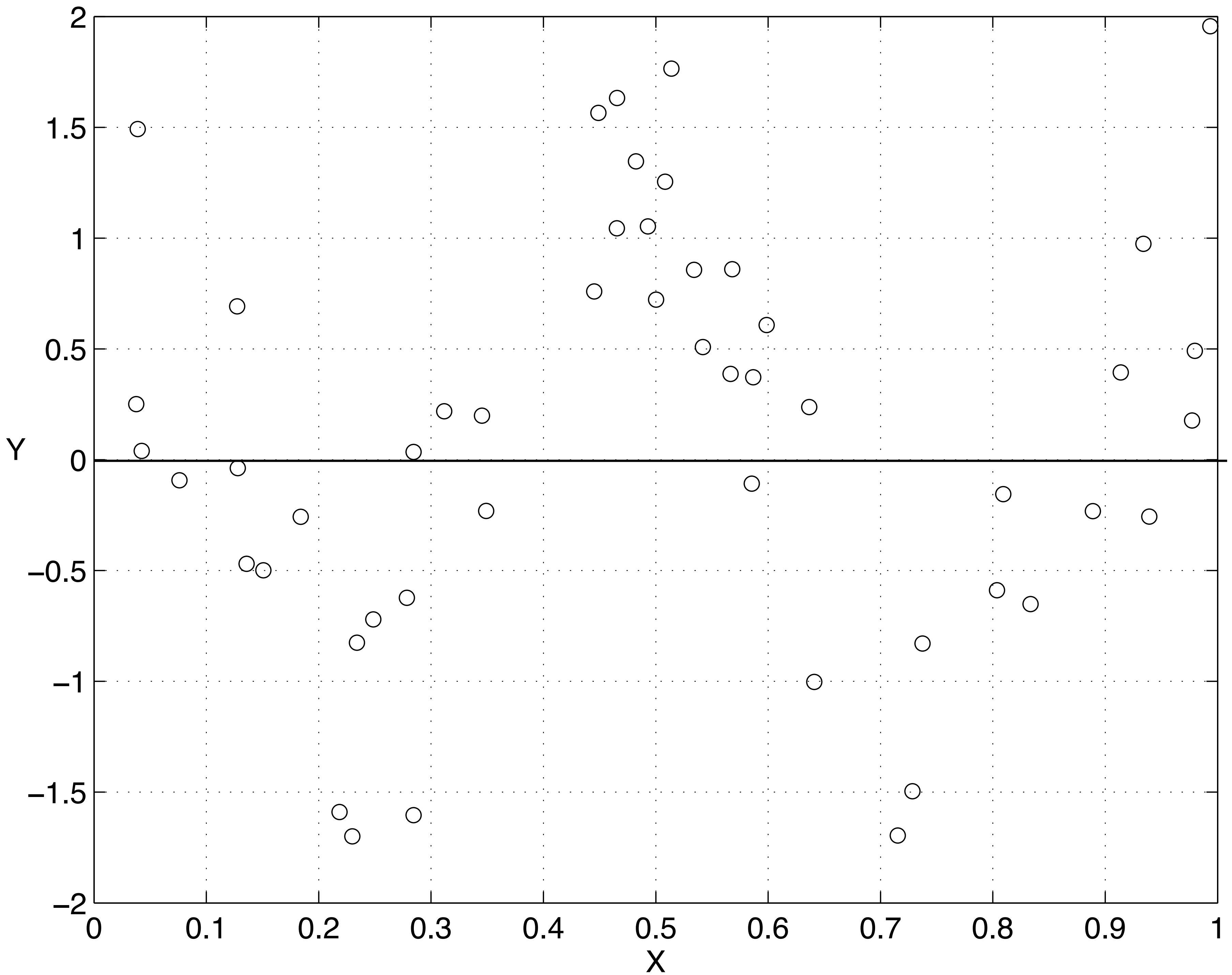
Estimator

$$f^t = \sum_{i=1}^{n} \alpha_i^t K(x_i, \cdot)$$

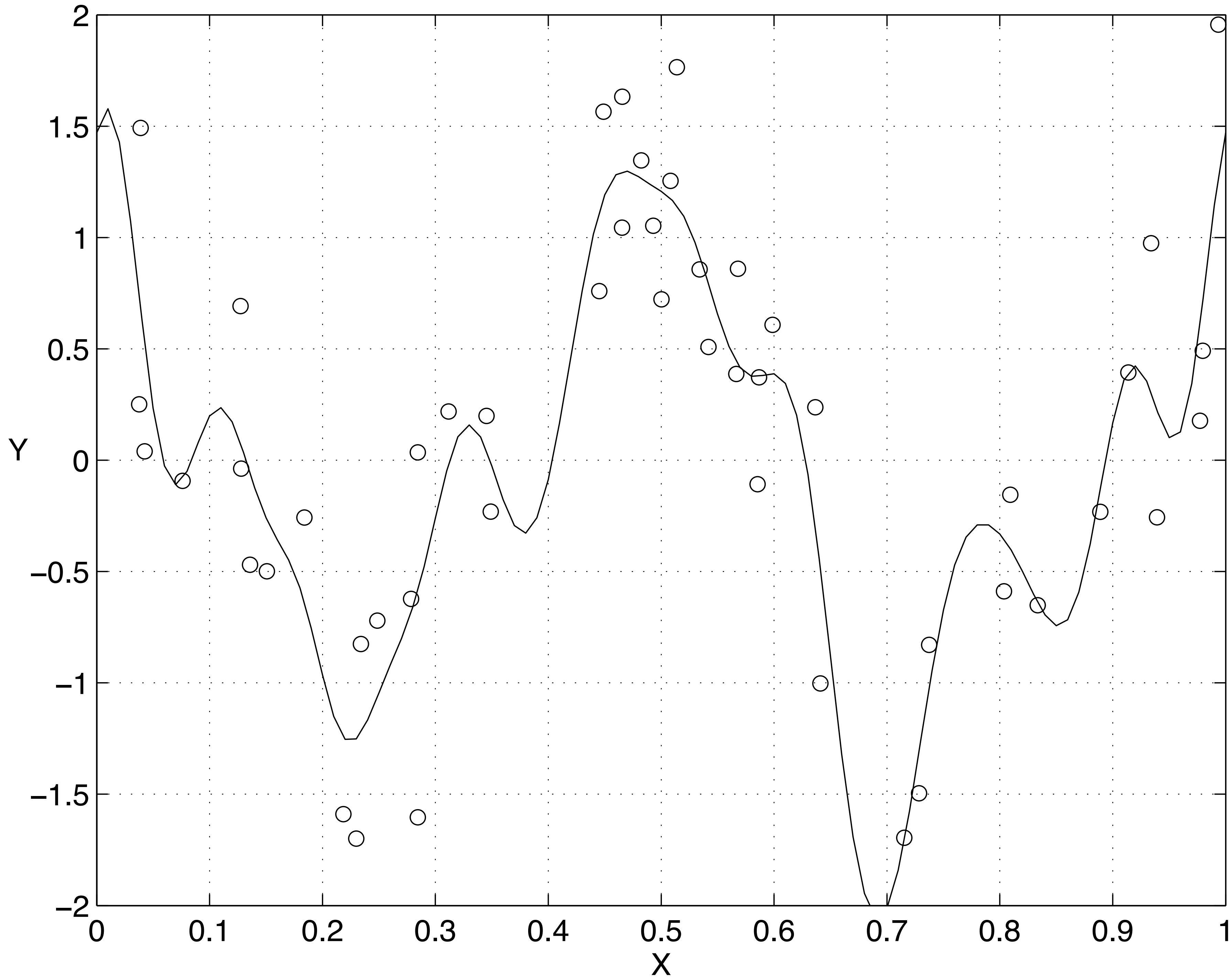
Early stopping at work



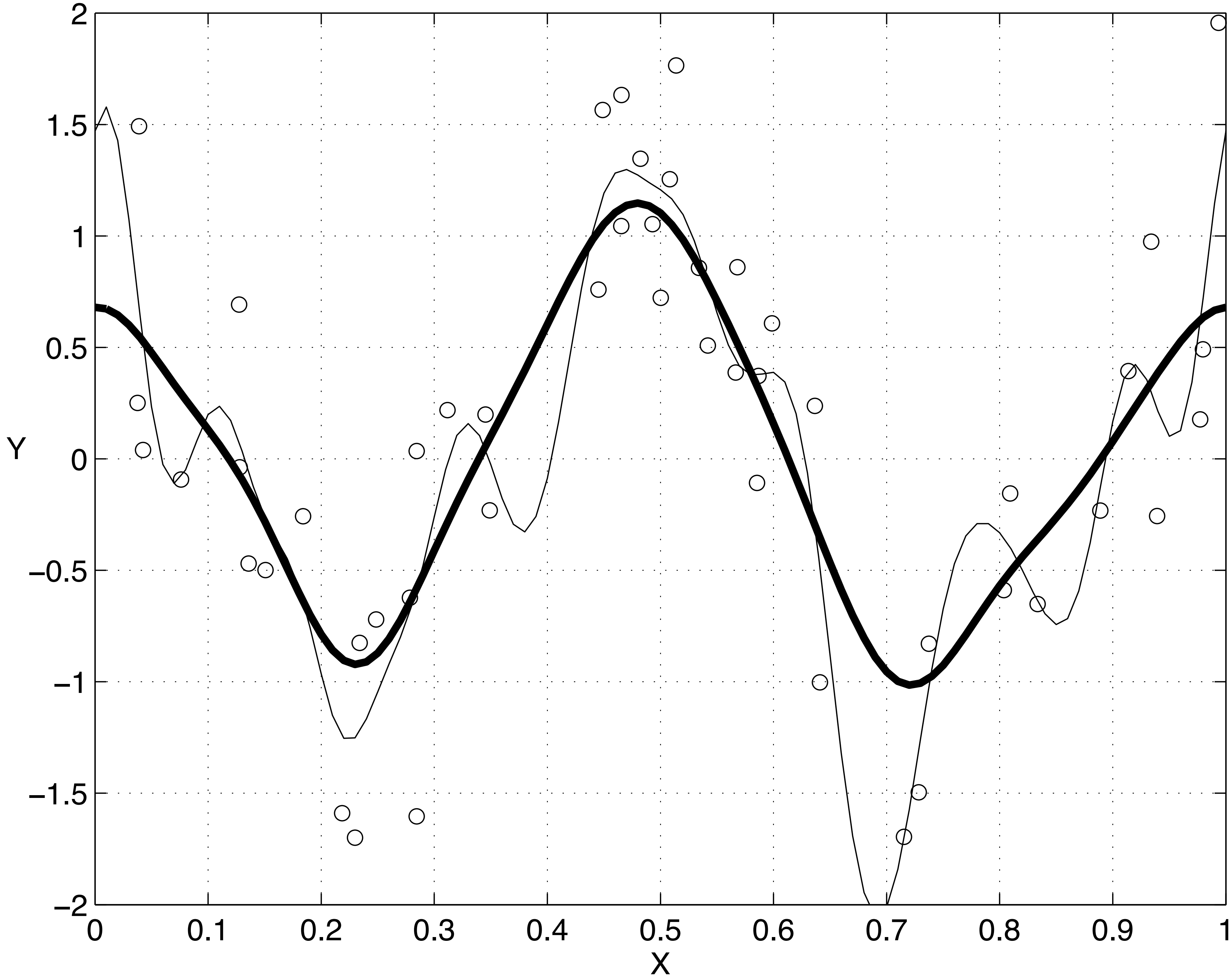
Early stopping at work



Early stopping at work



Early stopping at work



Remarks

- Empirical risk minimization with no constraints
- Regularization parameter is t: iteration regularizes
- No need of SVD
- Only matrix/vector multiplication

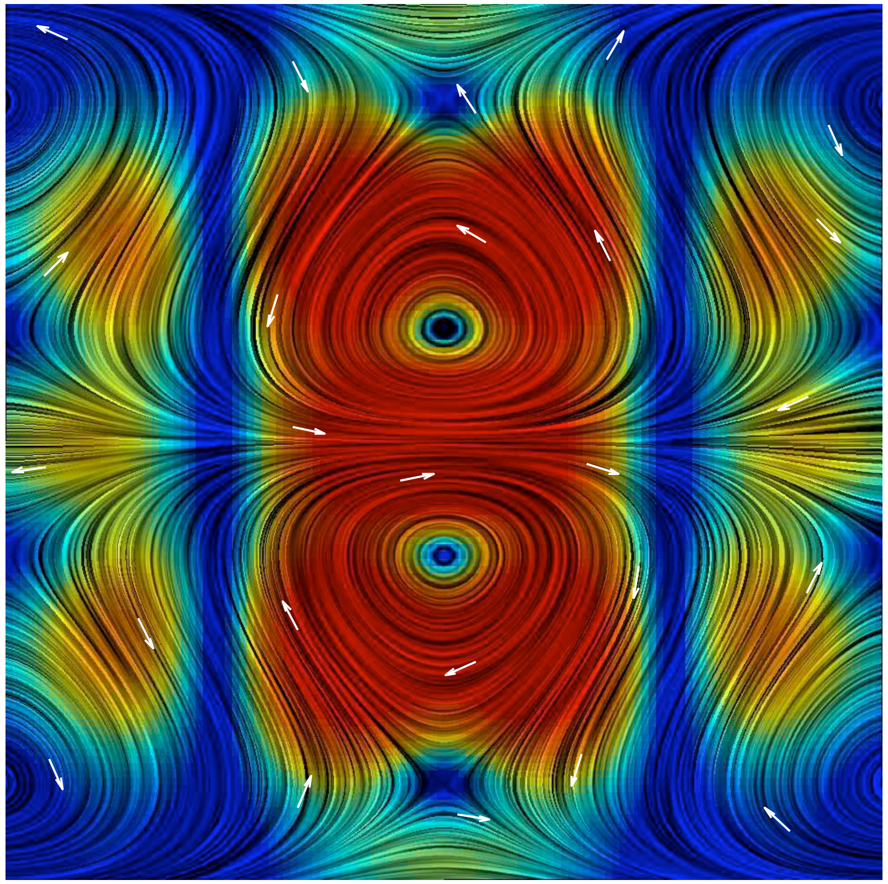
$$O(N(Tn_T)^2)$$

Fast Solution for Tikhonov Regularization

For Kernel of the form $K(x, x') = k(x, x')A$
we can diagonalize A and rotate data.

Tikhonov Regularization can be solved at the price of a single task!

Vector fields



$$v^1(x,y) = 2sin(3x)sin(1.5y)$$
$$v^2(x,y) = 2cos(3y)cos(1.5x)$$

+

Convolution with a Gaussian

Useful Kernels

Divergence Free

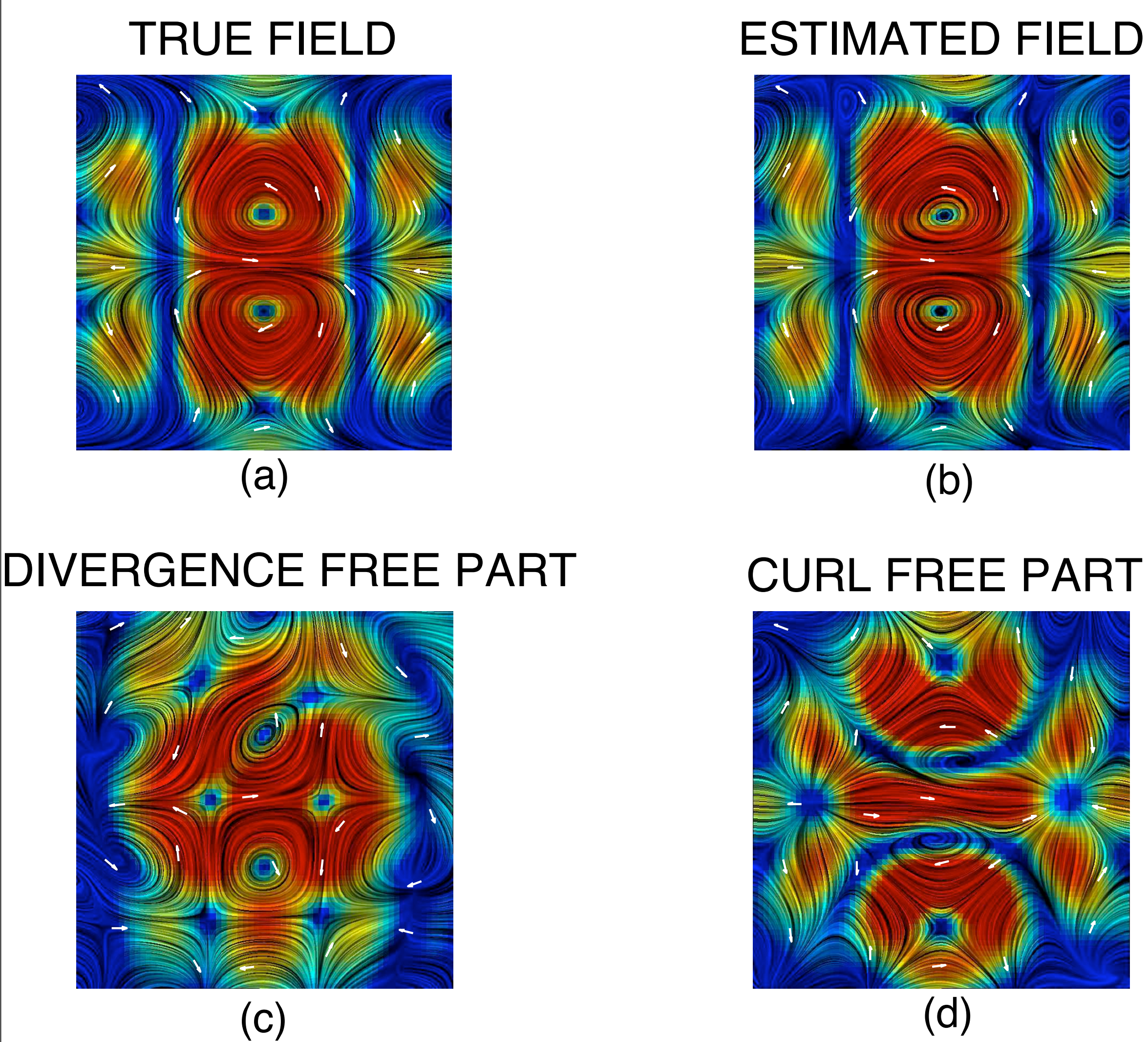
$$\Gamma_{df}(x,x') = \frac{1}{\sigma^2} e^{-\frac{||x-y||^2}{2\sigma^2}} A_{x,x'}$$

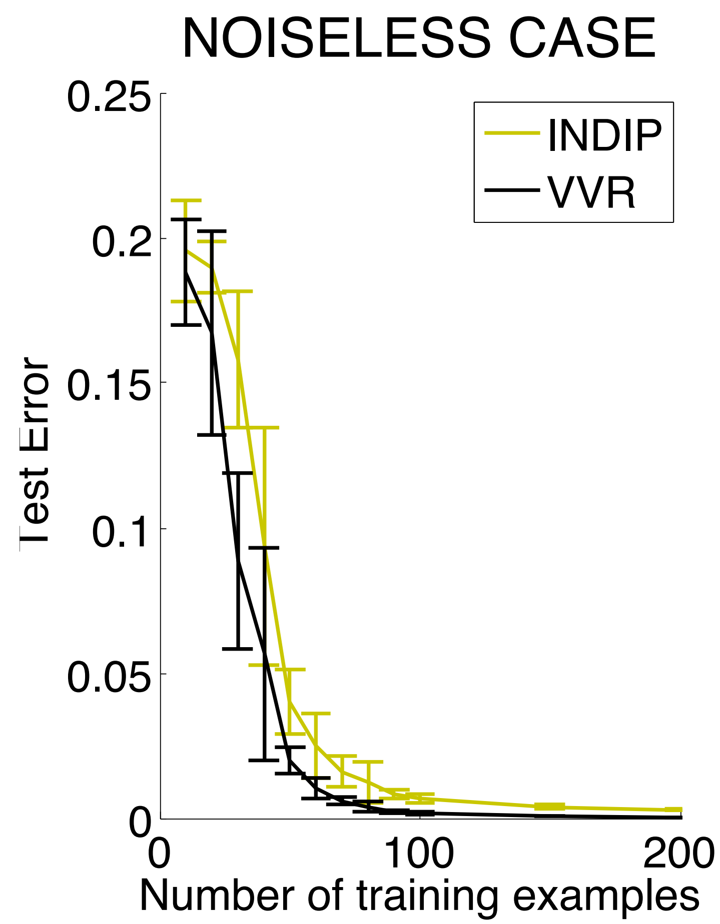
$$A_{x,x'} = \frac{(x-x')(x-x')^T}{\sigma^2} + \left((T-1) - \frac{||x-x'||^2}{\sigma^2}\right)\mathbf{I}$$

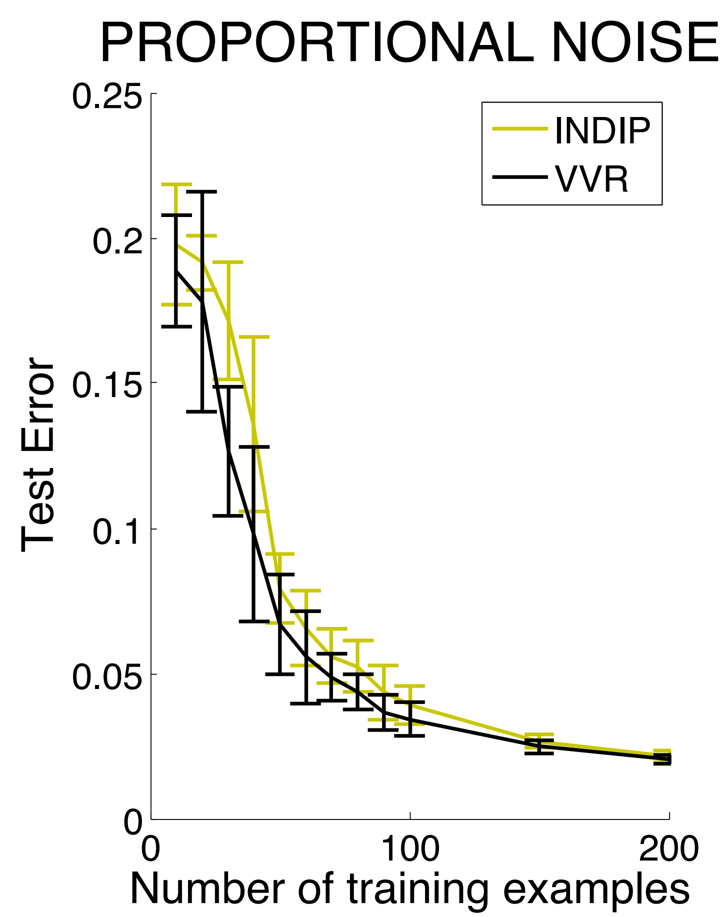
Curl Free

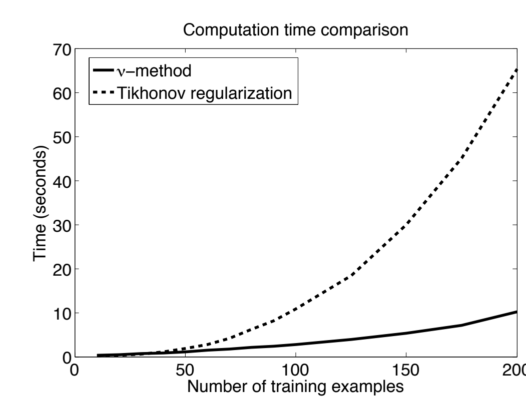
$$\mathit{\Gamma}_{cf}(x,x') = \frac{1}{\sigma^2} e^{-\frac{||x-x'||^2}{2\sigma^2}} \left(\mathbf{I} - \left(\frac{x-x'}{\sigma}\right)\left(\frac{x-x'}{\sigma}\right)^T\right)$$

Numerical Results

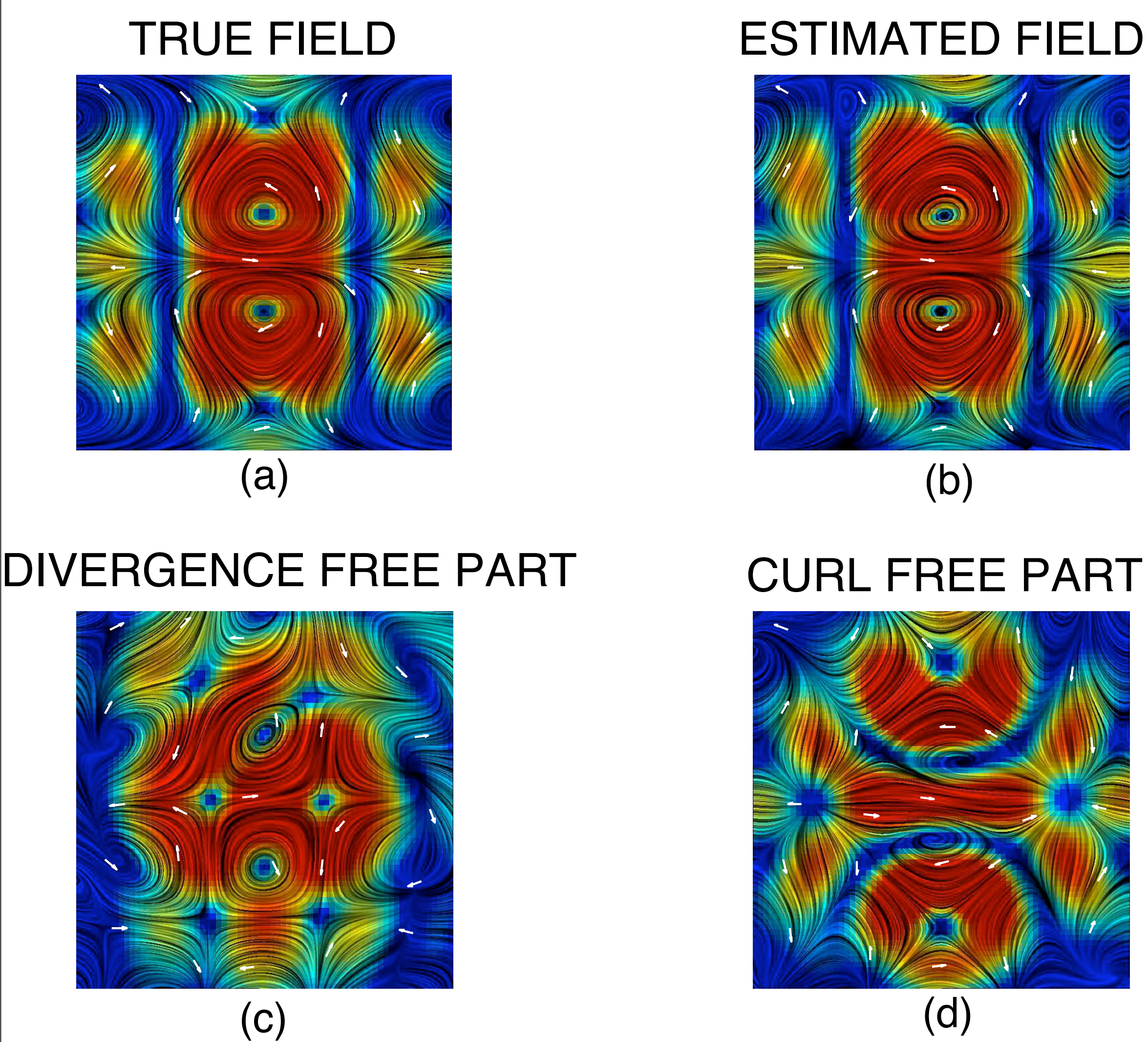


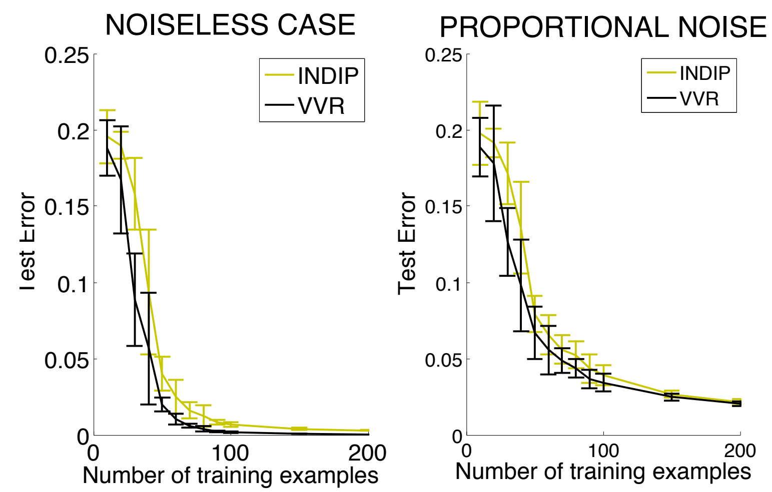


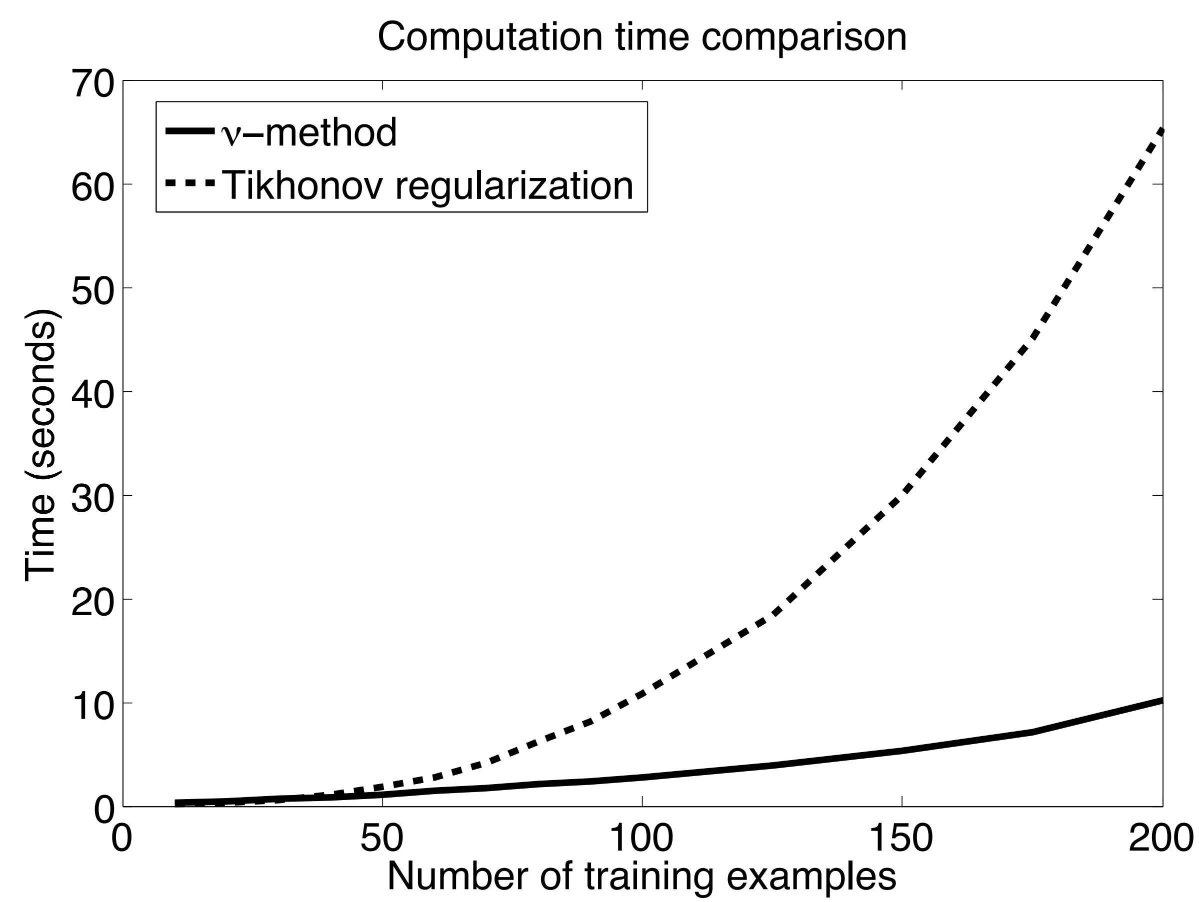




Numerical Results







Some Theory: Random Operators

$$T_{\mathbf{x}} f(x) = \frac{1}{n} \sum_{i=1}^{n} K(x, x_i) f(x_i)$$

$$T f(x) = \int K(x, x) f(x) d\rho(x)$$

$$P\left(\|T - T_{\mathbf{x}}\| \leq \frac{Ct}{\sqrt{n}} \right) \geq 1 - e^{-t^2}$$

The above result implies convergence of eigenfunctions and eigenvalues

Learning Rates

Theorem

If

$$\|T^{-r} f_\rho\| \leq R$$

with $r > 1/2$ and $\sigma_i \sim i^{-1/b}$, $b > 1$, then

$$\mathbb{P}\left(\|f_n - f_\rho\|_\rho^2 \leq C\sqrt{\tau} n^{-\frac{2rb}{2rb+1}} \right) \geq 1 - e^{-\tau^2}$$

for $\lambda = n^{-\frac{1}{2rb+1}}$.

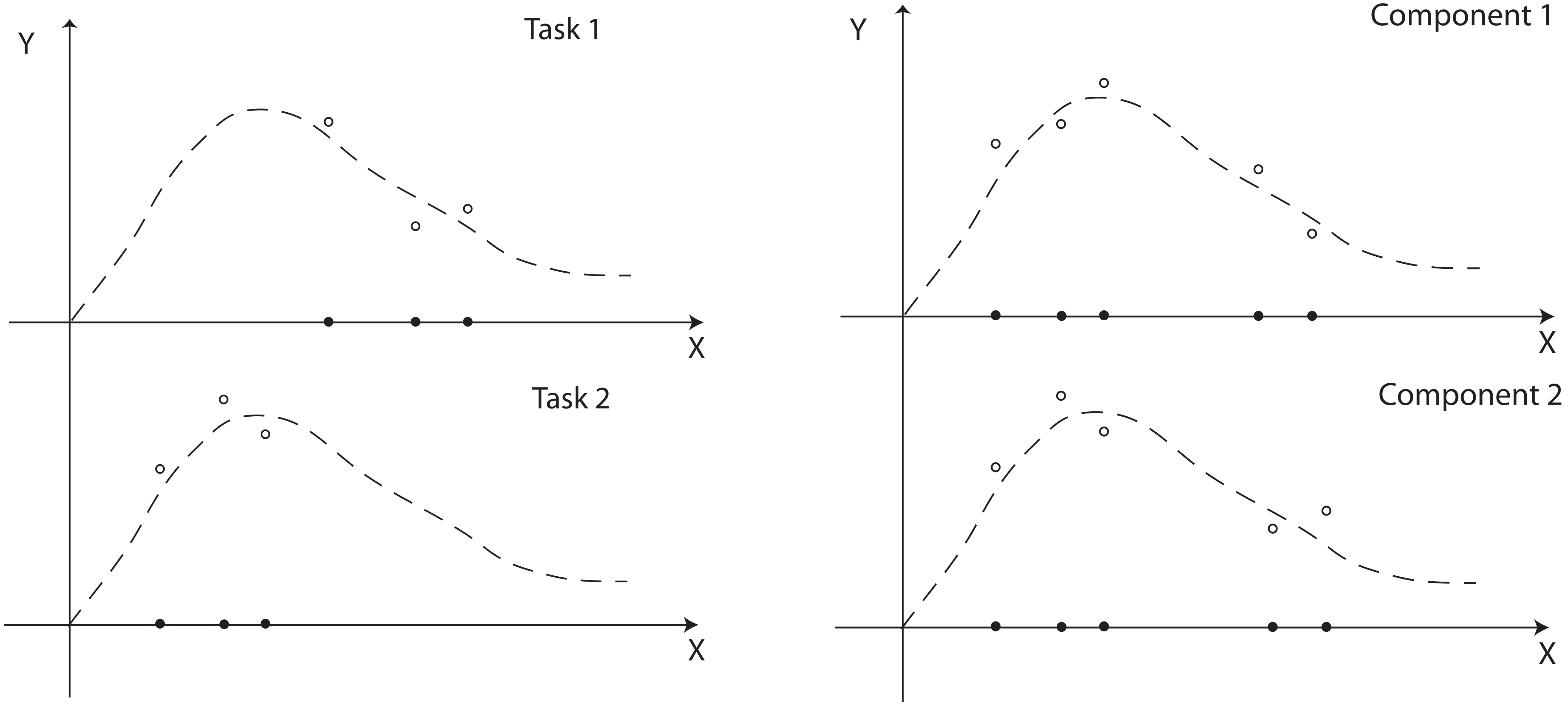
- The above estimate is optimal in a minimax sense.
- Parameter choice can be done adaptively

(Caponnetto et al., b=1 Bauer at al.)

Comments

- One name, 3 problems?
- Learning the kernels?

Vector Fields and Multi-tasks



Different Regimes?

- n>d>T (classical)
- d>n (high dimensional inference)
- T>n, n>T (??)

- curse of dimensionality vs blessing of smoothness
    - smoothness/d should be big

Multiple Classes

Inputs belong to one of T classes

Journal of Machine Learning Research 5 (2004) 101-141 Submitted 4/03; Revised 8/03; Published 1/04

In Defense of One-Vs-All Classification

Ryan Rifkin RIF@ALUM.MIT.EDU
Honda Research Institute USA
145 Tremont Street
Boston, MA 02111-1208, USA

Aldebaro Klautau A.KLAUTAU@IEEE.ORG
UC San Diego
La Jolla, CA 92093-0407, USA

One Versus All

Coding

$$1 = (1, -1, -1, \cdots), 2 = (-1, 1, -1, \cdots)...$$

Regression of Coding Vectors

$$\min_{f=(f^1,\cdots,f^T)} \sum_{i=1}^{n} \|y_i - f(x_i)\|_T^2 + \lambda \sum_{j=1}^{T} \|f^j\|^2$$

Classification Rule

$$c(x) = \max_{j=1,\cdots,T} f^j(x)$$

Remarks

- No correlation among classes
- How can we estimate it?
- In simulation one observe improvement in probability estimation but NOT in classification performances.

Regression vs Classification

- the components of the regression function are proportional of the conditional probabilities of each class
- the obtained estimator is Bayes Consistent

Learning the Kernel?

- Bayesian Approaches (consistency guarantees/stability/computability?)
- Regularization (what is the underlying Kernel? How are the outputs related?)