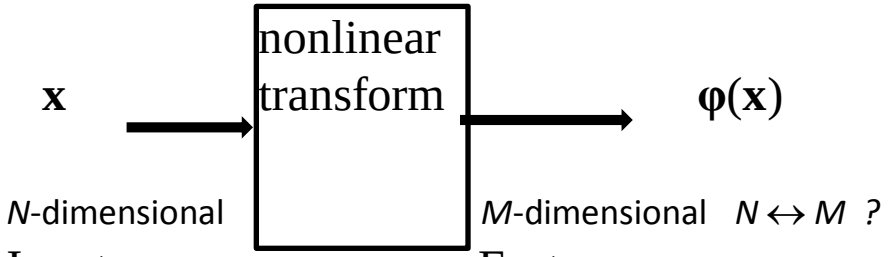
Nonlinear regression

Linear-in-the-parameters model:

Nonlinear transformation of the input data $\mathbf{x} \rightarrow \boldsymbol{\varphi}(\mathbf{x})$.



Nonlinear regression in the input space linear regression in the feature space.

Everything is valid but $\mathbf{X}$ should be replaced by $\boldsymbol{\Phi}$

$$\mathbf{w}_{\mathrm{LS}}^{*} = \left(\boldsymbol{\Phi}^{T}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^{T}\mathbf{d}$$

Regularized LS regression

$$\mathbf{w} = \left(\lambda\mathbf{I} + \boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^{\mathrm{T}} \quad \mathbf{d}$$

ML and Bayesian solutions too

Important questions:

- How to select the $\boldsymbol{\varphi}(\mathbf{x})$ basis functions?
- How many basis functions (*M*) should be used?
- How to select the basis function parameters (hyperparameters) ?

Nonlinear regression

Basis functions: Gaussian, spline, B-spline, wavelet, …
Parameters of basis functions:
e.g. for Gaussian centre point ($\mathbf{c}_i$) and width (σ)
the number of basis functions (M)
Possible solutions:

- all training points (input vectors)
- clustering, centre points of the clusters (k-means, Kohonen, centre of Voronoi regions)
- OLS

Nonlinear regression

Kernel methods (adifferent representation of the modeling problem)
A simple kernel representation (starting from the LS solution)

$$\hat{\mathbf{w}}^* = \hat{\mathbf{X}}^T (\hat{\mathbf{X}}\hat{\mathbf{X}}^T)^{-1}\mathbf{d}$$

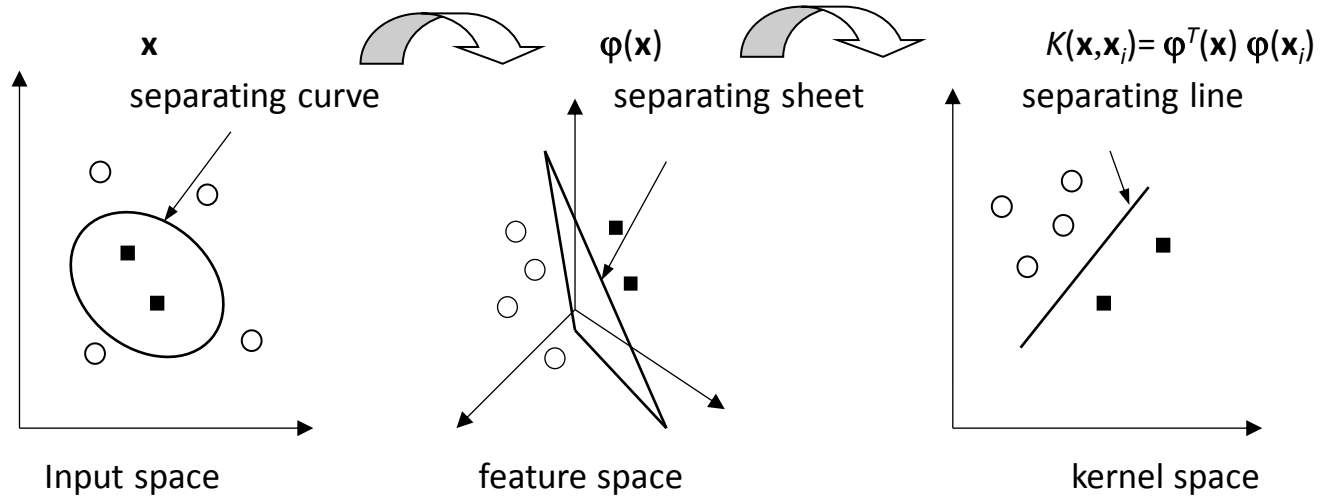
$$y(\hat{\mathbf{x}}) = \hat{\mathbf{x}}^T \hat{\mathbf{w}}^* = \hat{\mathbf{x}}^T \hat{\mathbf{X}}^T (\hat{\mathbf{X}}\hat{\mathbf{X}}^T)^{-1}\mathbf{d} \qquad y(\hat{\mathbf{x}}) = \hat{\mathbf{x}}^T \hat{\mathbf{X}}^T \boldsymbol{\alpha} = \sum_{i=1}^{P} \alpha_i (\hat{\mathbf{x}}^T \hat{\mathbf{x}}_i) = \sum_{i=1}^{P} \alpha_i K_i(\hat{\mathbf{x}}) \qquad \boldsymbol{\alpha} = (\hat{\mathbf{X}}\hat{\mathbf{X}}^T)^{-1}\mathbf{d}$$

$$\hat{\mathbf{x}}^T \hat{\mathbf{X}}^T = [\hat{\mathbf{x}}^T \hat{\mathbf{x}}_1, \hat{\mathbf{x}}^T \hat{\mathbf{x}}_2, \ldots, \hat{\mathbf{x}}^T \hat{\mathbf{x}}_i, \ldots, \hat{\mathbf{x}}^T \hat{\mathbf{x}}_P]$$

$$K_i(\hat{\mathbf{x}}) = \hat{\mathbf{x}}^T \hat{\mathbf{x}}_i$$

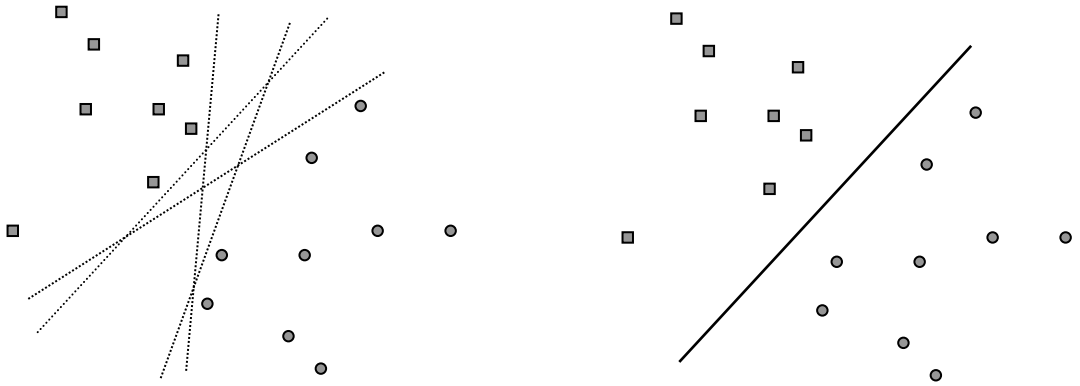
Nonlinear regression

Kernel trick



Nonlinear regression

SVM (support vector machines)
More natural to start with the linear classification version!
A simple linear classification task:



$$\mathbf{w}^T\mathbf{x}_i + b \geq a > 0 \quad \text{if} \quad d_i = +1$$
$$\mathbf{w}^T\mathbf{x}_i + b \leq -a < 0 \quad \text{if} \quad d_i = -1$$

i=1,,…, *P*

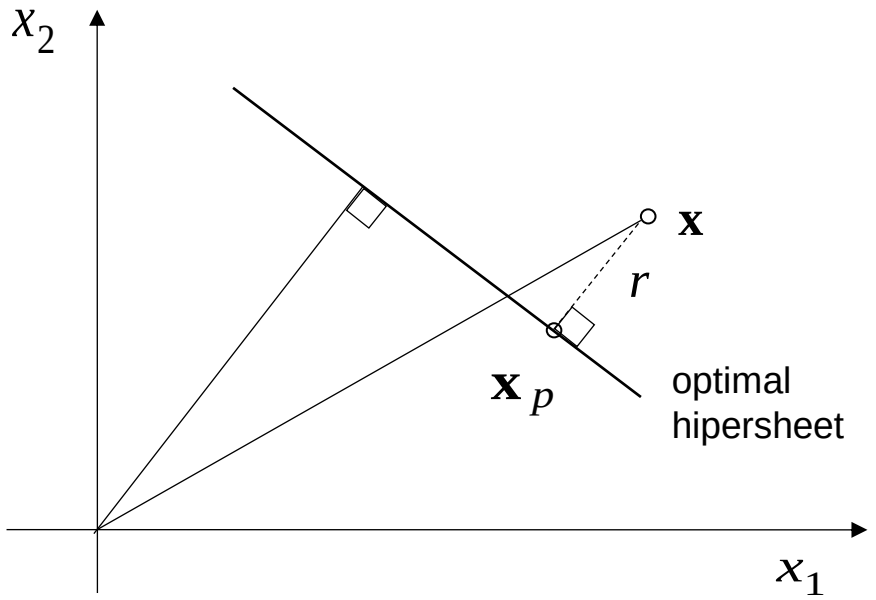
After a simple rescaling:

$$\mathbf{w}^T\mathbf{x}_i + b \geq +1 \quad \text{if} \quad d_i = +1$$
$$\mathbf{w}^T\mathbf{x}_i + b \leq -1 \quad \text{if} \quad d_i = -1$$

$$d_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 \qquad i = 1, 2, \ldots, P$$

,

SVM classification



$$g(\mathbf{x}_p) = 0 \qquad g(\mathbf{x}) = \mathbf{w}^{*T}\mathbf{x} + b^* = r\left\|\mathbf{w}^*\right\| \qquad \mathbf{x} = \mathbf{x}_p + r\frac{\mathbf{w}^*}{\left\|\mathbf{w}^*\right\|} \qquad r = \frac{g(\mathbf{x})}{\left\|\mathbf{w}^*\right\|} \qquad r = \frac{\rho}{2} = \frac{1}{\left\|\mathbf{w}^*\right\|}$$

$$d_i(\mathbf{w}^{*T}\mathbf{x}_i + b^*) \geq 1 \qquad i = 1, 2, \ldots, P \qquad \mathbf{w}^* = \arg\min_{\mathbf{w}}(\mathbf{w}^T\mathbf{w})$$

Lagrangian:

$$L(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\mathbf{w}^T\mathbf{w} - \sum_{i=1}^{P}\alpha_i\left[d_i(\mathbf{w}^T\mathbf{x}_i + b) - 1\right]$$

SVM classification

$$\frac{\partial L(\mathbf{w},b,\boldsymbol{\alpha})}{\partial \mathbf{w}} = 0 \rightarrow \mathbf{w} = \sum_{i=1}^{P} \alpha_i d_i \mathbf{x}_i \qquad \frac{\partial L(\mathbf{w},b,\boldsymbol{\alpha})}{\partial b} = 0 \rightarrow \sum_{i=1}^{P} \alpha_i d_i = 0 \qquad \alpha_i \geq 0$$

$$L(\mathbf{w},b,\boldsymbol{\alpha}) = \frac{1}{2}\mathbf{w}^T\mathbf{w} - \sum_{i=1}^{P} \alpha_i d_i \mathbf{w}^T \mathbf{x}_i - b\sum_{i=1}^{P} \alpha_i d_i + \sum_{i=1}^{P} \alpha_i$$

$$\mathbf{w}^T\mathbf{w} = \sum_{i=1}^{P} \alpha_i d_i \mathbf{w}^T \mathbf{x}_i = \sum_{i=1}^{P}\sum_{j=1}^{P} \alpha_i \alpha_j d_i d_j \mathbf{x}_i^T \mathbf{x}_j$$

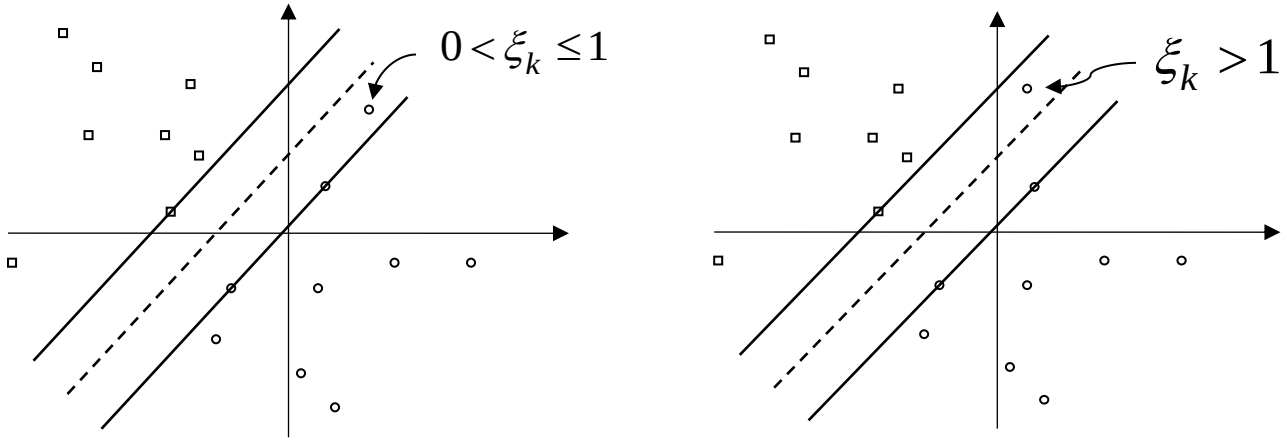
$$Q(\boldsymbol{\alpha}) = \sum_{i=1}^{P} \alpha_i - \frac{1}{2}\sum_{i=1}^{P}\sum_{j=1}^{P} \alpha_i \alpha_j d_i d_j \mathbf{x}_i^T \mathbf{x}_j \qquad \sum_{i=1}^{P} \alpha_i d_i = 0 \qquad \alpha_i \geq 0 \qquad y(\mathbf{x}) = \text{sign}\left[\sum_{i=1}^{P} \alpha_i^* d_i \mathbf{x}_i^T \mathbf{x} + b^*\right]$$

How to solve $Q(\boldsymbol{\alpha})$: Quadratic programming (Q.P.)

SVM classification

Using slacking variables

$$d_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 \quad i = 1, 2, \ldots, P \quad \rightarrow \quad d_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i,$$



$$J(\mathbf{w}) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{P}\xi_i$$

$$L(\mathbf{w}, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\gamma}, b) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{P}\xi_i - \sum_{i=1}^{P}\alpha_i\{d_i(\mathbf{w}^T\mathbf{x}_i + b) - 1 + \xi_i\} - \sum_{i=1}^{P}\gamma_i\xi_i$$

SVM classification

$$\frac{\partial L(\mathbf{w},\boldsymbol{\xi},\boldsymbol{\alpha},\boldsymbol{\gamma},b)}{\partial \mathbf{w}} = 0 \quad \rightarrow \quad \mathbf{w} = \sum_{i=1}^{P} \alpha_i d_i \mathbf{x}_i$$

$$\frac{\partial L(\mathbf{w},\boldsymbol{\xi},\boldsymbol{\alpha},\boldsymbol{\gamma},b)}{\partial b} = 0 \quad \rightarrow \quad \sum_{i=1}^{P} \alpha_i d_i = 0$$

$$\frac{\partial L(\mathbf{w},\boldsymbol{\xi},\boldsymbol{\alpha},\boldsymbol{\gamma},b)}{\partial \boldsymbol{\xi}} = 0 \quad \rightarrow \quad \gamma_i + \alpha_i = C$$

$$Q(\boldsymbol{\alpha}) = \sum_{i=1}^{P} \alpha_i - \frac{1}{2} \sum_{i=1}^{P} \sum_{j=1}^{P} \alpha_i \alpha_j d_i d_j \mathbf{x}_i^T \mathbf{x}_j$$

$$\sum_{i=1}^{P} d_i \alpha_i = 0, \quad 0 \le \alpha_i \le C, \quad i = 1, 2, \ldots, P$$

Nonlinear version

$$\mathbf{x} \rightarrow \boldsymbol{\varphi}(\mathbf{x}) \qquad K(\mathbf{x}_i, \mathbf{x}) = \boldsymbol{\varphi}^T(\mathbf{x}_i)\boldsymbol{\varphi}(\mathbf{x})$$

SVM (kernel methods)

Typical kernel functions

Linear	$K(\mathbf{x},\mathbf{x}_i)=\mathbf{x}_i^T\mathbf{x}$
Polynomial (with degree d	$K(\mathbf{x},\mathbf{x}_i)=(\mathbf{x}_i^T\mathbf{x}+1)^d$
Gauss (RBF)	$K(\mathbf{x},\mathbf{x}_i)=\exp\{-\|\mathbf{x}-\mathbf{x}_i\|^2/\sigma^2\}$
And many others	e.g spline, tanh,

General properties of the kernel functions

$K(\mathbf{x}_i,\mathbf{x}_j)\geq 0$

$K(\mathbf{x}_i,\mathbf{x}_j)=K(\|\mathbf{x}_i-\mathbf{x}_j\|)$

$K(\mathbf{x},\mathbf{x})=\max K(\mathbf{x},\mathbf{x}_j)$ $\quad$ i, j = 1,...,P.

$\lim_{t\to\infty} K(t)=0,\quad \text{ha}\quad t=\|\mathbf{x}_i-\mathbf{x}_j\|$

$K(\mathbf{x},\mathbf{z})=K_1(\mathbf{x},\mathbf{z})+K_2(\mathbf{x},\mathbf{z})$

$K(\mathbf{x},\mathbf{z})=aK_1(\mathbf{x},\mathbf{z})$

$K(\mathbf{x},\mathbf{z})=K_1(\mathbf{x},\mathbf{z})K_2(\mathbf{x},\mathbf{z})$

Nonlinear regression

SVM regression

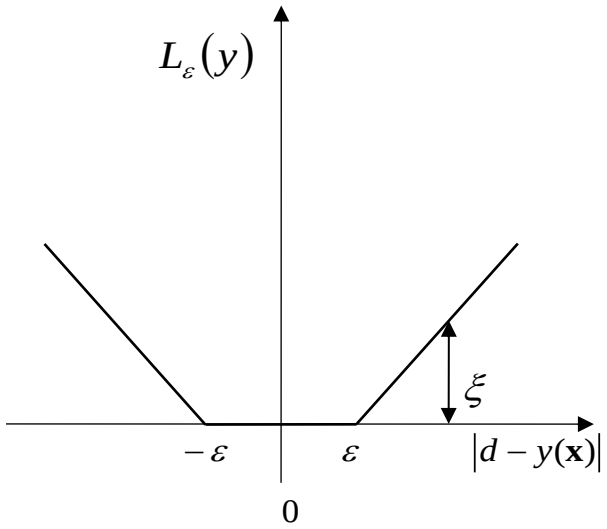
$$L_\varepsilon(y) = \begin{cases} 0 & \text{if} \quad |d - y(\mathbf{x})| < \varepsilon \\ |d - y(\mathbf{x})| - \varepsilon & \text{otherwise} \end{cases}$$

$$y(\mathbf{x}) = \sum_{j=1}^{M} w_j \varphi_j(\mathbf{x}) + b = \mathbf{w}^T \boldsymbol{\varphi}(\mathbf{x}) + b \qquad \mathbf{w} = [w_1, w_2, \ldots w_M]^T \qquad \boldsymbol{\varphi}(\mathbf{x}) = [\varphi_1(\mathbf{x}), \varphi_2(\mathbf{x}), \ldots \varphi_M(\mathbf{x})]^T$$

$$\begin{aligned} d_i - (\mathbf{w}^T \boldsymbol{\varphi}(\mathbf{x}_i) + b) &\le \varepsilon + \xi_i, \\ (\mathbf{w}^T \boldsymbol{\varphi}(\mathbf{x}_i) + b) - d_i &\le \varepsilon + \xi_i', \\ \xi_i &\ge 0, \\ \xi_i' &\ge 0, \end{aligned} \qquad i = 1, 2, \ldots, P$$

$$J(\mathbf{w}, \boldsymbol{\xi}, \boldsymbol{\xi}') = \frac{1}{2} \mathbf{w}^T \mathbf{w} + C \left(\sum_{i=1}^{P} (\xi_i + \xi_i') \right)$$

$$\begin{aligned} L(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\xi}', \boldsymbol{\alpha}, \boldsymbol{\alpha}', \boldsymbol{\gamma}, \boldsymbol{\gamma}') &= C \sum_{i=1}^{P} (\xi_i + \xi_i') + \frac{1}{2} \mathbf{w}^T \mathbf{w} \\ &- \sum_{i=1}^{P} \alpha_i [\mathbf{w}^T \boldsymbol{\varphi}(\mathbf{x}_i) + b - y_i + \varepsilon + \xi_i] \\ &- \sum_{i=1}^{P} \alpha_i' [y_i - \mathbf{w}^T \boldsymbol{\varphi}(\mathbf{x}_i) - b + \varepsilon + \xi_i'] - \sum_{i=1}^{P} (\gamma_i \xi_i + \gamma_i' \xi_i') \end{aligned}$$



Nonlinear regression

SVM regression

$$\mathbf{w} = \sum_{i=1}^{P} \left(\alpha_i - \alpha_i'\right) \boldsymbol{\varphi}(\mathbf{x}_i)$$

$$\gamma_i = C - \alpha_i$$

$$\gamma_i' = C - \alpha_i'$$

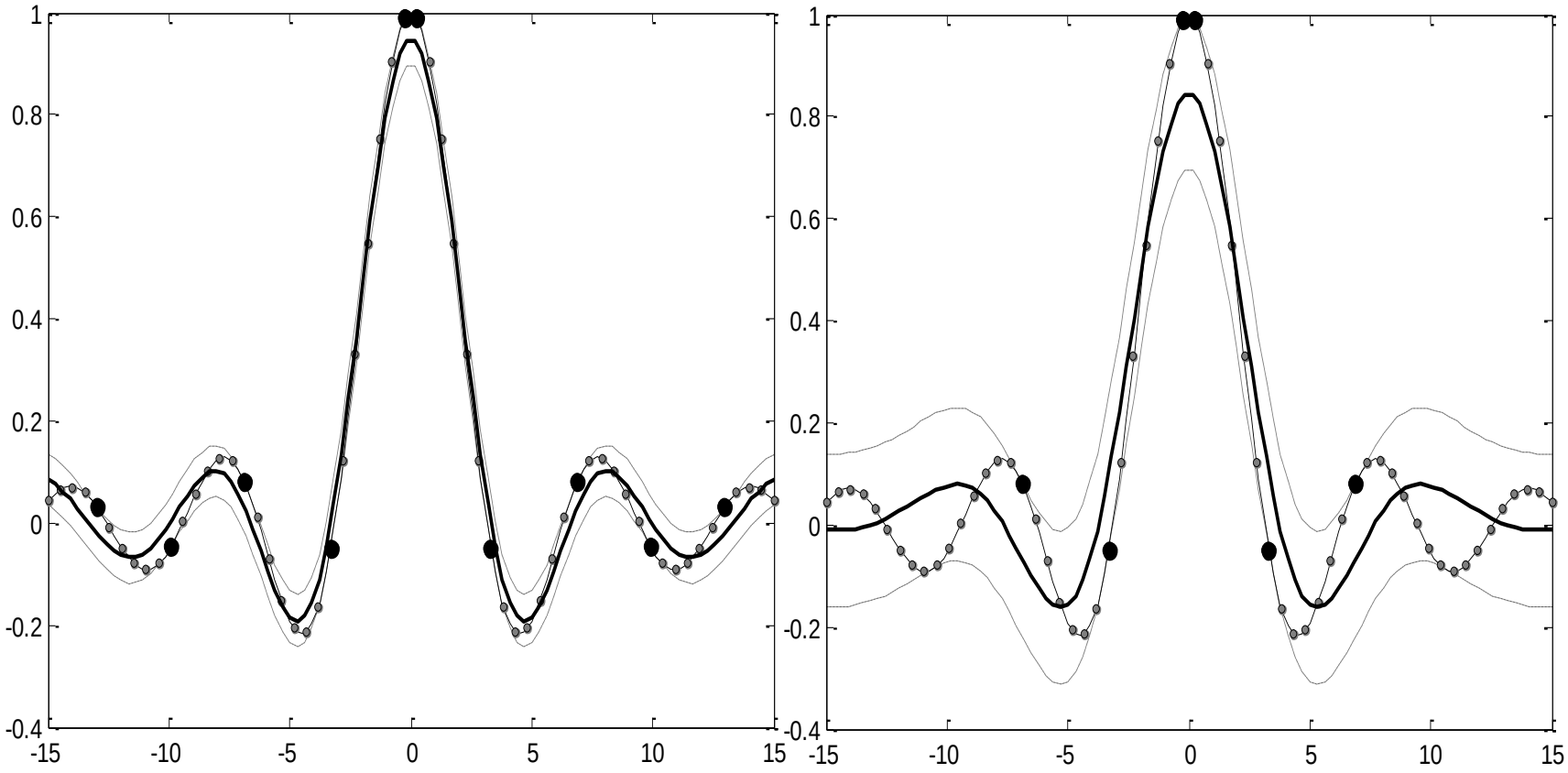
$$Q(\boldsymbol{\alpha}, \boldsymbol{\alpha}') = \sum_{i=1}^{P} d_i \left(\alpha_i - \alpha_i'\right) - \varepsilon \sum_{i=1}^{P} \left(\alpha_i + \alpha_i'\right)$$

$$-\frac{1}{2} \sum_{i=1}^{P} \sum_{j=1}^{P} \left(\alpha_i - \alpha_i'\right)\left(\alpha_j - \alpha_j'\right) K\left(\mathbf{x}_i, \mathbf{x}_j\right)$$

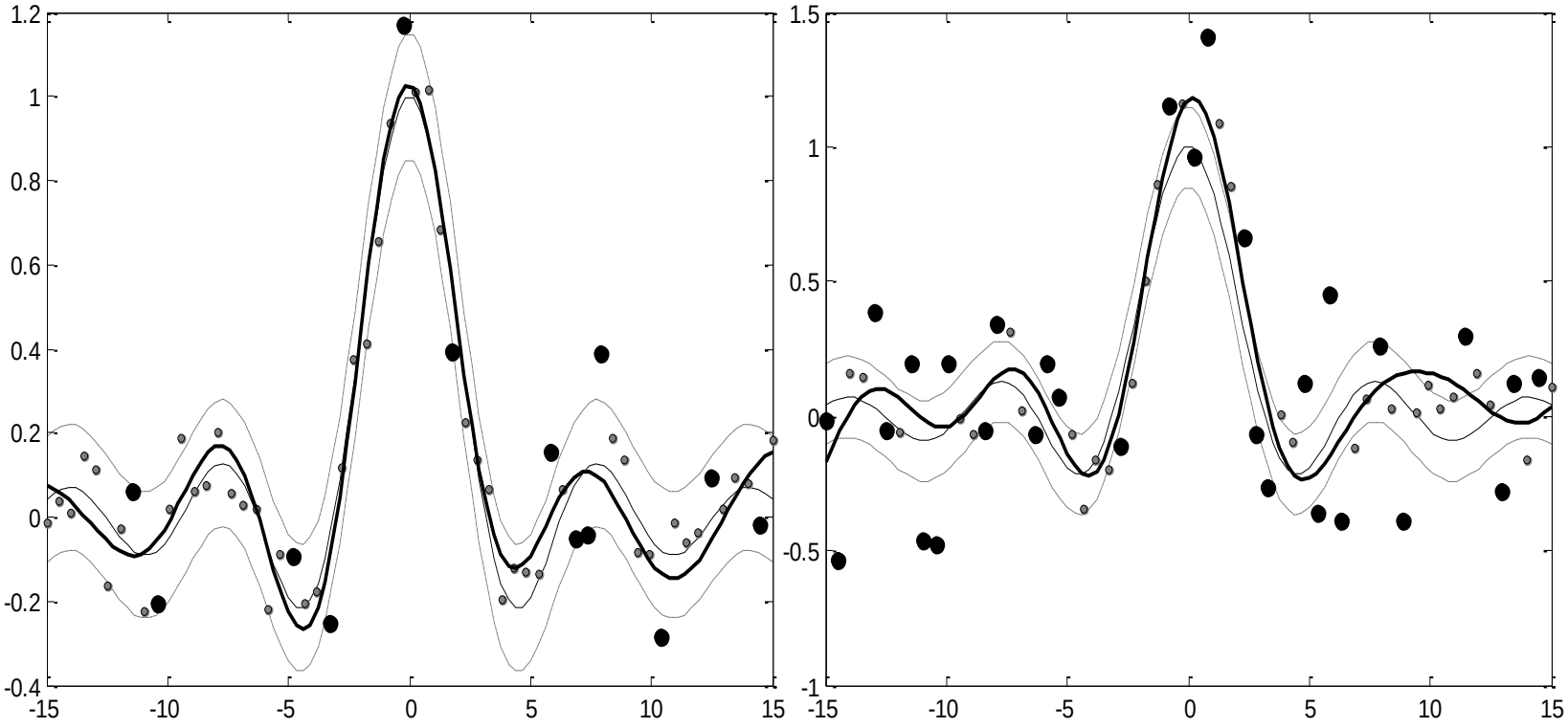
$$\sum_{i=1}^{P} \left(\alpha_i - \alpha_i'\right) = 0, \quad 0 \le \alpha_i \le C, \quad 0 \le \alpha_i' \le C, \quad i = 1,2,\ldots.P$$

$$y(\mathbf{x}) = \sum_{i=1}^{P} \left(\alpha_i - \alpha_i'\right) K\left(\mathbf{x}, \mathbf{x}_i\right) + b$$

Nonlinear regression



Nonlinear regression



Nonlinear regression

SVM variants: LS-SVM: seeking of conditional extremum using Lagrange multipliers

$$\min_{\mathbf{w},b,\mathbf{e}} J(\mathbf{w},b,\mathbf{e}) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\frac{1}{2}\sum_{i=1}^{P} e_i^2 \qquad y(\mathbf{x}) = \mathbf{w}^T\boldsymbol{\varphi}(\mathbf{x}) + b$$

$$d_i = \mathbf{w}^T\boldsymbol{\varphi}(\mathbf{x}_i) + b + e_i \qquad i = 1,\ldots,P$$

Lagrangian

$$L(\mathbf{w},b,\mathbf{e};\boldsymbol{\alpha}) = J(\mathbf{w},b,\mathbf{e}) - \sum_{i=1}^{P}\alpha_i\left\{\mathbf{w}^T\boldsymbol{\varphi}(\mathbf{x}_i) + b + e_i - d_i\right\}$$

$$\frac{\partial L}{\partial \mathbf{w}} = 0 \quad \rightarrow \quad \mathbf{w} = \sum_{i=1}^{P}\alpha_i\boldsymbol{\varphi}(\mathbf{x}_i)$$

$$\frac{\partial L}{\partial b} = 0 \quad \rightarrow \quad \sum_{i=1}^{P}\alpha_i = 0$$

$$\frac{\partial L}{\partial e_i} = 0 \quad \rightarrow \quad \alpha_i = C\,e_i \qquad i = 1,\ldots,P$$

$$\frac{\partial L}{\partial \alpha_i} = 0 \quad \rightarrow \quad \mathbf{w}^T\boldsymbol{\varphi}(\mathbf{x}_i) + b + e_i - d_i = 0 \qquad i = 1,\ldots,P$$

$$\begin{bmatrix} 0 & \mathbf{1}^T \\ \mathbf{1} & \boldsymbol{\Omega} + C^{-1}\mathbf{I} \end{bmatrix}\begin{bmatrix} b \\ \boldsymbol{\alpha} \end{bmatrix} = \begin{bmatrix} 0 \\ \mathbf{d} \end{bmatrix}$$

$$\Omega_{i,j} = \boldsymbol{\varphi}(\mathbf{x}_i)^T\boldsymbol{\varphi}(\mathbf{x}_j) = K(\mathbf{x}_i,\mathbf{x}_j)$$

$$y(\mathbf{x}) = \sum_{i=1}^{P}\alpha_i K(\mathbf{x},\mathbf{x}_i) + b$$