Support Vector Regression (SVR)

Support Vector Regression: Principle

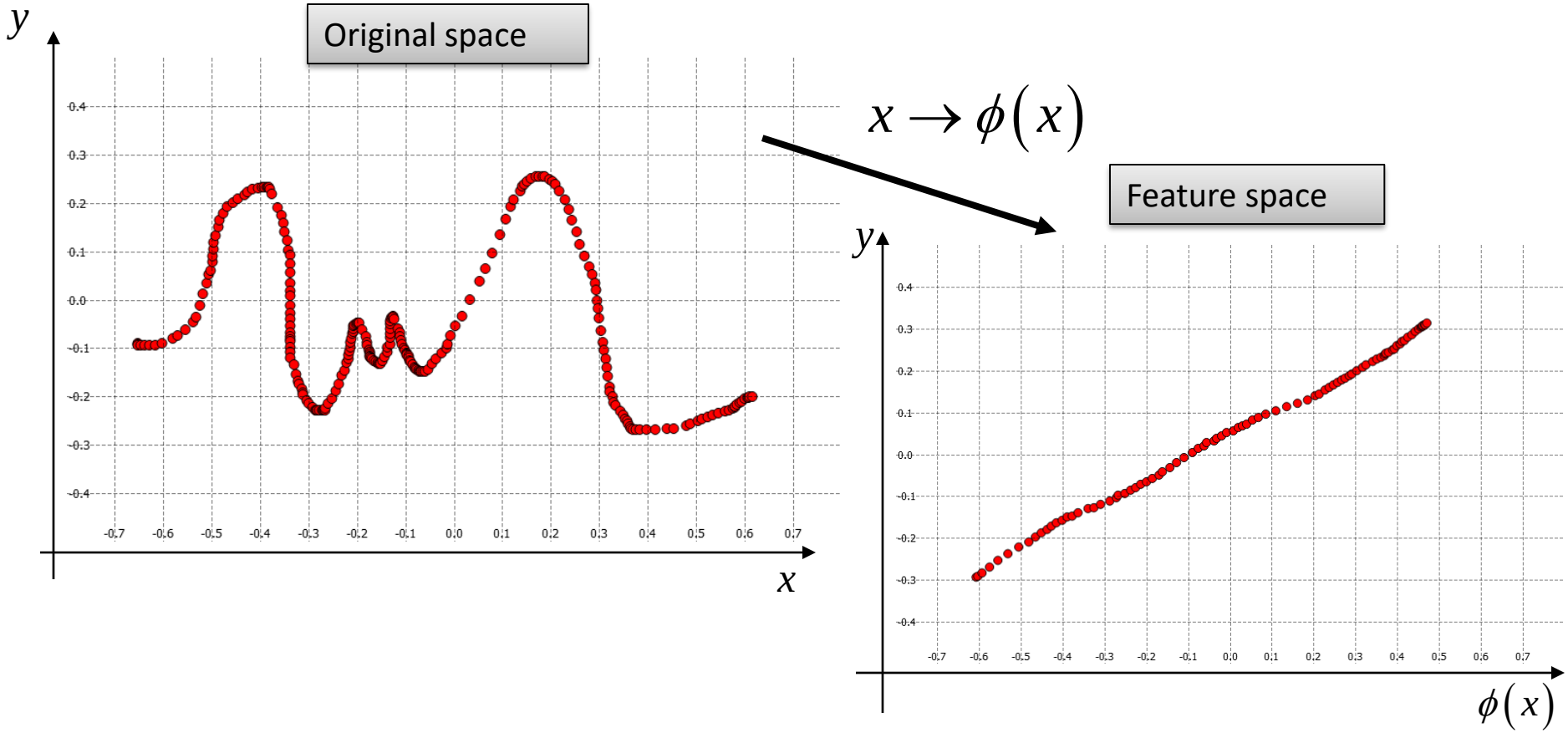
We seek to find a function f , s.t. $y = f(x)$.

$y \in \mathbb{R}, \quad x \in \mathbb{R}^N$

Generalize the support vector machine framework for classification to estimate continuous functions

Assume a non-linear mapping through feature space and then perform *linear regression* in feature space.

Support Vector Regression: Principle



Principle: Assume there exists a transformation ϕ such that the problem becomes linear

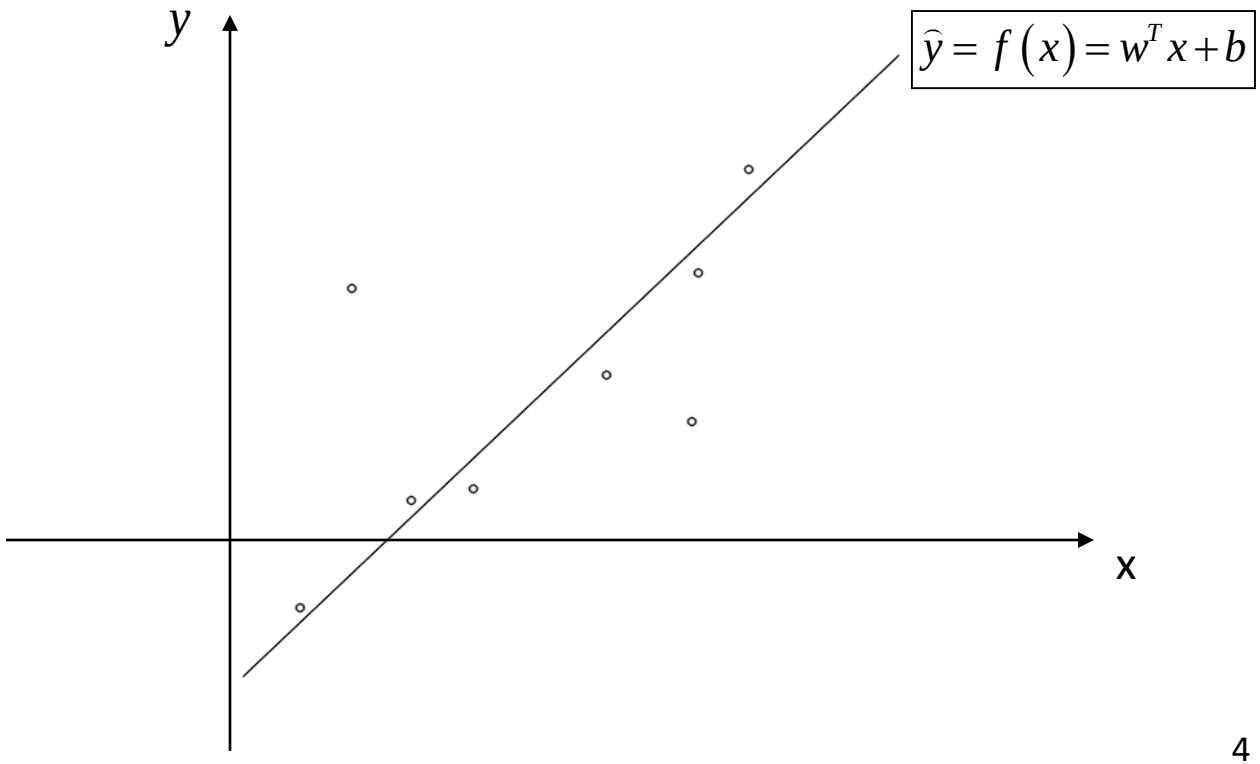
→ Perform linear regression in feature space

Support Vector Regression: Linear Case

Assume a linear mapping f , s.t. $y = f(x) = w^T x + b.$

$x \in \mathbb{R}^N, y \in \mathbb{R}$

Need to estimate w and b to best predict the pair of training points $\{x^i, y^i\}_{i=1,\ldots M}$.

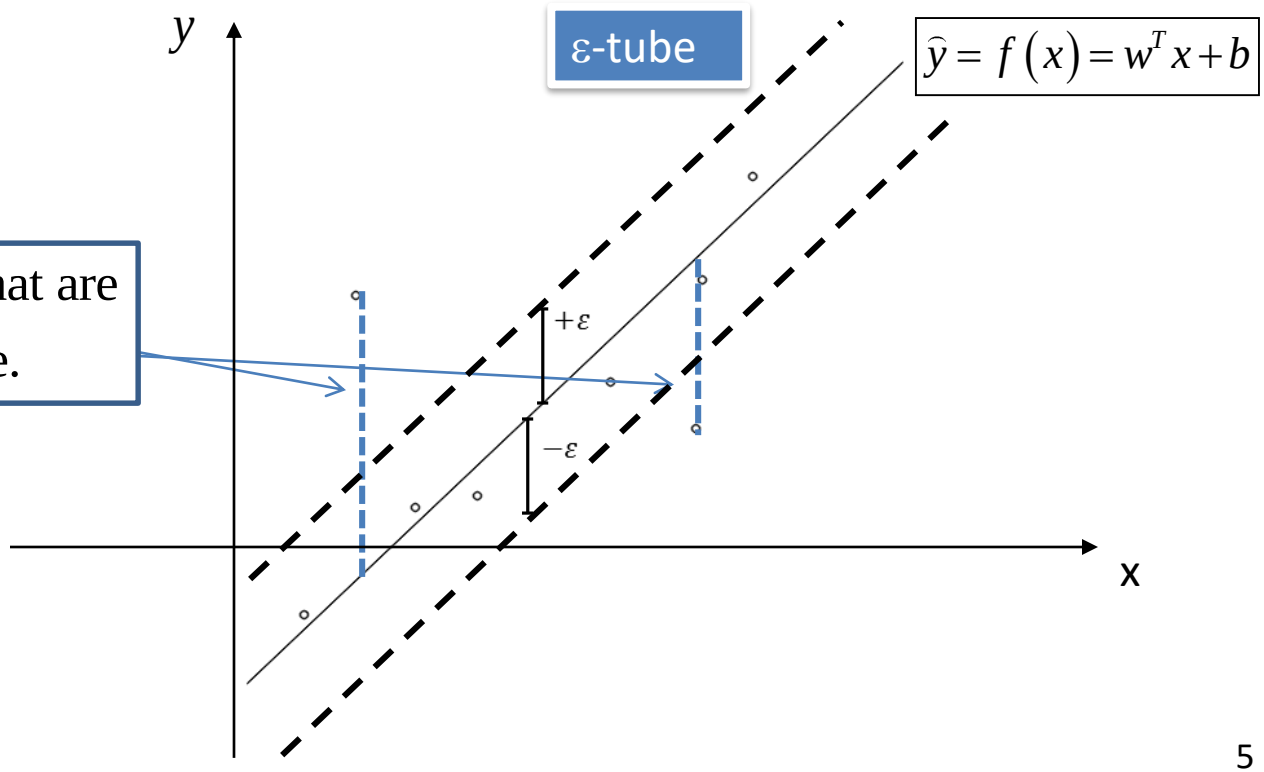


Support Vector Regression: ε-tube

Assume deterministic noise model: $y \pm \varepsilon$

Consider as correctly fit all points such that $f(x) - y \leq \varepsilon$.

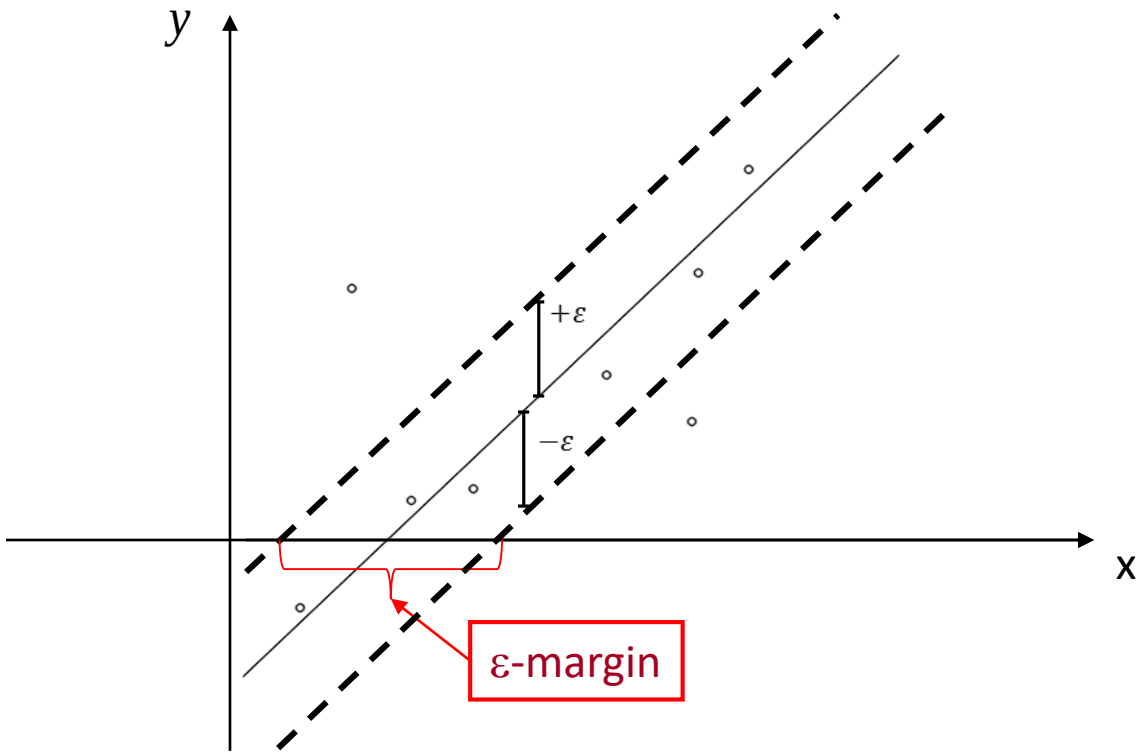
Penalize only datapoints that are not contained in the ε-tube.



Support Vector Regression: ε-margin

How to assess how well we do for a choice of ε-tube.

The ε-margin is a measure of the width of the ε-insensitive tube.
It is a measure of the precision of the regression.

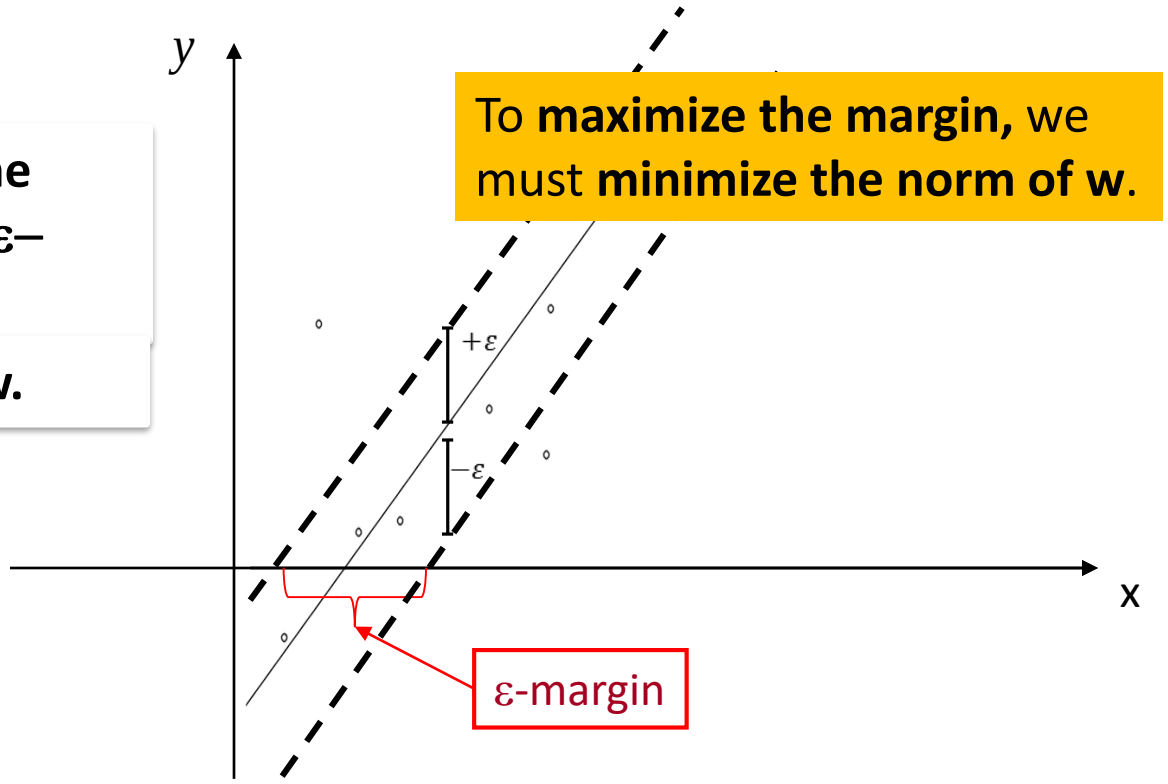


Support Vector Regression: ε-margin

The ε-margin is a measure of the width of the ε-insensitive tube.
It is a measure of the precision of the regression.

The flatter the slope of the function *f*, the larger the ε–margin

the smaller the norm of w.



To **maximize the margin,** we must **minimize the norm of w.**

Support Vector Regression: ε-margin

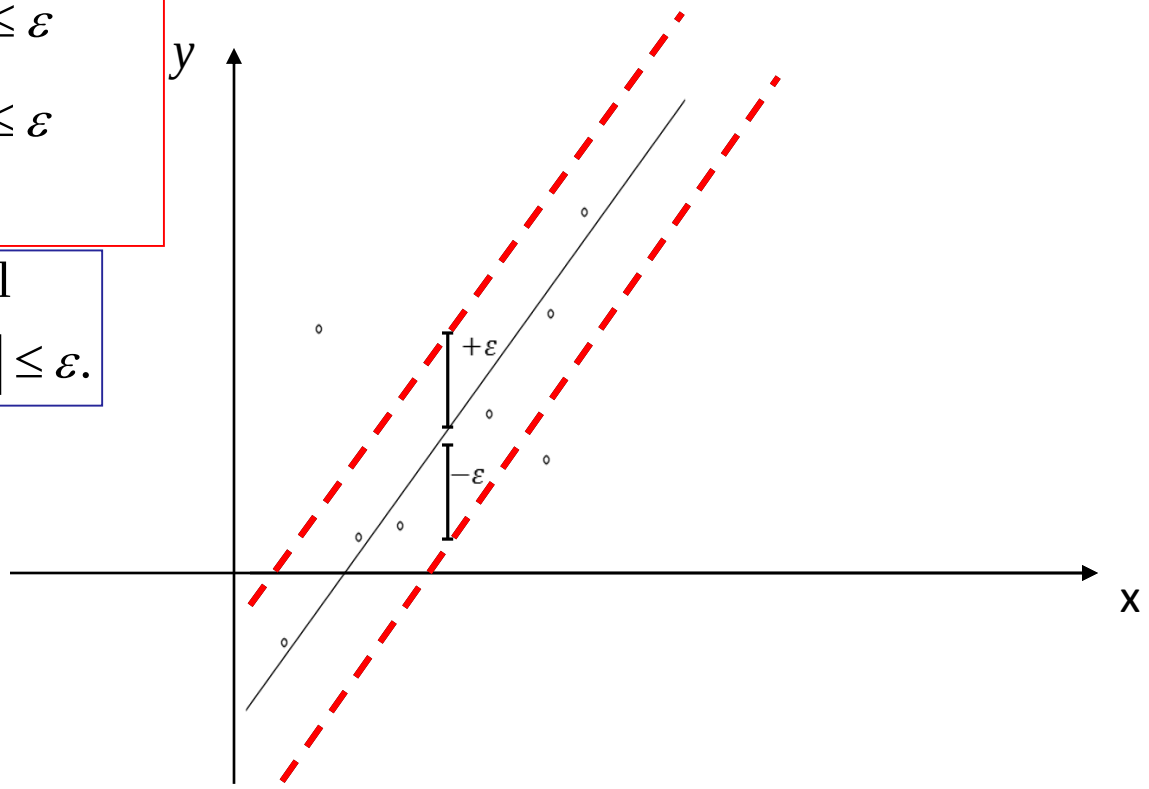
This can be rephrased as a constraint-based optimization problem:

minimize $\frac{1}{2}\|w\|^2$

subject to $\begin{cases} \langle w, x^i \rangle + b - y^i \leq \varepsilon \\ y^i - \langle w, x^i \rangle - b \leq \varepsilon \end{cases}$

$\forall i = 1, \ldots M$

Consider as correctly fit all points such that $|f(x) - y| \leq \varepsilon.$



Support Vector Regression: ε-margin

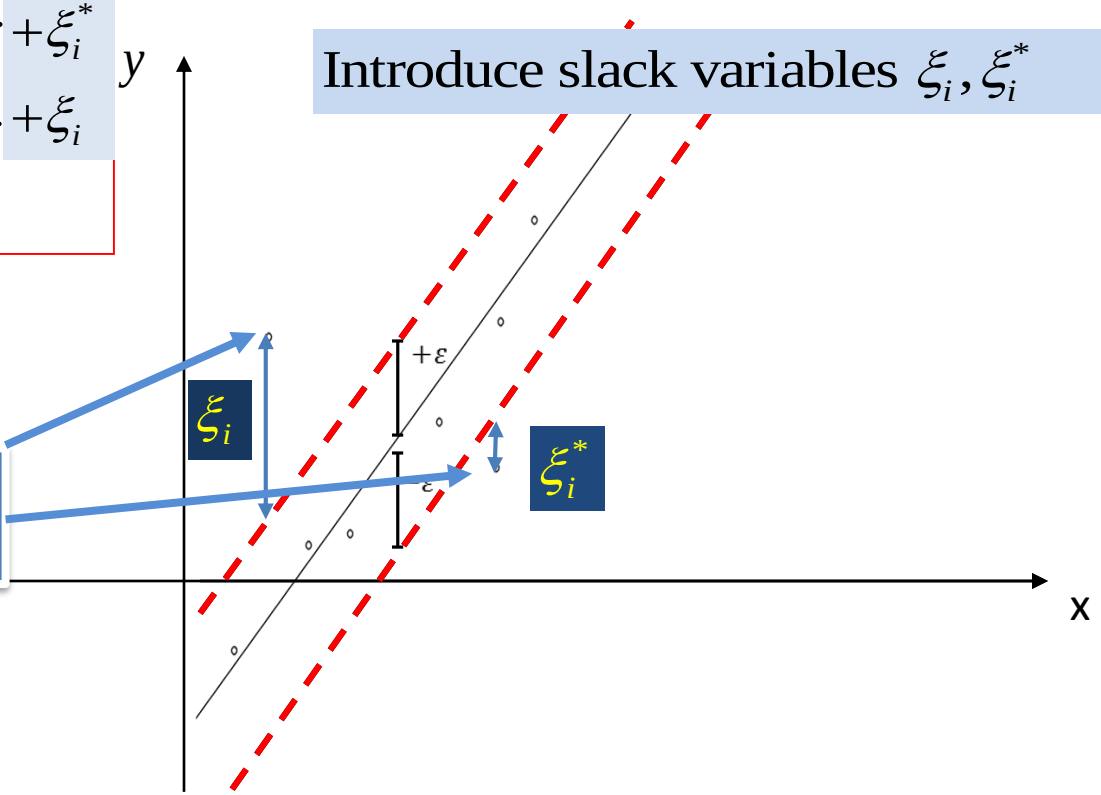
This can be rephrased as a constraint-based optimization problem:

$$\text{minimize } \frac{1}{2}\|w\|^2 + \frac{C}{M}\sum_{i=1}^{M}\left(\xi_i + \xi_i^*\right), \qquad C > 0$$

$$\text{subject to } \begin{cases} \langle w, x^i \rangle + b - y^i \le \varepsilon + \xi_i^* \\ y^i - \langle w, x^i \rangle - b \le \varepsilon + \xi_i \end{cases}$$

$$\forall i = 1, \ldots M \qquad \xi_i \ge 0, \quad \xi_i^* \ge 0$$

Introduce slack variables ξ_i, ξ_i^*



Need to penalize points outside the ε-insensitive tube.

Support Vector Regression: optimization

We can solve this quadratic problem by introducing sets of $\alpha,\ \eta \in \mathbb{R}$ Lagrange multipliers and writing the Lagrangian :

Lagrangian = Objective function + multipliers * constraints

$$\mathrm{L}\left(w,\xi,\xi^{*},b\right)=\frac{1}{2}\|w\|^{2}+\frac{\mathrm{C}}{M}\sum_{i=1}^{M}\left(\xi_{i}+\xi_{i}^{*}\right)-\frac{\mathrm{C}}{M}\sum_{i=1}^{M}\left(\eta_{i}\xi_{i}+\eta_{i}^{*}\xi_{i}^{*}\right)$$
$$-\sum_{i=1}^{M}\alpha_{i}\left(\varepsilon+\xi_{i}-y^{i}+\left\langle w,x^{i}\right\rangle+b\right)$$
$$-\sum_{i=1}^{M}\alpha_{i}^{*}\left(\varepsilon+\xi_{i}^{*}+y^{i}-\left\langle w,x^{i}\right\rangle-b\right)$$

Support Vector Regression: solution

Requiring that the partial derivatives are all zero:

$$\frac{\partial \mathrm{L}}{\partial b} = \sum_{i=1}^{M} \left(\alpha_i - \alpha_i^* \right) = 0. \qquad \rightarrow \sum_{i=1}^{M} \alpha_i = \sum_{i=1}^{M} \alpha_i^*$$

Rebalancing the effect of the support vectors on both sides of the ε-tube

$$\frac{\partial \mathrm{L}}{\partial w} = w - \sum_{i=1}^{M} \left(\alpha_i - \alpha_i^* \right) x^i = 0. \qquad \Rightarrow w = \sum_{i=1}^{M} \left(\alpha_i - \alpha_i^* \right) x^i.$$

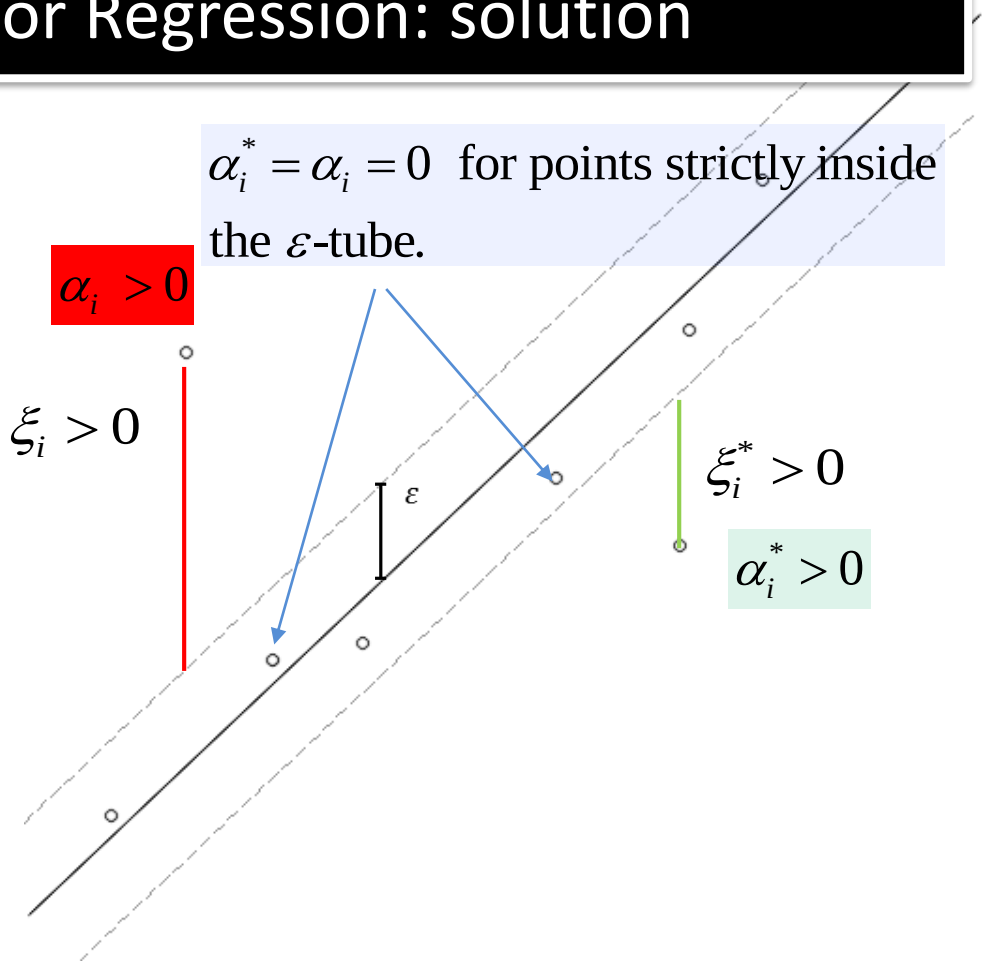
Linear combination of support vectors

$$\begin{aligned} y &= f(x) \\ &= \langle w, x \rangle + b \\ &= \sum_{i=1}^{M} \left(\alpha_i - \alpha_i^* \right) \langle x^i, x \rangle + b \end{aligned}$$

α_i or $\alpha_i^* > 0$ for points outside ε-tube

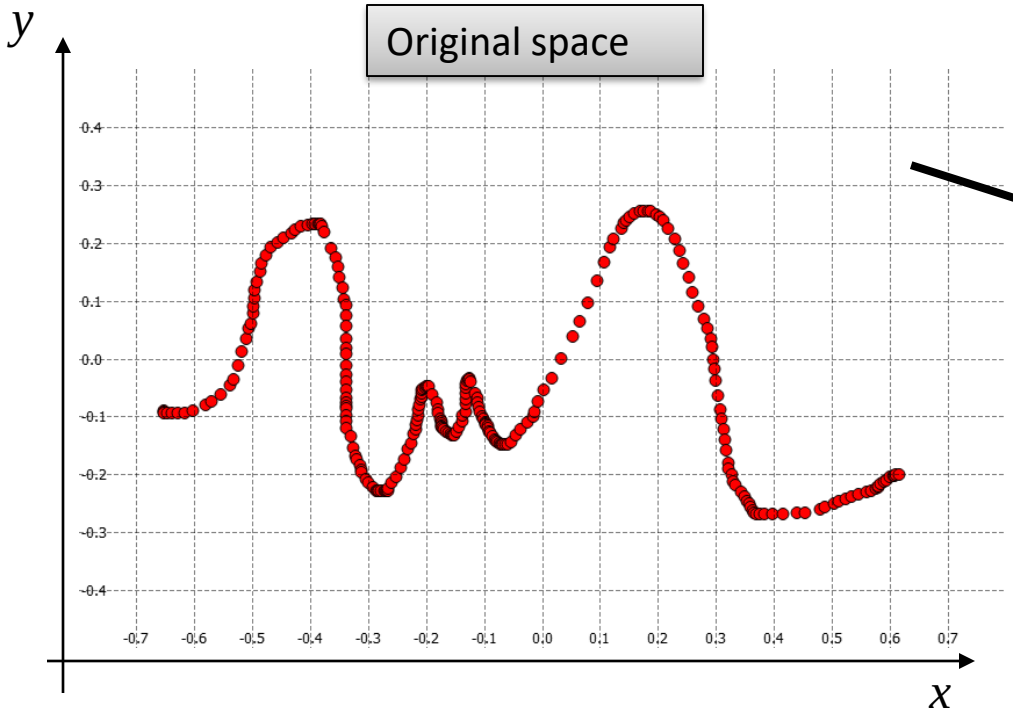
Support Vector Regression: solution

$\alpha_i^* = \alpha_i = 0$ for points strictly inside the ε-tube.

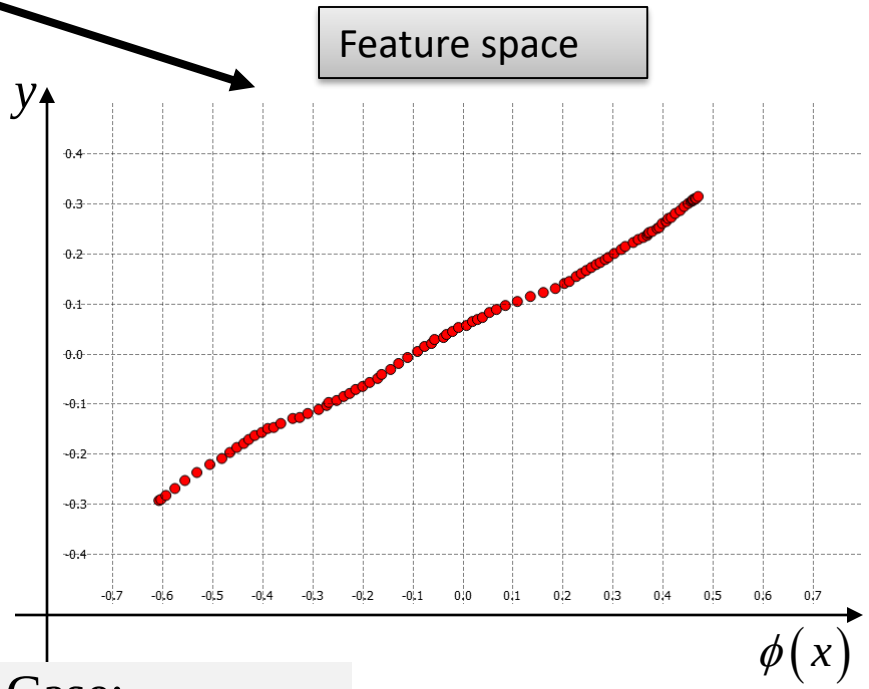


α_i or $\alpha_i^* > 0$ for points outside ε-tube

Support Vector Regression: non-linear regression



$$x \rightarrow \phi(x)$$



Linear Case:

$$y = f(x) = \langle w, x \rangle + b$$

Non-Linear Case:

$$x \rightarrow \phi(x)$$

$$y = f(\phi(x)) = \langle w, \phi(x) \rangle + b$$

w lives in feature space!

Support Vector Regression: non-linear regression

Replacing in the primal Lagrangian, we get the Dual optimization:

$$\max_{\alpha,\alpha^*} \begin{cases} -\frac{1}{2}\sum_{i,j=1}^{M}\left(\alpha_i^* - \alpha_i\right)\left(\alpha_j^* - \alpha_j\right)\left\langle \phi\left(x^i\right), \phi\left(x^j\right)\right\rangle \\ -\varepsilon\sum_{i=1}^{M}\left(\alpha_i^* + \alpha_i\right) + \sum_{i=1}^{M} y^i\left(\alpha_i^* + \alpha_i\right) \end{cases}$$

$$\text{subject to } \sum_{i=1}^{M}\left(\alpha_i^* - \alpha_i\right) = 0 \text{ and } \alpha_i^*, \alpha_i{}^i \in [0, C]$$

Kernel Trick

$$k\left(x^i, x^j\right) = \left\langle \phi\left(x^i\right), \phi\left(x^j\right)\right\rangle$$

Support Vector Regression: non-linear regression

The solution is given by:

$$y = f(x) = \sum_{i=1}^{M} (\alpha_i - \alpha_i^*) k(x^i, x) + b$$

An estimate of b can be computed from the KKT conditions:

$$\Rightarrow b = \frac{1}{M} \sum_{j=1}^{M} \left(y^j - \sum_{i=1}^{M} (\alpha_i - \alpha_i^*) k(x^j, x^i) \right)$$

using only the SVs on the ε-tube.

Support Vector Regression: non-linear regression

The solution is given by:

$$y = f(x) = \sum_{i=1}^{M} (\alpha_i - \alpha_i^*) k(x^i, x) + b$$

Linear Coefficients (Lagrange multipliers for each constraint).

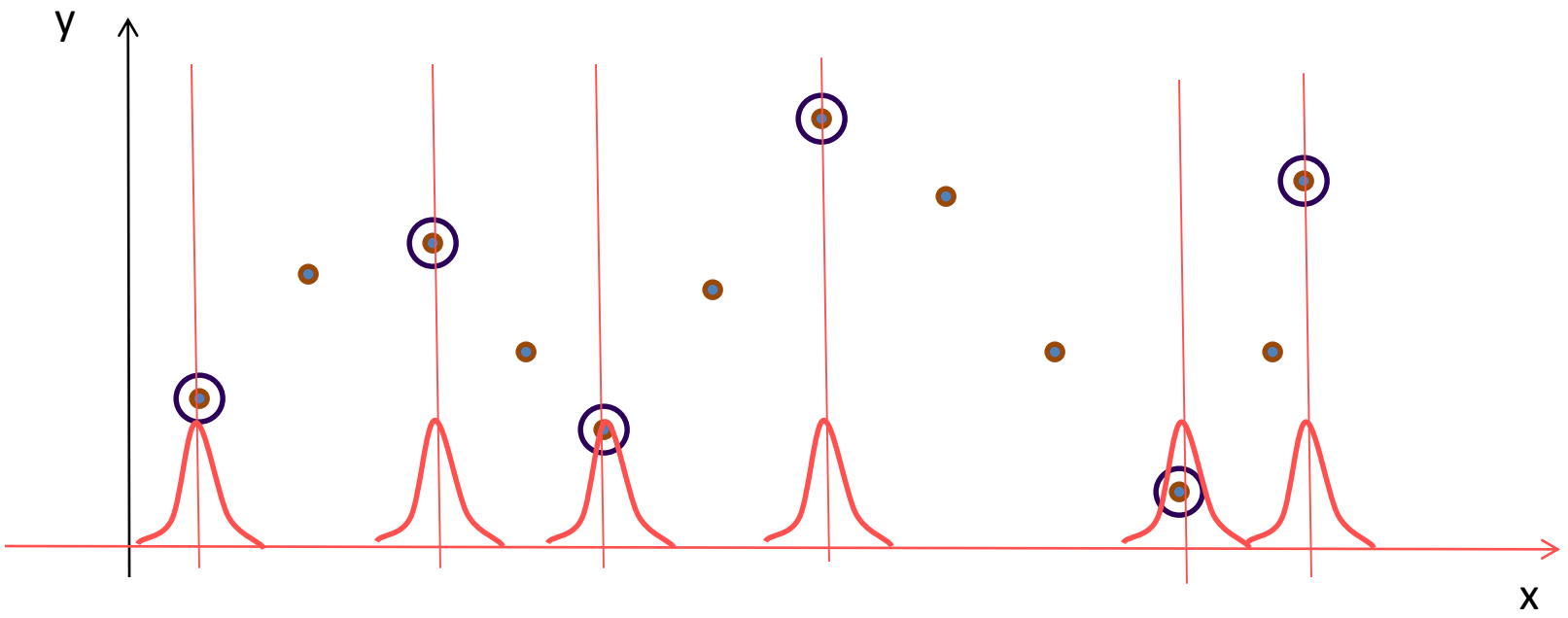
If one uses RBF Kernel, *M* un-normalized isotropic Gaussians centered on each training datapoint.

Support Vector Regression: interpretation

The solution is given by:

$$y = f(x) = \sum_{i=1}^{M} (\alpha_i - \alpha_i^*) k(x^i, x) + b$$

Kernel places a Gauss function on each SV



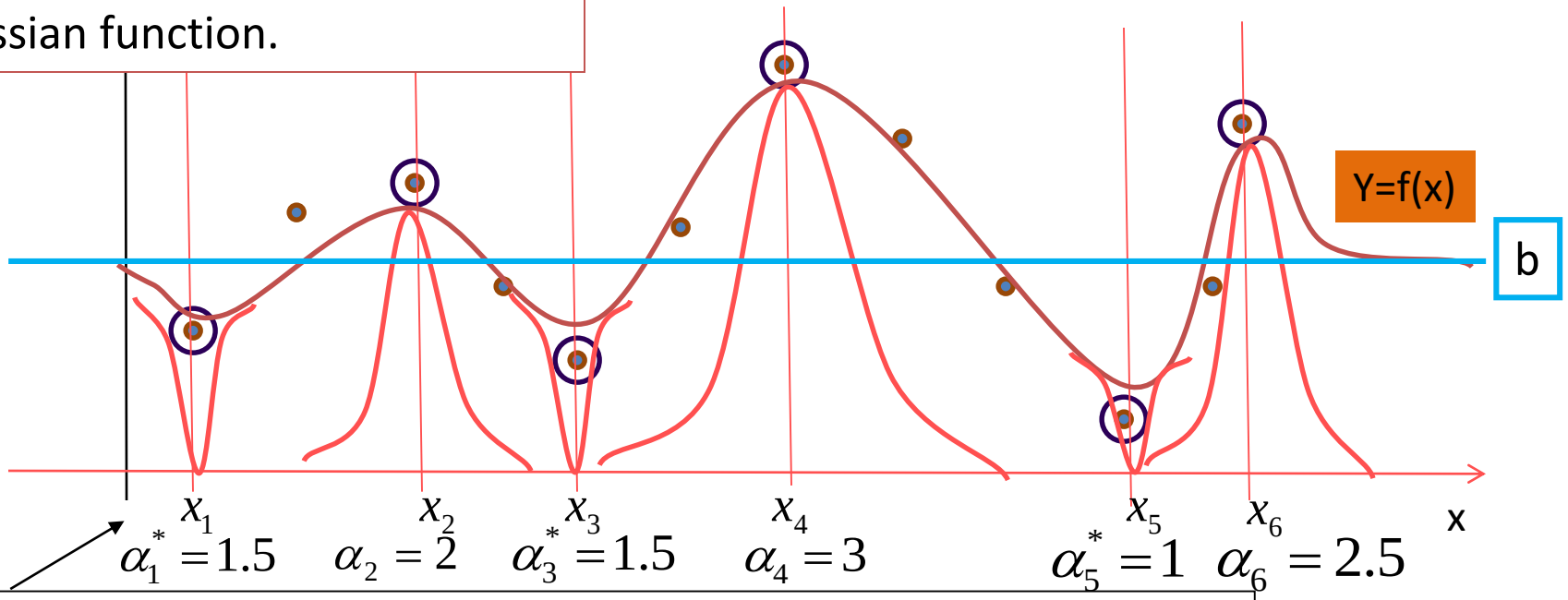
Support Vector Regression: interpretation

The solution is given by:

$$y = f(x) = \sum_{i=1}^{M} (\alpha_i - \alpha_i^*) k(x^i, x) + b$$

Converges to b when SV effect vanishes.

The Lagrange multipliers define the importance of each Gaussian function.



For illustrative purpose, we plot the negative Gauss function next to the SV, but they are distributed on the negative y axis.

ε–SVR: 3 Hyperparameters

The solution to SVR we just saw is referred to as ε–SVR

Two Hyperparameters for optimization

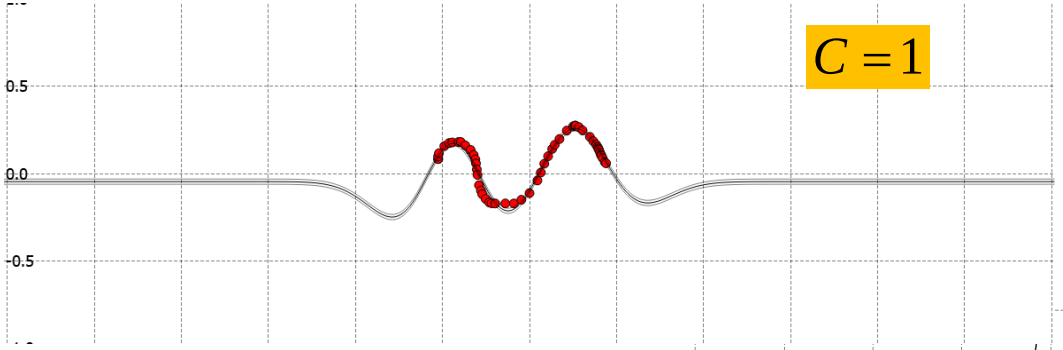
$$\text{minimize} \quad \frac{1}{2}\|w\|^2 + \frac{C}{M}\sum_{i=1}^{M}\left(\xi_i + \xi_i^*\right)$$

$$\text{subject to} \quad \begin{cases} \langle w, x^i \rangle + b - y^i \leq \varepsilon + \xi_i^* \\ y^i - \langle w, x^i \rangle - b \leq \varepsilon + \xi_i \\ \xi_i \geq 0, \quad \xi_i^* \geq 0 \end{cases}$$

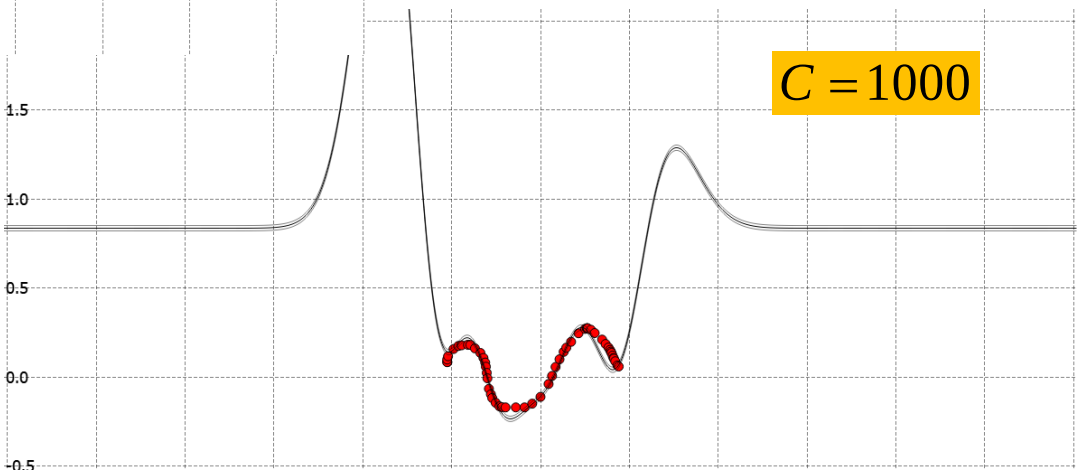
1. C controls the penalty term on poor fit.
2. ε determines the minimal required precision for the fit.
3. The kernel width for the RBF kernel determines the locality of the regression

Hyperparameters' Influence

$\alpha_i, \alpha_i^* \in [0, C]$



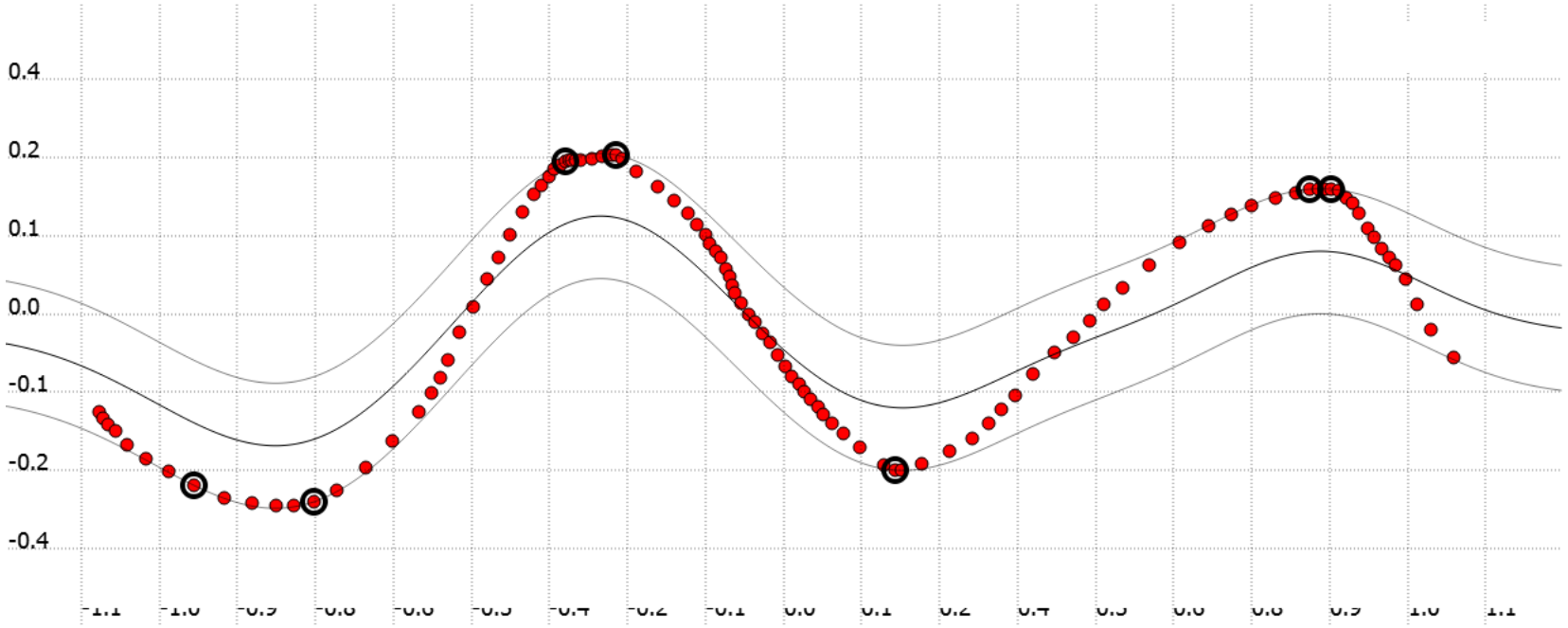
$C = 1$



$C = 1000$

Effect of C on the fit.

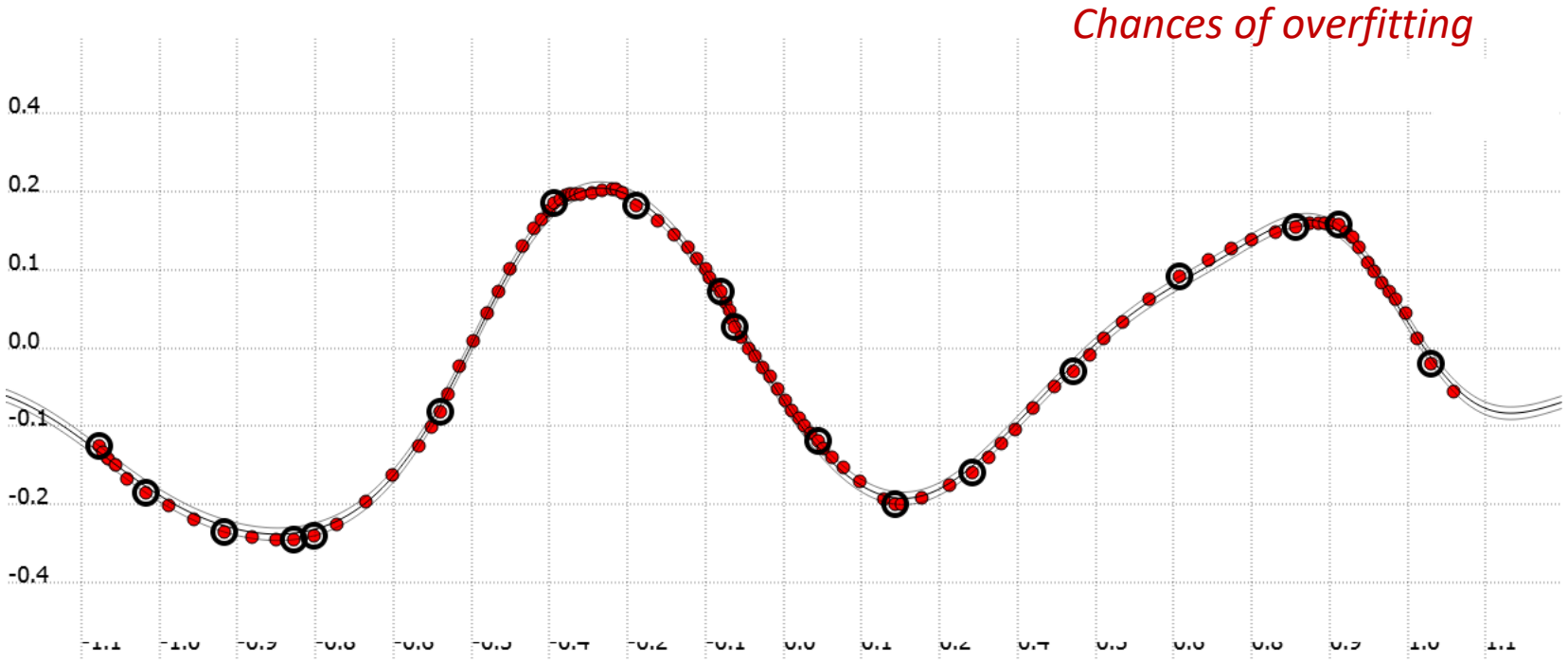
Hyperparameters' Influence



Here fit using C=1, ε=0.1, kernel width=0.01.

Effect of the ε on the fit.

Hyperparameters' Influence

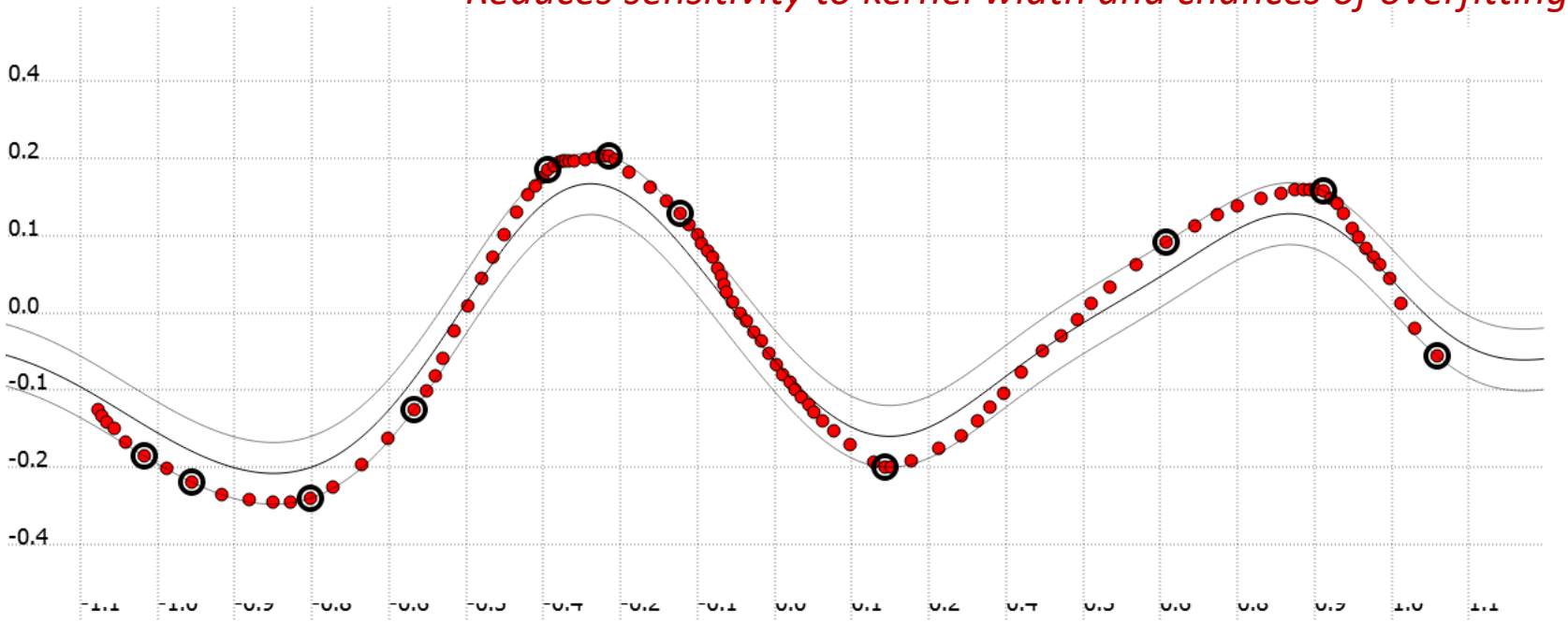


Here fit using C=1, ε=0.01, kernel width=0.01

Effect of the ε on the fit.

Hyperparameters' Influence

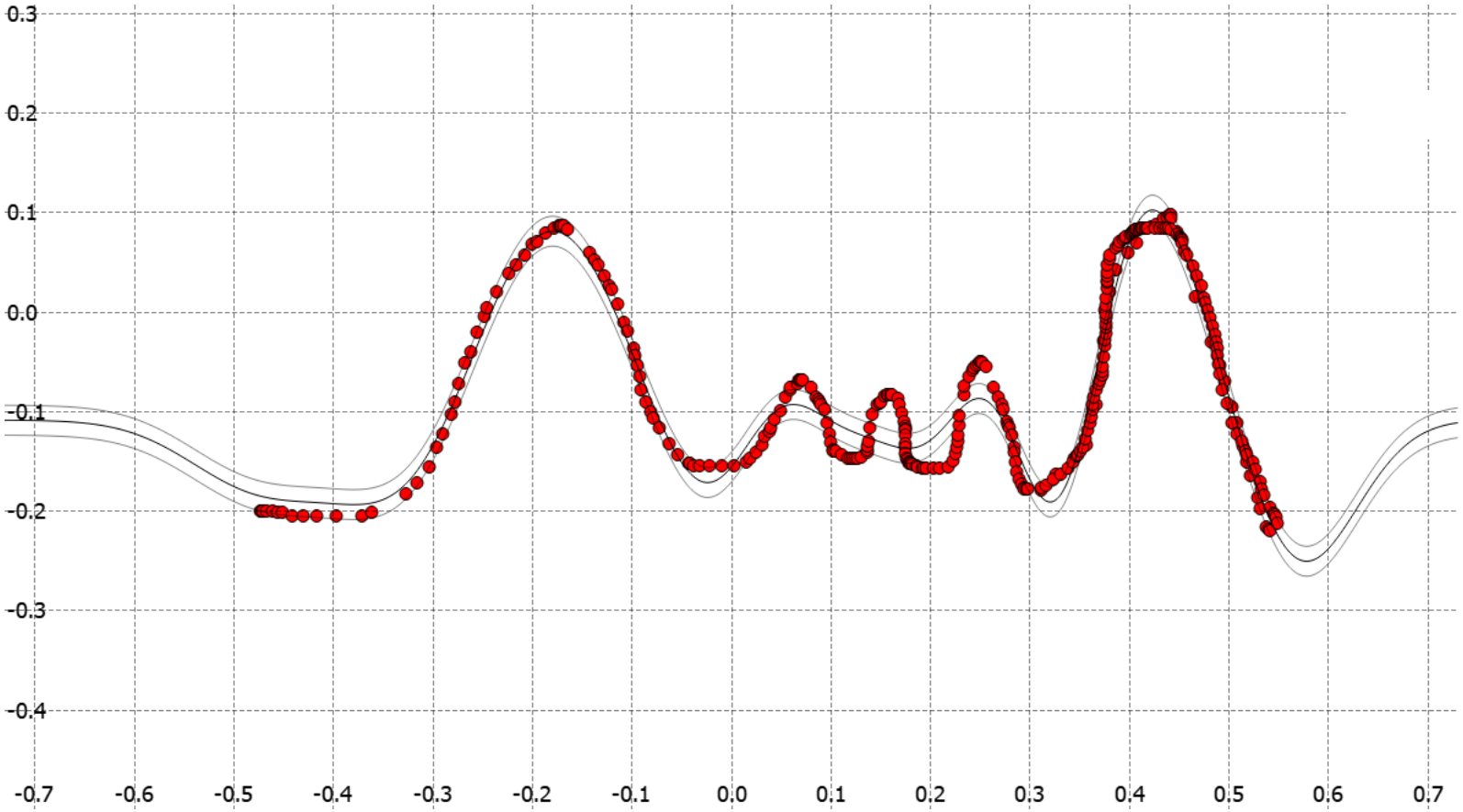
Reduces sensitivity to kernel width and chances of overfitting



Here fit using C=1, ε=0.05, kernel width=0.01

Effect of the ε on the fit.

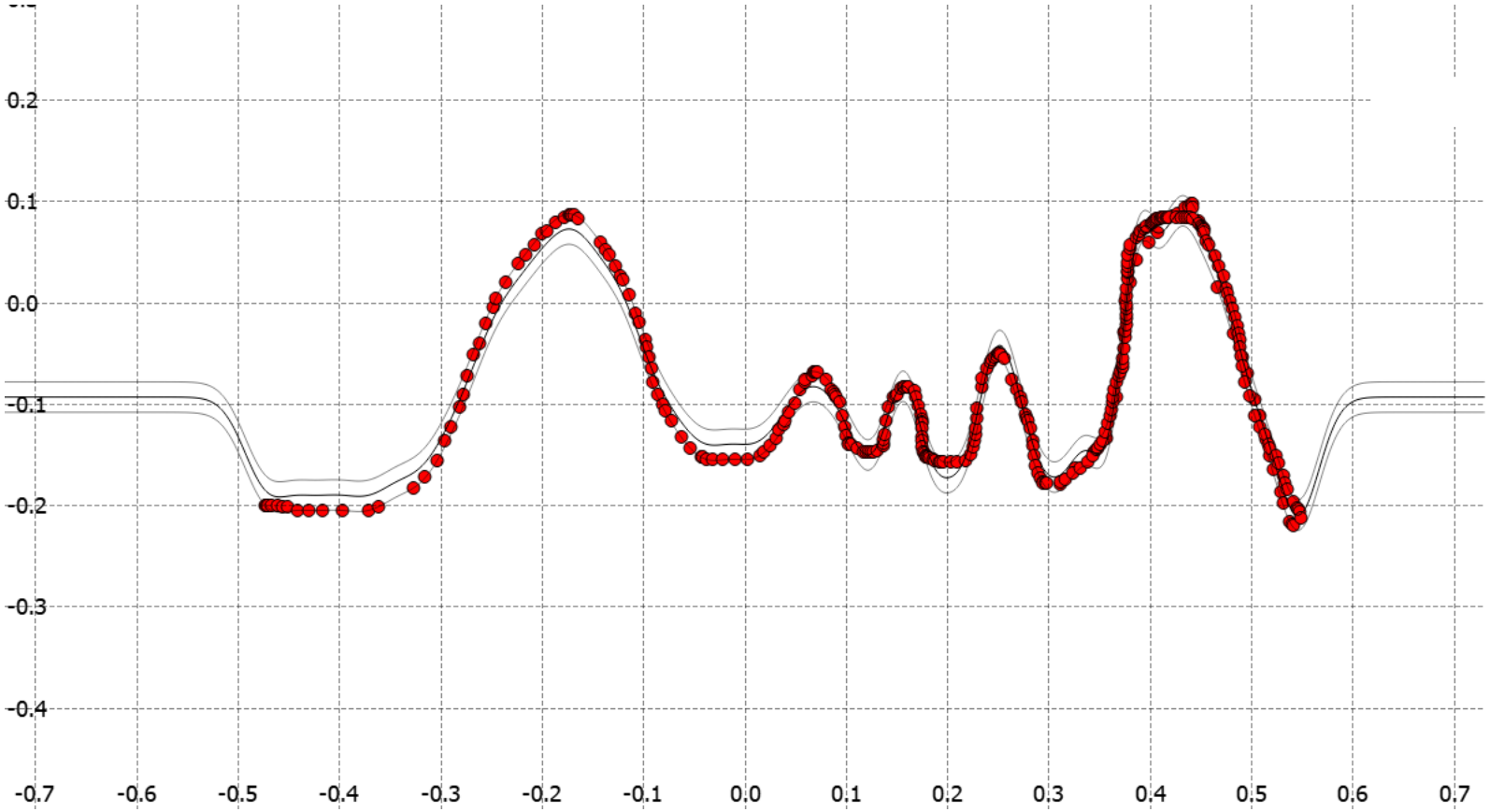
Hyperparameters' Influence



Kernel width=0.05. Too large the fit.

Effect of the RBF kernel width on the fit.

Hyperparameters' Influence



Kernel width=0.001. Perfect fit.

Effect of the RBF kernel width on the fit.

Summary

Linear regression can be solved through Least-Mean-Square estimation and yields an optimal analytical solution.

Weighted regression offers the possibility to perform a local regression and yields also an optimal analytical solution.

The estimate is no longer global and is computed around each group of data point!

Support Vector Regression: performs regression on a non-linear function. Determines automatically the important points. The estimate is globally optimal.