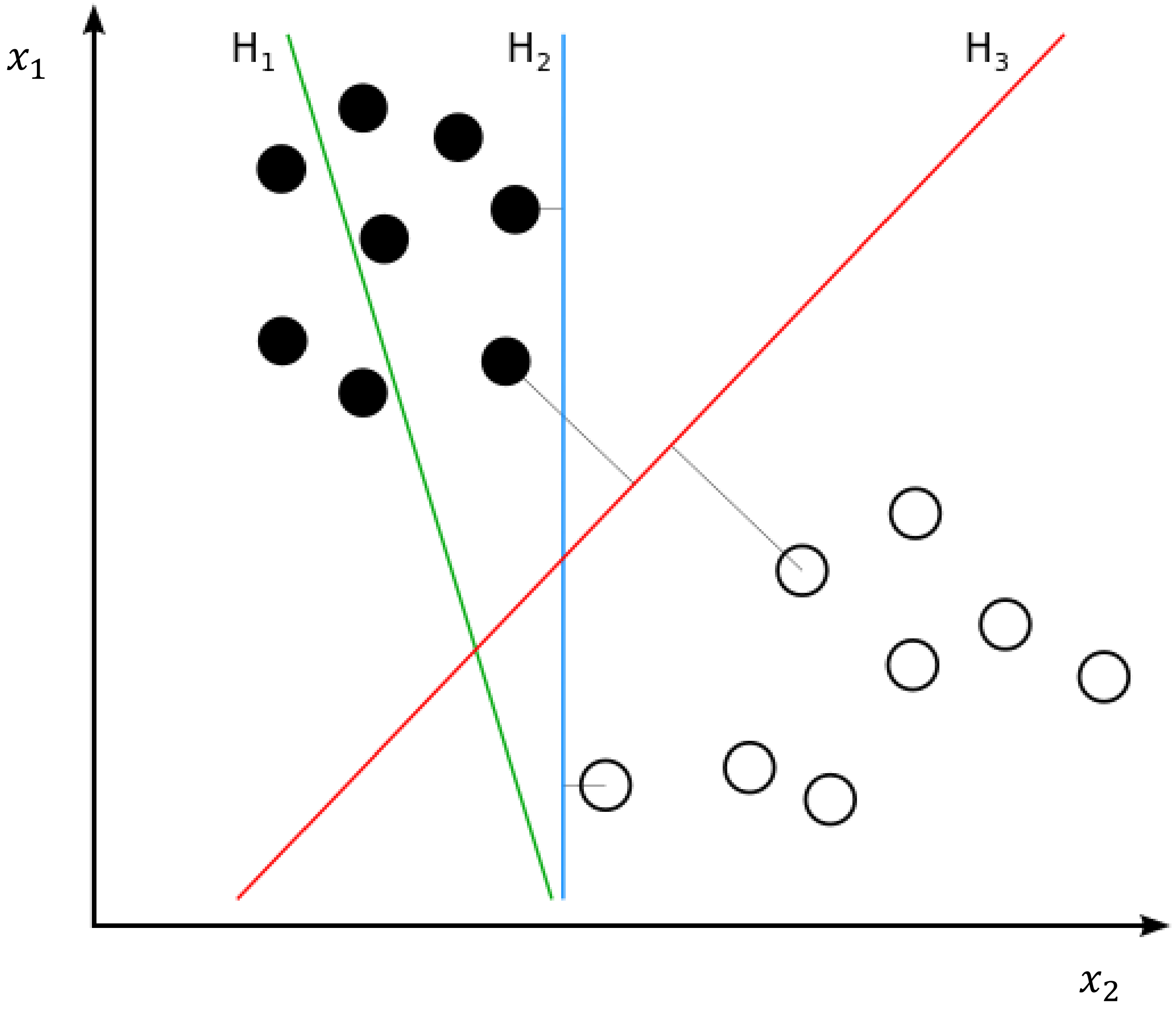
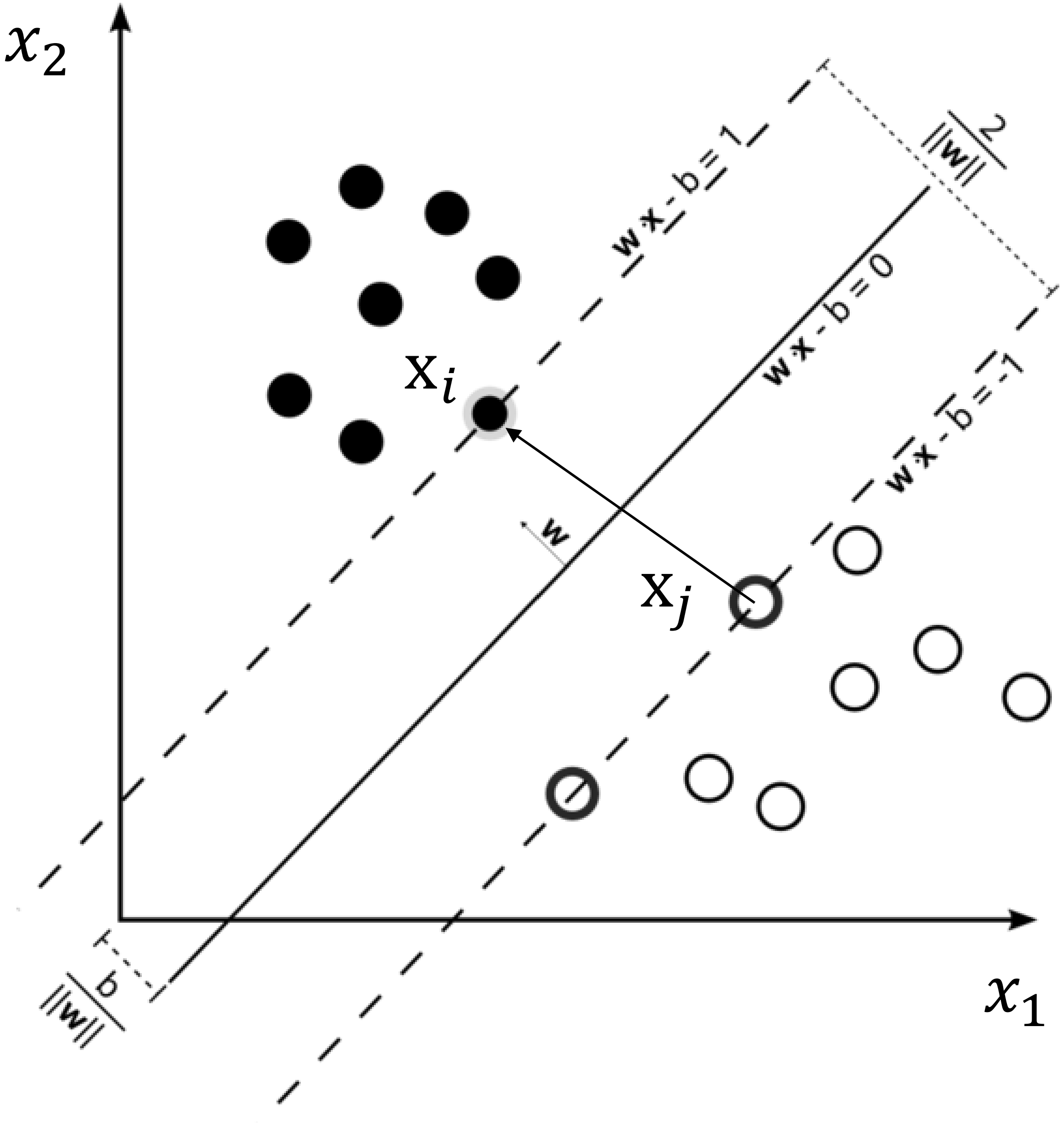
SVM

Support Vector Machines

SVM (Support Vector Machines)



Maximize the margin
Linearly separable case



Scale **w** and ***b*** *so that:*

$$\min\left|\mathbf{w}^{\mathrm{T}}\mathbf{x}_i - \boldsymbol{b}\right| = \min\left(\boldsymbol{y}_i\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i - \boldsymbol{b}\right)\right) = \mathbf{1},$$

Then the width of the margin is:

$$\frac{\mathbf{w}^{\mathrm{T}}\left(\mathbf{x}_i - \mathbf{x}_j\right)}{||\mathbf{w}||} = \frac{\mathbf{w}^{\mathrm{T}}\mathbf{x}_i - \boldsymbol{b} - \left(\mathbf{w}^{\mathrm{T}}\mathbf{x}_j - \boldsymbol{b}\right)}{||\mathbf{w}||} = \frac{\mathbf{2}}{||\mathbf{w}||}$$

Our task become the maximizing of $\frac{\mathbf{2}}{||\boldsymbol{w}||}$,

or minimizing $\left||\boldsymbol{w}|\right|$, or $\mathbf{w}^{\mathrm{T}}\mathbf{w}$

Under $\boldsymbol{y}_i\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i - \boldsymbol{b}\right) \geq \mathbf{1}$ constraints.

Optimization task

$$\begin{cases} \frac{1}{2}\mathrm{w}^{\mathrm{T}}\mathrm{w} \to \min \\ y_i\left(\mathrm{w}^{\mathrm{T}}\mathrm{x}_i - b\right) \geq 1 \end{cases}$$

Karush-Kuhn-Tucker conditions

Our (primal) optimization under constraints problem:
$$\begin{cases} \min\limits_{\mathrm{z}} f(\mathrm{z}) \\ g_i(\mathrm{z}) \le 0 \\ h_i(\mathrm{z}) = 0 \end{cases}$$

If z^* – is a local minimum, then the exist Lagrangian multipliers α_i^* *and* β_j^* for:

$$\mathcal{L}(\mathrm{z}, \upalpha, \upbeta) = f(\mathrm{z}) + \sum_{i=1}^{m} \alpha_i g_i(\mathrm{z}) + \sum_{j=1}^{k} \beta_j h_j(\mathrm{z}),$$

such that:
$$\begin{cases} \frac{\partial}{\partial z_i} \mathcal{L}(\mathrm{z}^*, \upalpha^*, \upbeta^*) = 0, \\ \frac{\partial}{\partial \beta_i} \mathcal{L}(\mathrm{z}^*, \upalpha^*, \upbeta^*) = 0, \\ \alpha_i g_i(\mathrm{z}^*) = 0, \\ \alpha_i^* \ge 0 \end{cases}$$

And the solution of the primal problem is the solution of the dual problem: $\max\limits_{\alpha,\beta} \mathcal{L}(\mathrm{z}, \upalpha, \upbeta)$

Dual problem

$$\begin{cases} \frac{1}{2}\mathrm{w}^{\mathrm{T}}\mathrm{w} \to \min \\ y_i(\mathrm{w}^{\mathrm{T}}\mathrm{x}_i - b) \geq 1 \end{cases} \to \begin{cases} \frac{1}{2}\mathrm{w}^{\mathrm{T}}\mathrm{w} \to \min \\ -(y_i(\mathrm{w}^{\mathrm{T}}\mathrm{x}_i - b) - 1) \leq 0 \end{cases}$$

$$\downarrow$$

$$\mathcal{L}(\overbrace{\mathrm{w}, b}^{\mathrm{z}}, \alpha) = \frac{1}{2}(\mathrm{w}^{\mathrm{T}}\mathrm{w}) - \sum_{i=1}^{N} \alpha_i \left(y_i(\mathrm{w}^{\mathrm{T}}\mathrm{x}_i - b) - 1\right)$$

$$\alpha_i \geq 0; \qquad \alpha_i(y_i(\mathrm{w}^{\mathrm{T}}\mathrm{x}_i - b) - 1) = 0$$

Dual problem solution

$$\mathcal{L}(\mathrm{w}, \alpha, b) = \frac{1}{2}\left(\mathrm{w}^{\mathrm{T}}\mathrm{w}\right) - \sum_{i=1}^{N} \alpha_i \left(y_i\left(\mathrm{w}^{\mathrm{T}}\mathrm{x}_i - b\right) - 1\right); \qquad \alpha_i \geq 0; \qquad \alpha_i\left(y_i\left(\mathrm{w}^{\mathrm{T}}\mathrm{x}_i - b\right) - 1\right) = 0$$

$$\nabla_{\mathrm{w}}\mathcal{L}(\mathrm{w}, \alpha, b) = \mathrm{w} - \sum_{i=1}^{N} \alpha_i\, y_i \mathrm{x}_i = 0 \;\; \rightarrow \;\; \mathrm{w} = \sum_{i=1}^{N} \alpha_i\, y_i \mathrm{x}_i$$

$$\frac{\partial}{\partial b}\mathcal{L}(\mathrm{w}, \alpha, b) = \sum_{i=1}^{N} \alpha_i\, y_i = 0$$

$$\mathcal{L}(\mathrm{w}, \alpha, b) = \frac{1}{2}\left(\mathrm{w}^{\mathrm{T}}\mathrm{w}\right) - \sum_{i=1}^{N} \alpha_i \left(y_i\left(\mathrm{w}^{\mathrm{T}}\mathrm{x}_i - b\right) - 1\right) = \frac{1}{2}\sum\sum y_i y_j \alpha_i \alpha_j \mathrm{x}_i^{\mathrm{T}}\mathrm{x}_j - \sum\sum y_i y_j \alpha_i \alpha_j \mathrm{x}_i^{\mathrm{T}}\mathrm{x}_j + \sum \alpha_i =$$

$$= \mathcal{L}(\mathrm{w}, \alpha, b) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{N} y_i y_j\, \alpha_i \alpha_j \mathrm{x}_i^{\mathrm{T}}\mathrm{x}_j$$

Quadratic optimization problem under linear constraints is efficiently solved by quadratic programming.

Support vectors

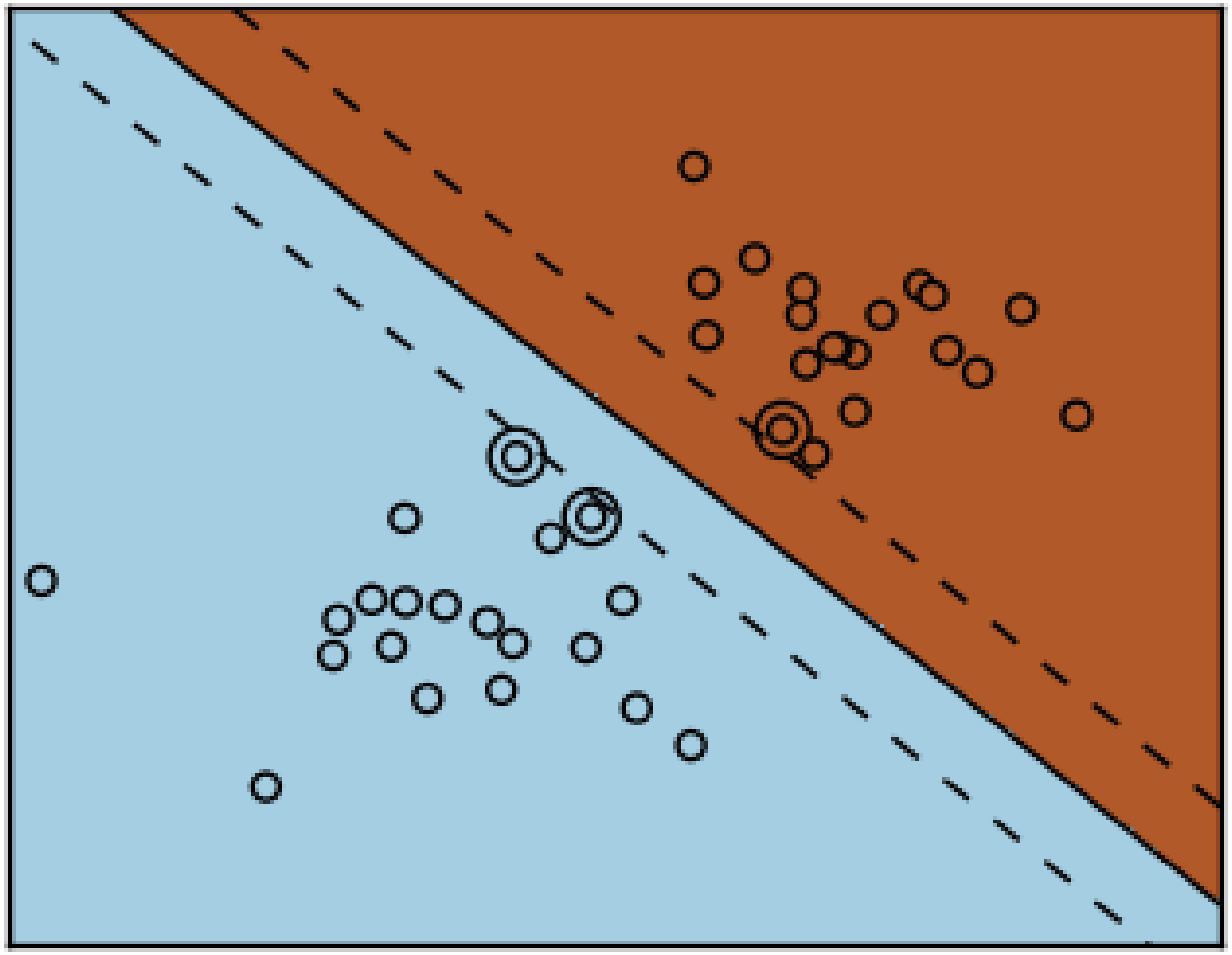
$$\mathrm{w} = \sum_{i=1}^{N} \alpha_i \, y_i \mathrm{x}_i$$

This is how we can find ***b***.

$$\alpha_i \left(y_i \left(\mathrm{w}^{\mathrm{T}} \mathrm{x}_i - b \right) - 1 \right) = 0$$

$$\mathrm{x}_i \colon \alpha_i > 0$$

$$y_i \left(\mathrm{w}^{\mathrm{T}} \mathrm{x}_i - b \right) - 1 = 0$$



Dual problem solution with CVXOPT Package

Task that QP solver solves:
$$\begin{cases} \frac{1}{2}\alpha^{\mathrm{T}} P\alpha + q^T \alpha \to \min \\ G\alpha \le h \\ A\alpha = b \end{cases}$$

$$\begin{cases} \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{N} y_i y_j\, \alpha_i \alpha_j \mathrm{x}_i^{\mathrm{T}} \mathrm{x}_j \to \max \\ \alpha_i \ge 0 \\ \sum_{i=1}^{N} \alpha_i\, y_i = 0 \end{cases} \quad \to$$

$$P = \left[y_i y_j \mathrm{x}_i^{\mathrm{T}} \mathrm{x}_j\right] \qquad q = [-1]$$

$$G = -I \qquad h = [0]$$

$$A = \mathrm{y}^T \qquad b = 0$$

Kernel trick

Notice, that we only used dot products in all calculations (i.e. $\mathrm{x}_i^{\mathrm{T}}\mathrm{x}_j$). That means that we do not need to transition into higher dimensional space but rather only define a dot product operation there.

$K(\mathrm{x},\mathrm{x}')$ is a kernel function if $K(\mathrm{x},\mathrm{x}') = \psi(\mathrm{x})^{\mathrm{T}}\psi(\mathrm{x}')$,
where $\psi\colon \mathrm{X} \to \mathrm{H}$, and H is some Hilbert space.

Example:

$$K(\mathrm{x},\mathrm{x}') = \left(1+\mathrm{x}^{\mathrm{T}}\mathrm{x}'\right)^2 = 1 + x_1^2 x_1'^2 + x_2^2 x_2'^2 + 2x_1 x'_1 + 2x_2 x'_2 + 2x_1 x'_1 x_2 x'_2$$

$$\psi(\mathrm{x}) = (1, x_1^2, x_2^2, \sqrt{2}x_1, \sqrt{2}x_2, \sqrt{2}x_1 x_2)$$

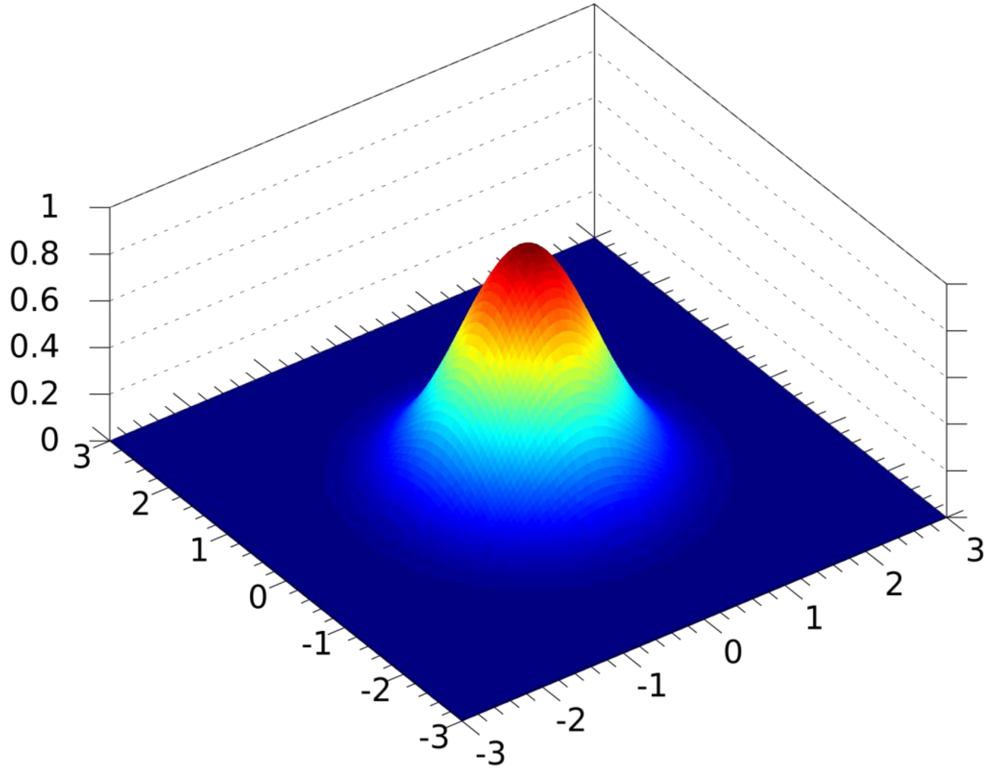
Kernels

Linear kernel: $\langle \mathrm{x}, \mathrm{x}' \rangle$

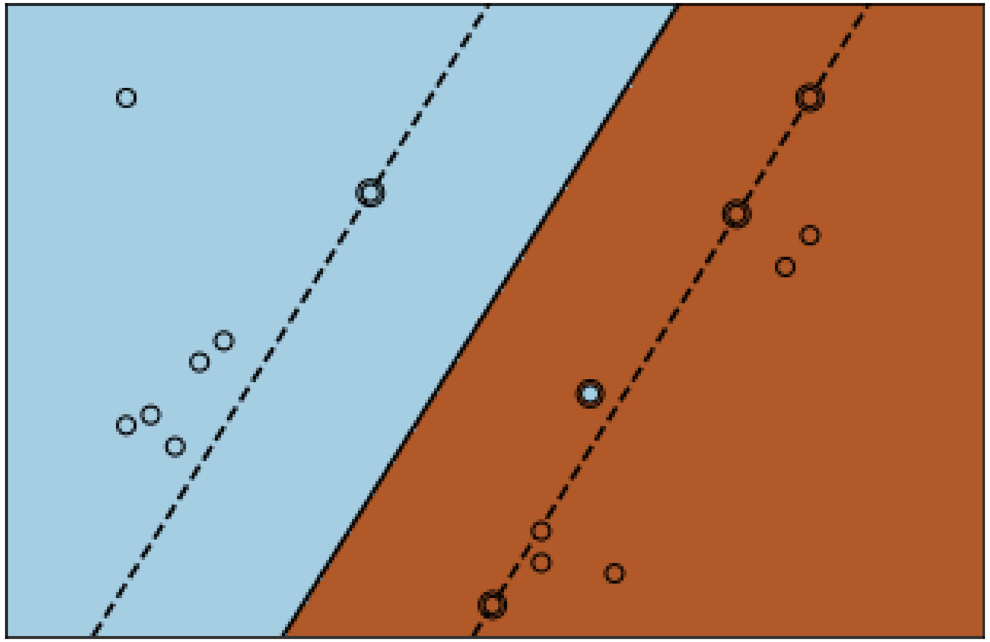
Polynomial kernel: $(r + \gamma \langle \mathrm{x}, \mathrm{x}' \rangle)^d$

Radial basis function: $e^{-\gamma |\mathrm{x} - \mathrm{x}'|^2}$

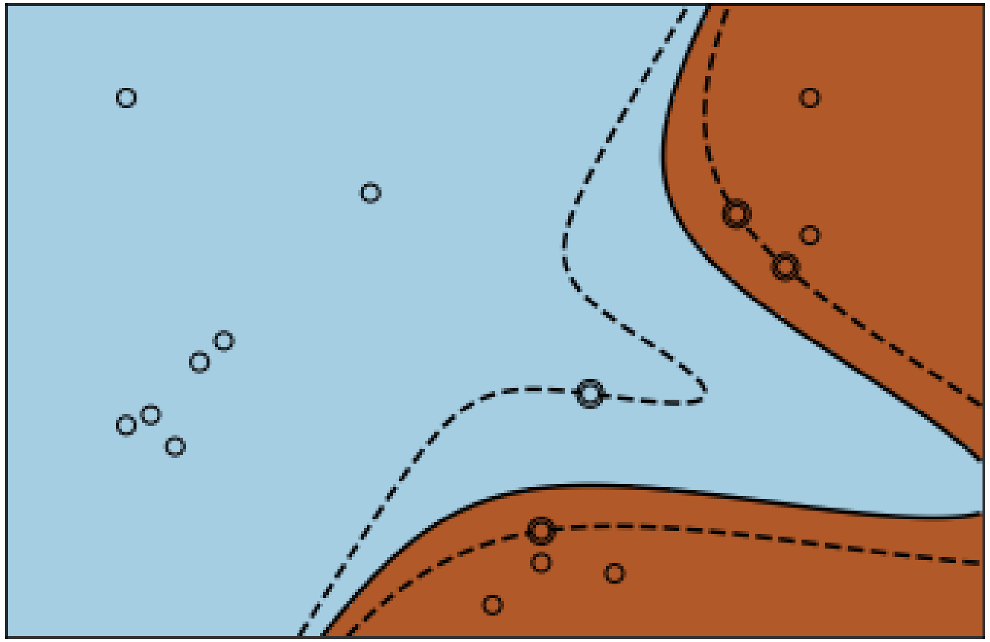
Sigmoid kernel: $\mathrm{th}(\gamma \langle \mathrm{x}, \mathrm{x}' \rangle + r)$



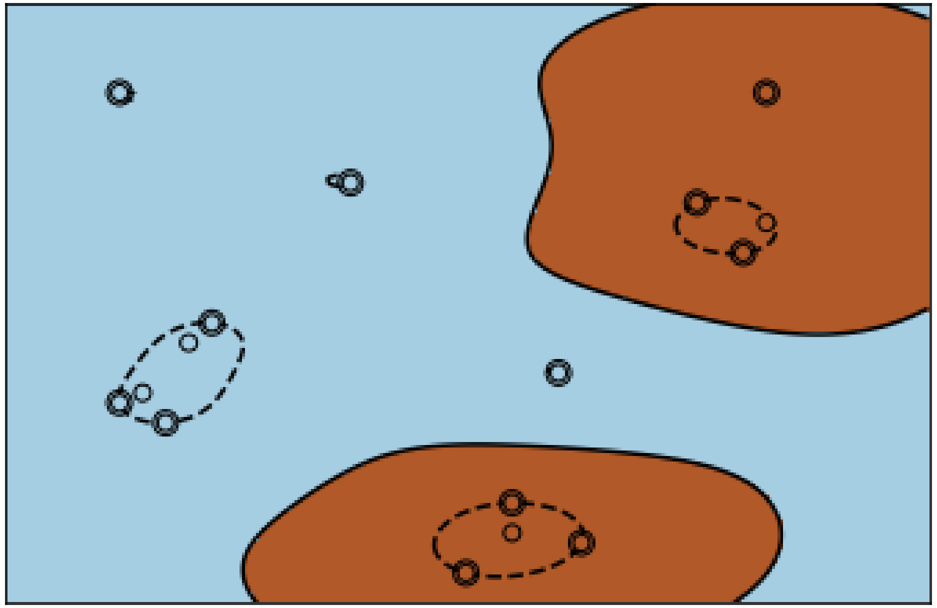
Radial Basis Function (RBF)



Linear kernel



Polynomial kernel



Radial Basis Function

Soft margin

Linearly separable case:

$$\begin{cases} \frac{1}{2}\mathrm{w}^{\mathrm{T}}\mathrm{w} \to \min \\ y_i\left(\mathrm{w}^{\mathrm{T}}\mathrm{x}_i - b\right) \geq 1 \end{cases}$$

Linearly inseparable case:

$$\begin{cases} \frac{1}{2}\mathrm{w}^{\mathrm{T}}\mathrm{w} + C\sum_{i=1}^{N}\xi_i \to \min \\ y_i\left(\mathrm{w}^{\mathrm{T}}\mathrm{x}_i - b\right) \geq 1 - \xi_i \\ \xi_i \geq 0 \end{cases}$$

Dual problem for soft margin

$$\mathcal{L}(\overbrace{\mathrm{w}, b, \xi}^{\mathrm{z}}, \underbrace{\alpha, r}_{\alpha}) = \frac{1}{2}\mathrm{w}^{\mathrm{T}}\mathrm{w} + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\alpha_i(y_i(\mathrm{w}^{\mathrm{T}}\mathrm{x}_i - b) - 1 + \xi_i) - \sum_{i=1}^{N} r_i\xi_i$$

$$= \frac{1}{2}\mathrm{w}^{\mathrm{T}}\mathrm{w} - \sum \alpha_i(y_i(\mathrm{w}^{\mathrm{T}}\mathrm{x}_i - b) - 1) - \sum \xi_i(r_i + \alpha_i - C)$$

$$\alpha_i, r_i \geq 0; \qquad \alpha_i\,(y_i(\mathrm{w}^{\mathrm{T}}\mathrm{x}_i - b) - 1 + \xi_i) = 0; \qquad r_i\xi_i = 0$$

$$\frac{\partial\mathcal{L}}{\partial\mathrm{w}} = \frac{\partial\mathcal{L}}{\partial b} = \frac{\partial\mathcal{L}}{\partial\xi} = 0 \Rightarrow \begin{cases} \mathrm{w} = \sum \alpha_i\, \mathrm{y}_i\mathrm{x}_i \\ \sum \alpha_i y_i = 0 \\ \alpha_i = C - r_i \Rightarrow 0 \leq \alpha_i \leq C \end{cases}$$

Dual problem for soft margin

$$\begin{cases} \mathrm{w} = \sum \alpha_i \, \mathrm{y}_i \mathrm{x}_i \\ \sum \alpha_i y_i = 0 \\ \alpha_i = C - r_i \end{cases} \Rightarrow \quad \mathcal{L}(\alpha) = \frac{1}{2}\mathrm{w}^{\mathrm{T}}\mathrm{w} - \sum \alpha_i \big(y_i\big(\mathrm{w}^{\mathrm{T}}\mathrm{x}_i - b\big) - 1\big) - \sum \xi_i (r_i + \alpha_i - C) =$$

$$= \frac{1}{2}\sum\sum \mathrm{y}_i \mathrm{y}_j \alpha_i \alpha_j \mathrm{x}_i^{\mathrm{T}} \mathrm{x}_j - \sum\sum \mathrm{y}_i \mathrm{y}_j \alpha_i \alpha_j \mathrm{x}_i^{\mathrm{T}} \mathrm{x}_j + \sum \alpha_i$$

$$\begin{cases} \max\limits_{\alpha} \mathcal{L}(\alpha) = \sum \alpha_i - \frac{1}{2}\sum\sum \mathrm{y}_i \mathrm{y}_j \alpha_i \alpha_j \mathrm{x}_i^{\mathrm{T}} \mathrm{x}_j \\ 0 \le \alpha_i \le C \\ \sum \alpha_i y_i = 0 \end{cases}$$

Vector types

$$\mathrm{w} = \sum \alpha_i\, \mathrm{y}_i \mathrm{x}_i; \qquad y_i\left(\mathrm{w}^{\mathrm{T}}\mathrm{x}_i - b\right) \geq 1 - \xi_i$$

$$\alpha_i = C - r_i; \qquad \alpha_i\left(y_i\left(\mathrm{w}^{\mathrm{T}}\mathrm{x}_i - b\right) - 1 + \xi_i\right) = 0; \qquad r_i \xi_i = 0$$

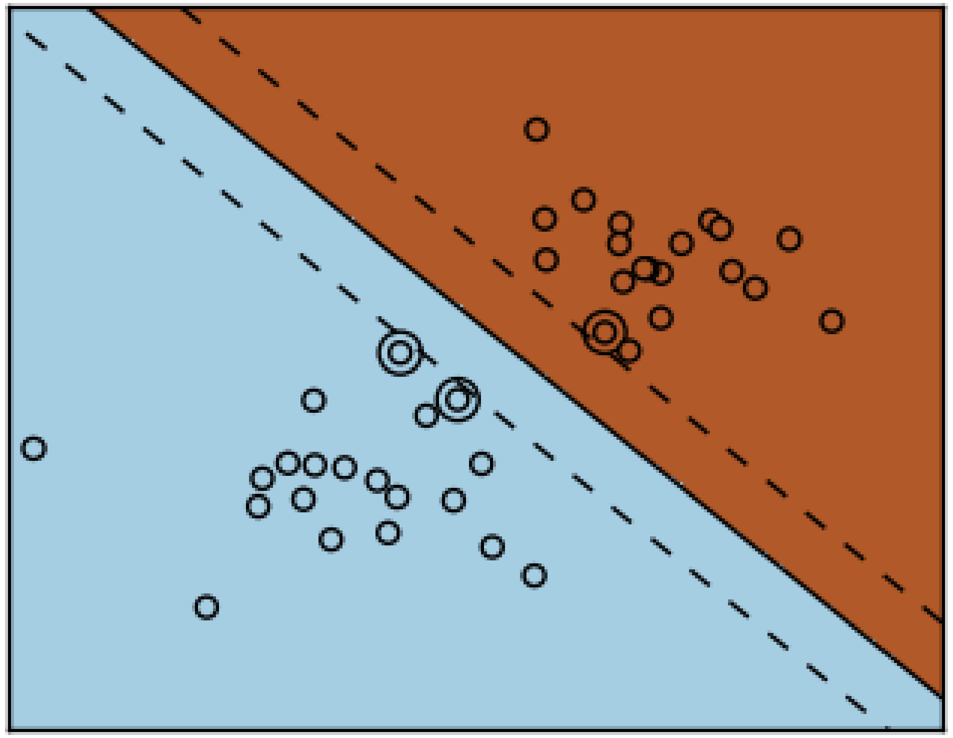
1. $\alpha_i = 0;\ \xi_i = 0;\ y_i\left(\mathrm{w}^{\mathrm{T}}\mathrm{x}_i - b\right) \geq 1$ – Inside Vectors

2. $0 < \alpha_i < C;\ \xi_i = 0;\ y_i\left(\mathrm{w}^{\mathrm{T}}\mathrm{x}_i - b\right) = 1$ – **"Good" support vectors**

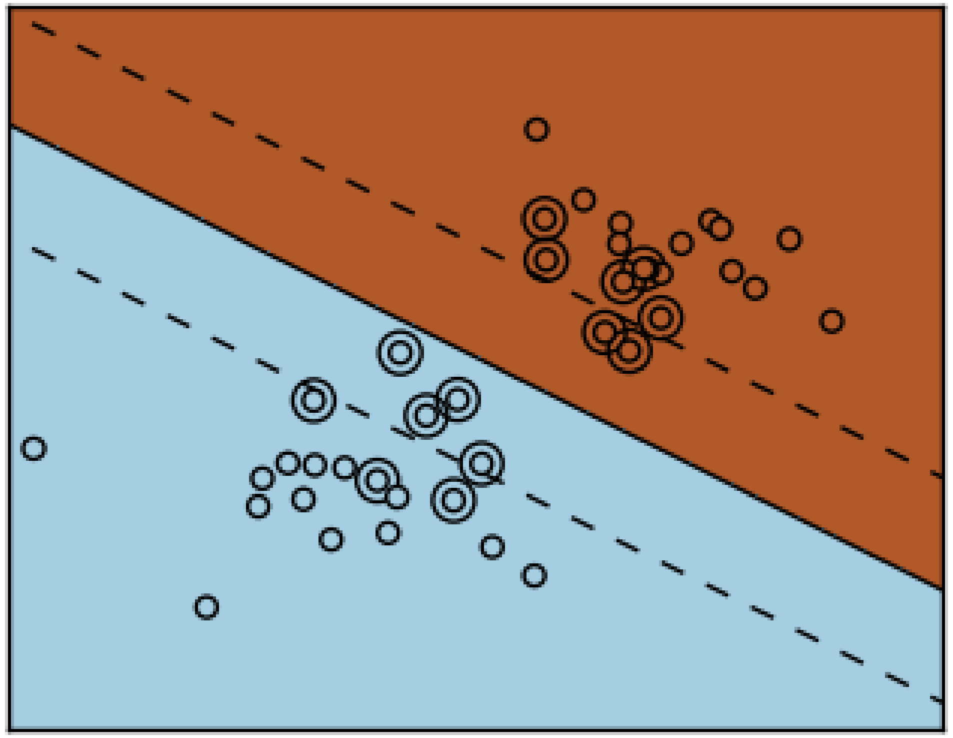
3. $\alpha_i = C;\ \xi_i > 0;\ y_i\left(\mathrm{w}^{\mathrm{T}}\mathrm{x}_i - b\right) \leq 1$ – *"Bad" support vectors*

Soft margin coefficient

$$\min\left(\frac{1}{2}\mathrm{w}^{\mathrm{T}}\mathrm{w} + C\sum_{i=1}^{N}\xi\right)$$

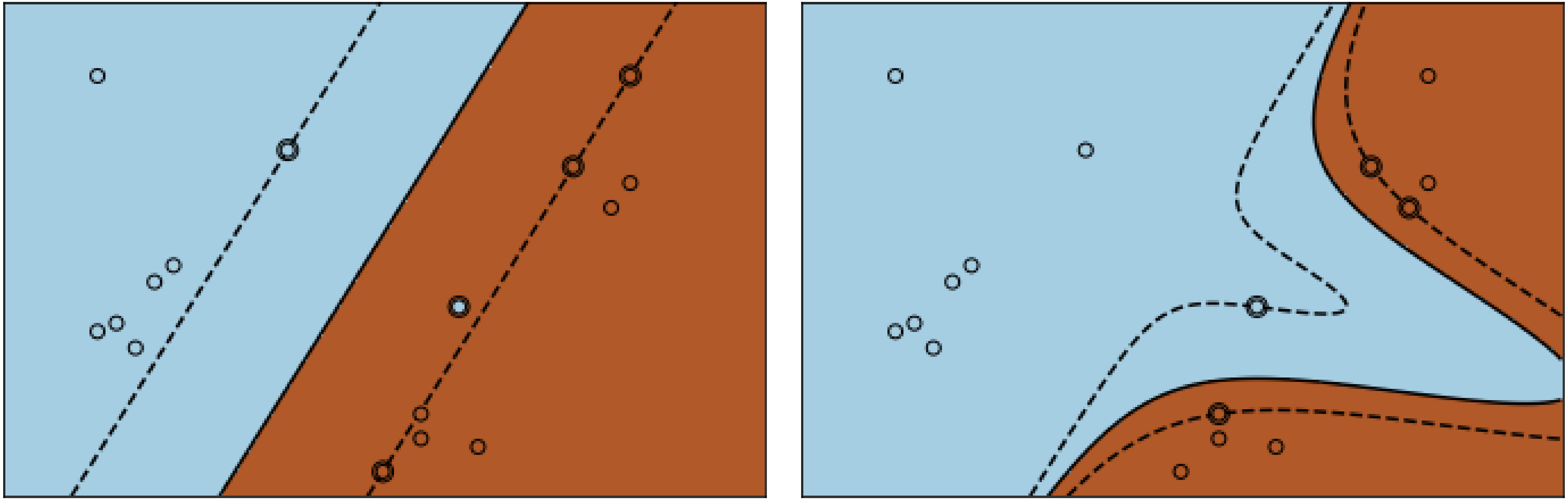


$C \uparrow$



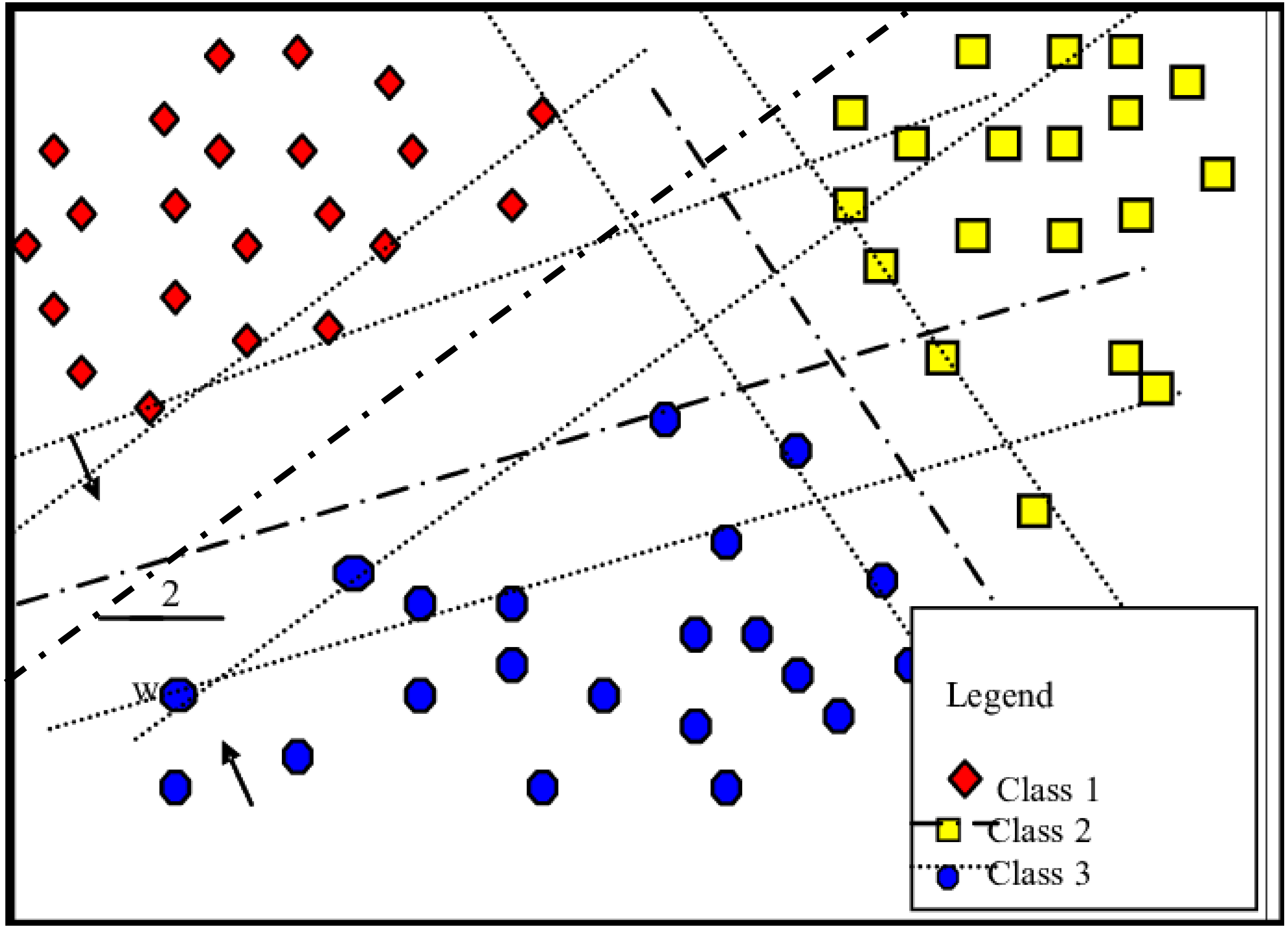
$C \downarrow$

Higher dimensionality and generalization



$$E_{gen} \leq \frac{\#\ of\ support\ vectors}{\#\ of\ training\ vectors}$$

Multi-class

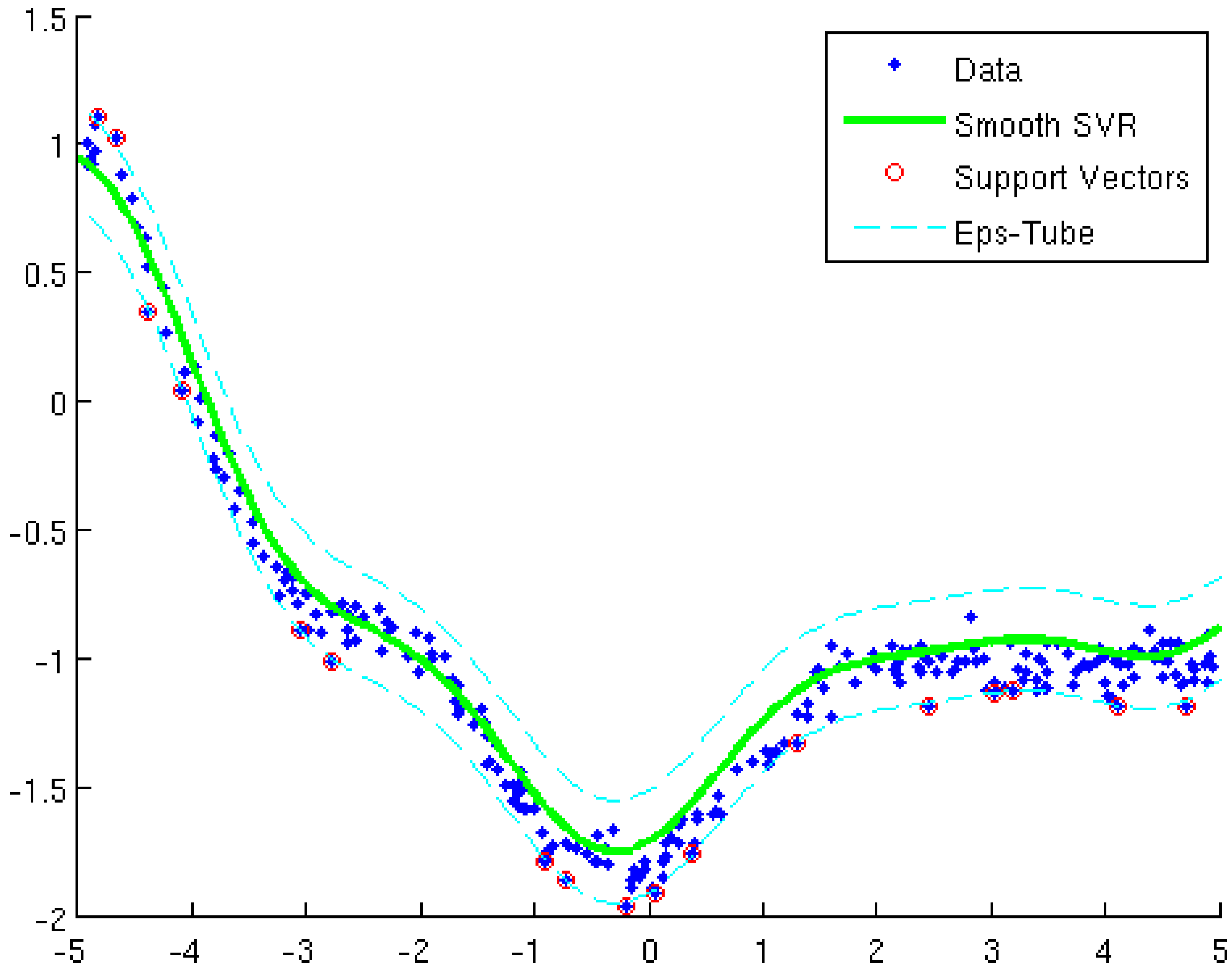


1) One-vs-all classifiers

2) $\max_y \left(\mathrm{w}_y^{\mathrm{T}} \mathrm{x} - b_y \right)$

Support Vector Regression Machine

$$\begin{cases} \frac{1}{2}\|\mathrm{w}\|^2 + C\sum(\xi_i + \xi_i^*) \to min \\ y_i - \mathrm{w}^{\mathrm{T}}\mathrm{x}_i - b \le \epsilon + \xi_i \\ \mathrm{w}^{\mathrm{T}}\mathrm{x}_i + b - y_i \le \epsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0 \end{cases}$$



Support Vector Networks

Cortes and Vapnik, 1995