

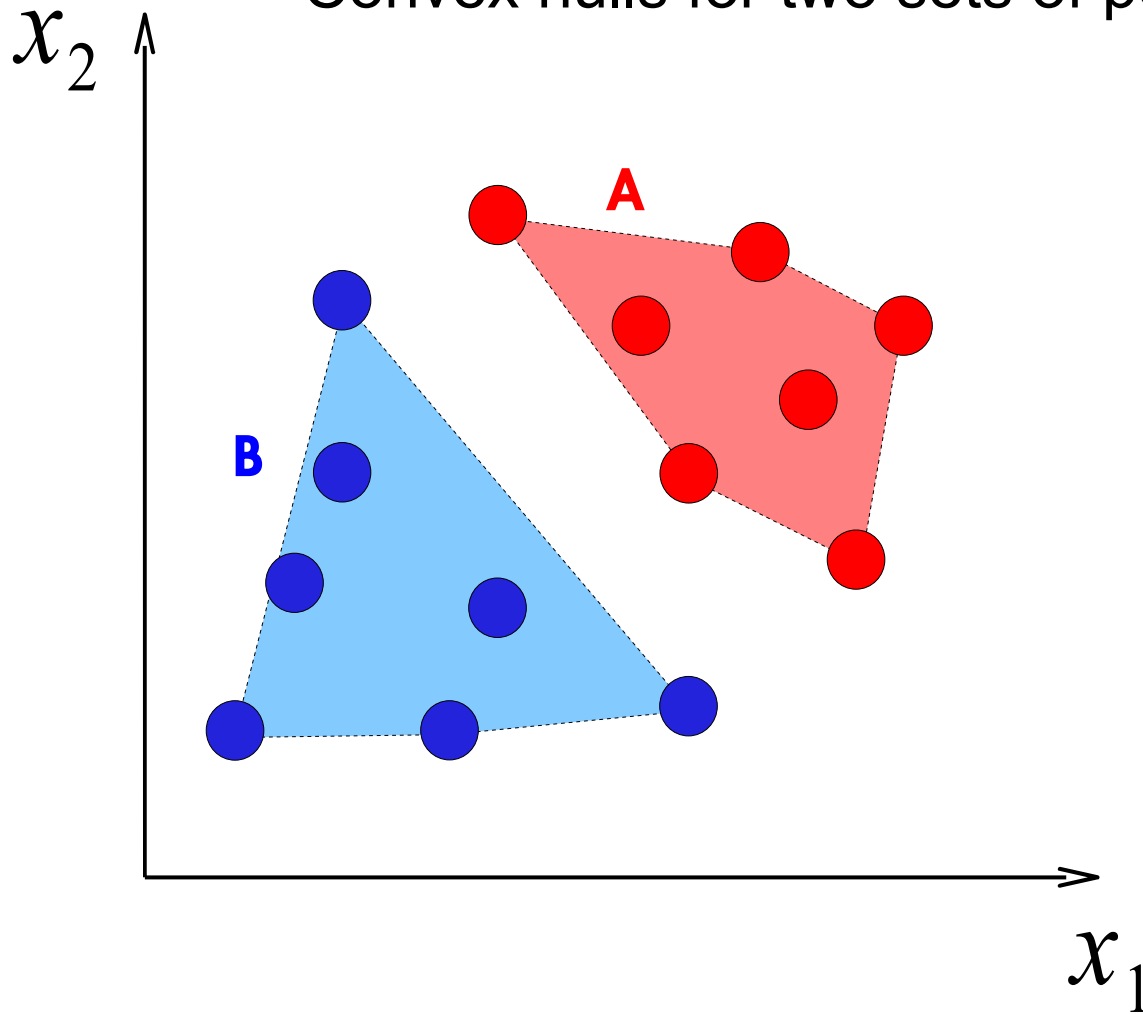
Geometrical intuition behind the dual problem

Based on:

KP Bennett, EJ Bredensteiner, “Duality and Geometry in SVM Classifiers”, *Proceedings of the International Conference on Machine Learning*, 2000

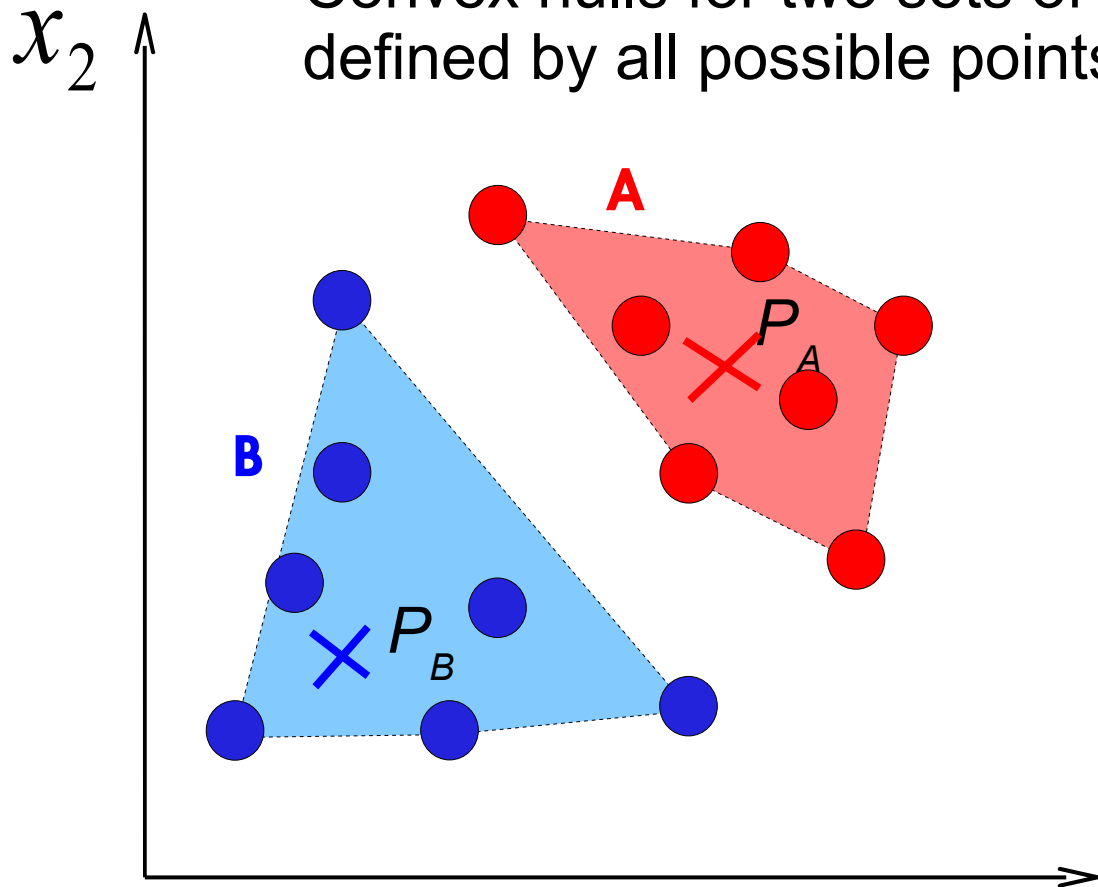
Geometrical intuition behind the dual problem

- Convex hulls for two sets of points **A** and **B**



Geometrical intuition behind the dual problem

- Convex hulls for two sets of points **A** and **B** defined by all possible points P_A and P_B



$$P_A = \sum_{i \in A} \lambda_i x_i \quad P_B = \sum_{i \in B} \lambda_i x_i$$

$$\sum_{i \in A} \lambda_i = 1$$

$$\lambda_i \geq 0, \quad i \in A$$

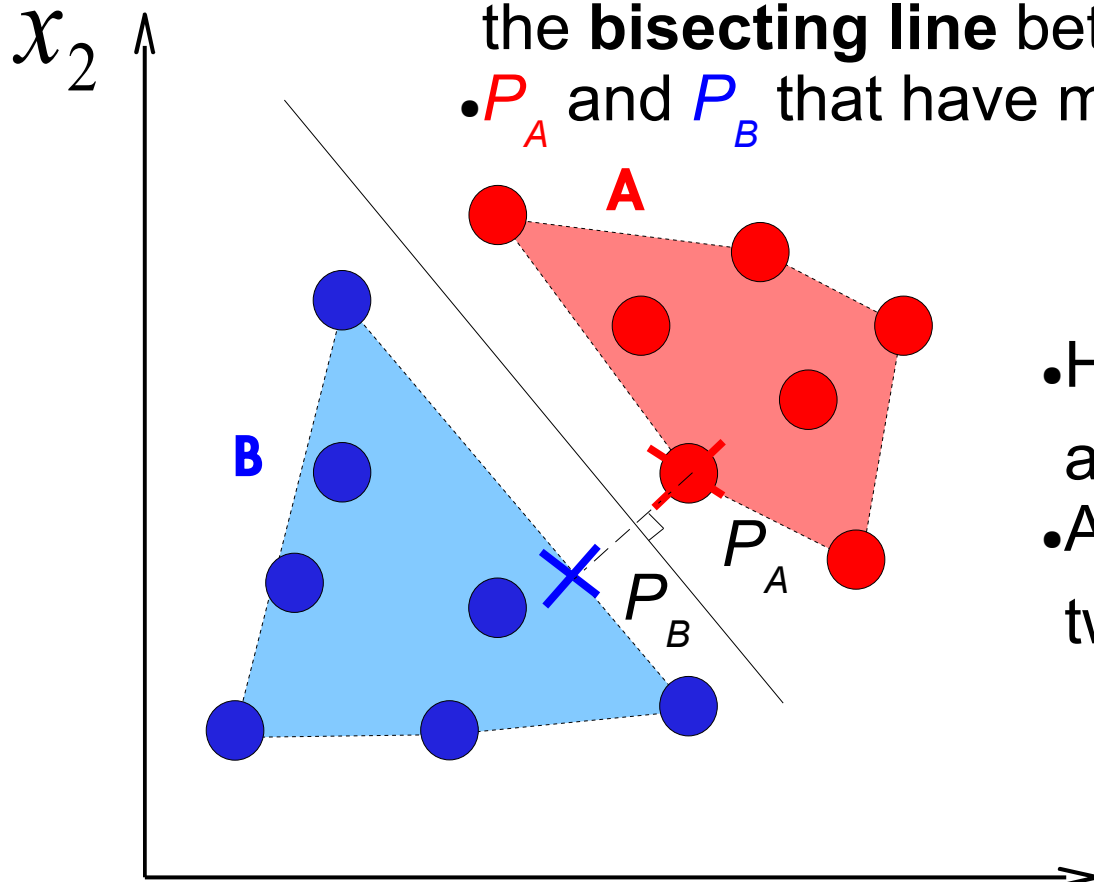
$$\sum_{i \in B} \lambda_i = 1$$

$$\lambda_i \geq 0, \quad i \in B$$

Geometrical intuition behind the dual problem

- For the plane separating **A** and **B** we choose the **bisecting line** between specific
- P_A and P_B that have minimal distance

$$\|P_A - P_B\|_2$$



- Here, for P_A , a single coeff. is non-zero
- And for P_B , two coeffs. are non-zero

$$P_A = \sum_{i \in A} \lambda_i x_i \quad P_B = \sum_{i \in B} \lambda_i x_i$$

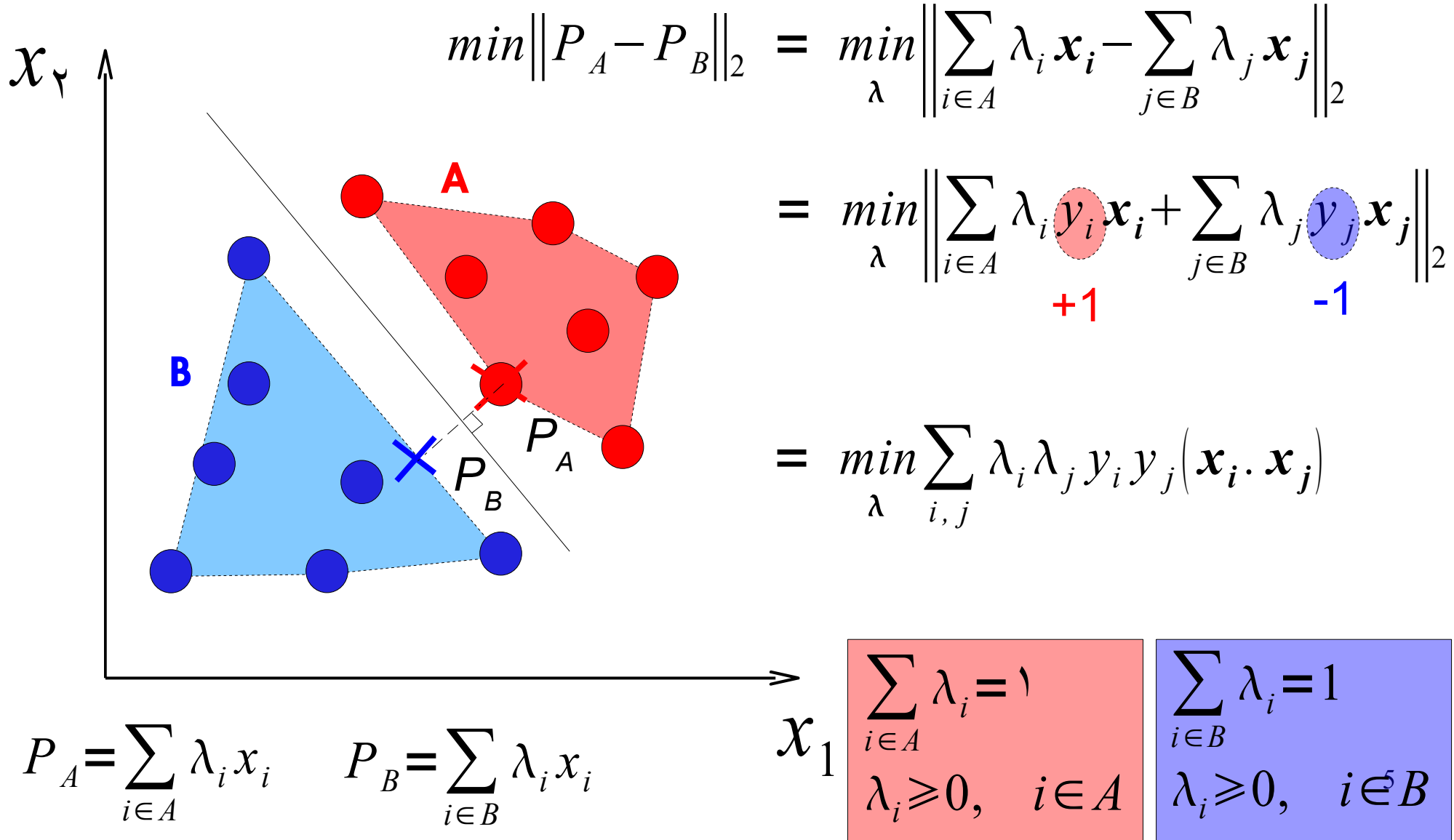
$$\sum_{i \in A} \lambda_i = 1$$

$$\lambda_i \geq 0, \quad i \in A$$

$$\sum_{i \in B} \lambda_i = 1$$

$$\lambda_i \geq 0, \quad i \in B$$

Geometrical intuition behind the dual problem



Going from the Primal to the Dual

- Constrained Optimization Problem $\min_w \frac{1}{2} \|\mathbf{w}\|^2$
s.t.:
 $y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1, \quad i=1, \dots, m$
label input

- A convex optimization problem (objective and constraints)
- Unique solution if datapoints are linearly separable

Going from the Primal to the Dual

- Constrained Optimization Problem $\min_w \frac{1}{2} \|\mathbf{w}\|^2$
s.t.:
 $y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1, \quad i=1, \dots, m$
label input

- Lagrange: $L(\mathbf{w}, b, \boldsymbol{\lambda}) = \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{i=1}^m \lambda_i [y_i(\mathbf{w} \cdot \mathbf{x}_i + b) - 1]$

Going from the Primal to the Dual

- **Constrained Optimization Problem**

$$\min_w \frac{1}{2} \|\mathbf{w}\|^2$$

s.t.:

$$y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1, \quad i=1, \dots, m$$

label input

- Lagrange:

$$L(\mathbf{w}, b, \boldsymbol{\lambda}) = \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{i=1}^m \lambda_i [y_i(\mathbf{w} \cdot \mathbf{x}_i + b) - 1]$$

- KKT

$$\nabla_{\mathbf{w}} L(\mathbf{w}, b, \boldsymbol{\lambda}) = \mathbf{w} - \sum_{i=1}^m \lambda_i y_i \mathbf{x}_i = 0$$

conditions:

$$\nabla_b L(\mathbf{w}, b, \boldsymbol{\lambda}) = - \sum_{i=1}^m \lambda_i y_i = 0$$

Going from the Primal to the Dual

- Constrained Optimization Problem**

$$\min_w \frac{1}{2} \|\mathbf{w}\|^2$$

s.t.:

$$y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1, \quad i=1, \dots, m$$

label input

- Lagrange:

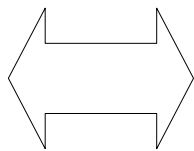
$$L(\mathbf{w}, b, \lambda) = \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{i=1}^m \lambda_i [y_i(\mathbf{w} \cdot \mathbf{x}_i + b) - 1]$$

- KKT

$$\nabla_{\mathbf{w}} L(\mathbf{w}, b, \lambda) = \mathbf{w} - \sum_{i=1}^m \lambda_i y_i \mathbf{x}_i = \mathbf{0}$$

conditions:

$$\nabla_b L(\mathbf{w}, b, \lambda) = - \sum_{i=1}^m \lambda_i y_i = 0$$



$$\mathbf{w} = \sum_{i=1}^m \lambda_i y_i \mathbf{x}_i$$

$$\sum_{i=1}^m \lambda_i y_i = 0$$

- Plus KKT:

$$\lambda_i [y_i(\mathbf{w} \cdot \mathbf{x}_i + b) - 1] = 0, \quad i=1, \dots, m$$

Going from the Primal to the Dual

- Constrained Optimization Problem**

$$\min_w \frac{1}{2} \|\mathbf{w}\|^2$$

s.t.:

$$y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1, \quad i=1, \dots, m$$

label input

- Lagrange:

$$L(\mathbf{w}, b, \boldsymbol{\lambda}) = \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{i=1}^m \lambda_i [y_i(\mathbf{w} \cdot \mathbf{x}_i + b) - 1]$$

$$L(\mathbf{w}, b, \boldsymbol{\lambda}) = \frac{1}{2} \left\| \sum_{i=1}^m \lambda_i y_i \mathbf{x}_i \right\|^2 - \sum_{i,j=1}^m \lambda_i \lambda_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j) - \sum_{i=1}^m \lambda_i y_i b + \sum_{i=1}^m \lambda_i$$

$$\mathbf{w} = \sum_{i=1}^m \lambda_i y_i \mathbf{x}_i$$

$$\sum_{i=1}^m \lambda_i y_i = 0$$

$$L(\mathbf{w}, b, \boldsymbol{\lambda}) = -\frac{1}{2} \sum_{i,j=1}^m \lambda_i \lambda_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j) + \sum_{i=1}^m \lambda_i$$

Going from the Primal to the Dual

- **Constrained Optimization Problem** $\min_w \frac{1}{2} \|\mathbf{w}\|^2$
s.t.:
 $y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1, \quad i=1, \dots, m$
label input

- **Equivalent
Dual Problem:**

$$\max_{\lambda} \left\{ -\frac{1}{2} \sum_{i,j=1}^m \lambda_i \lambda_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j) + \sum_{i=1}^m \lambda_i \right\}$$

s.t.:

$$\lambda_i \geq 0, \quad i=1, \dots, m$$
$$\sum_{i=1}^m \lambda_i y_i = 0$$

Solution to the Dual Problem

- The Dual Problem below admits the following solution:

$$\text{sign}(h(\mathbf{x})) = \text{sign}\left(\sum_{i=1}^m \lambda_i y_i (\mathbf{x}_i \cdot \mathbf{x}) + b\right)$$

$$b = y_i - \sum_{j=1}^m \lambda_j y_j (\mathbf{x}_j \cdot \mathbf{x}_i), \quad i=1, \dots, m$$

- Equivalent Dual Problem:

$$\max_{\lambda} \left\{ -\frac{1}{2} \sum_{i,j=1}^m \lambda_i \lambda_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j) + \sum_{i=1}^m \lambda_i \right\}$$

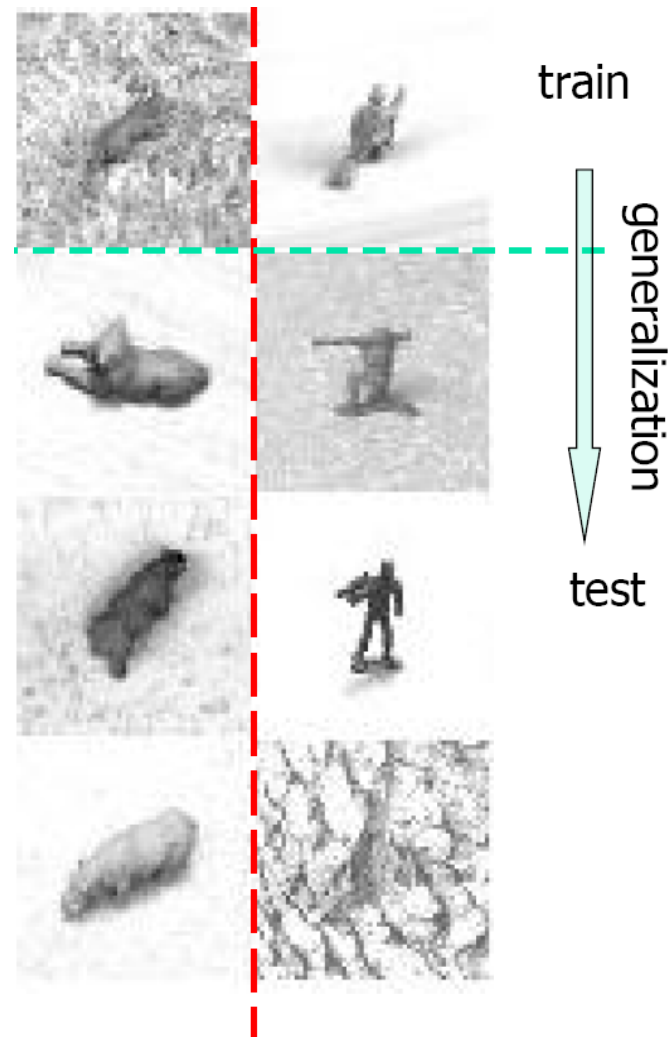
s.t.:

$$\lambda_i \geq 0, \quad i=1, \dots, m$$

$$\sum_{i=1}^m \lambda_i y_i = \cdot$$

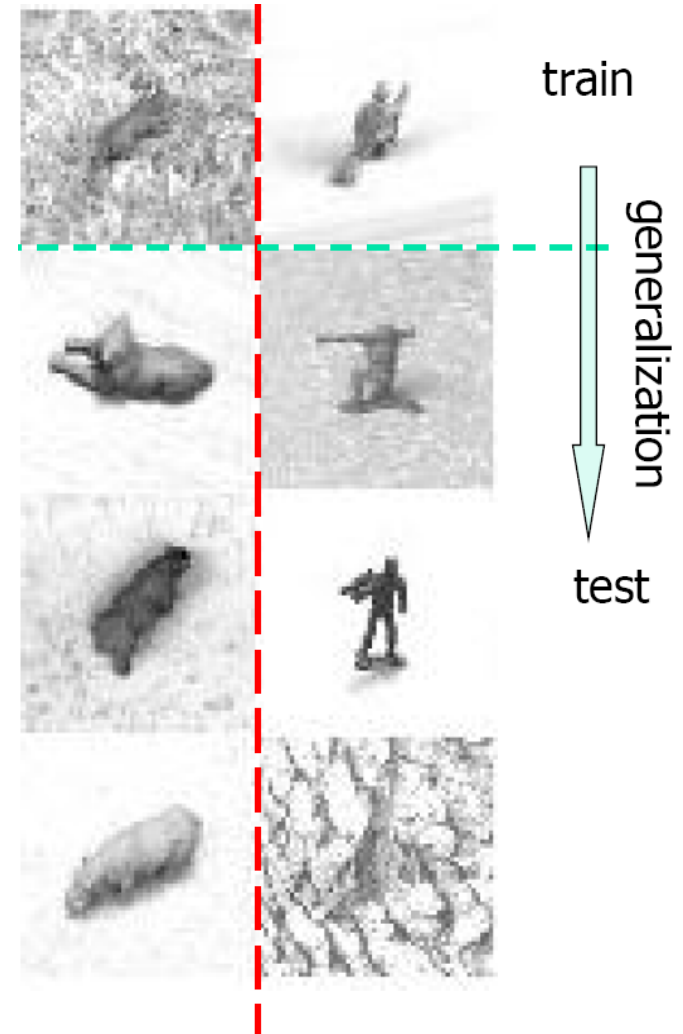
What are Support Vector Machines?

- Linear classifiers
- (Mostly) binary classifiers
- Supervised training
- Good generalization with explicit bounds




Main Ideas Behind Support Vector Machines

- Maximal margin
- Dual space
- Linear classifiers in high-dimensional space using non-linear mapping
- Kernel trick



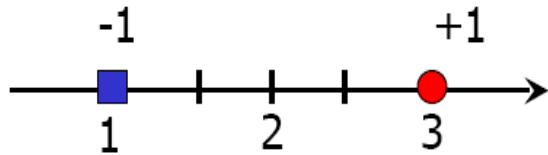
Quadratic Programming

$$\max_{w,b} \min_i \frac{|\mathbf{w}^T \mathbf{x}_i + b|}{\|\mathbf{w}\|}$$

$$\min_i |\mathbf{w}^T \mathbf{x}_i + b| = 1$$


$$\min_{w,b} \frac{1}{2} \langle \mathbf{w}^T \cdot \mathbf{w} \rangle$$

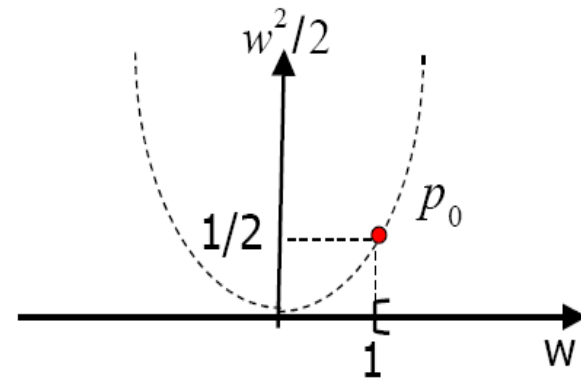
$$y_i (\langle \mathbf{w}^T \cdot \mathbf{x}_i \rangle + b) \geq 1$$



$$\min_w \frac{w^2}{2}$$

$$(+1)(w \cdot 3 + b) \geq 1$$

$$(-1)(w \cdot 1 + b) \geq 1$$



Using the Lagrangian

- Combine target and constraints
- Minimize over primal
- Maximize over dual

$$L(x, \boldsymbol{\lambda}) = f_0(x) - \sum \lambda_i f_i(x)$$

$$Q(\boldsymbol{\lambda}) = \min_x L(x, \boldsymbol{\lambda})$$

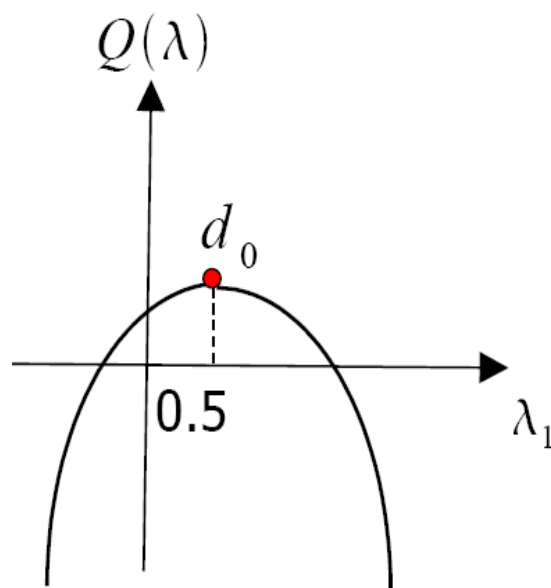
$$\max_{\boldsymbol{\lambda}} Q(\boldsymbol{\lambda}), \lambda > 0$$

Dual Space

$$\min_w \frac{w^2}{2}$$

$$(+1)(w \cdot 3 + b) \geq 1$$

$$(-1)(w \cdot 1 + b) \geq 1$$



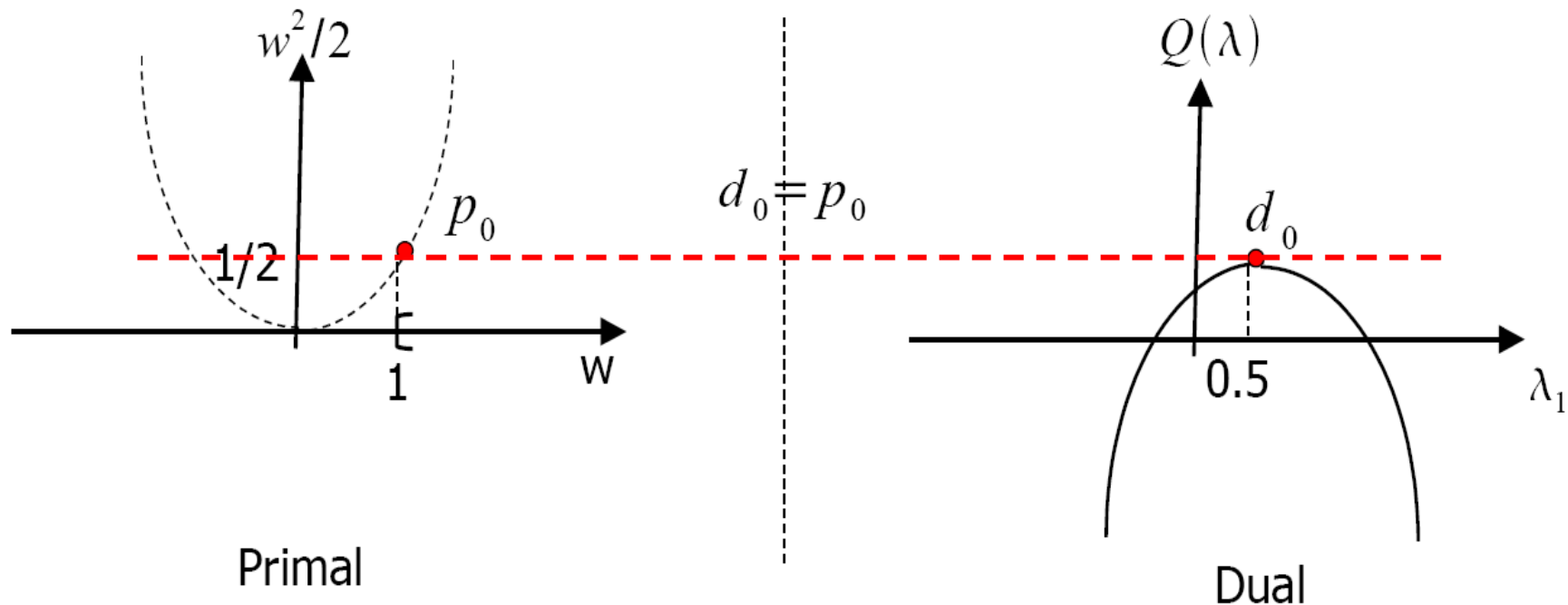
$$L(w, b, \lambda) = w^2/2 - \lambda_1(3w + b - 1) - \lambda_2(-w - b - 1)$$

$$\min_{w, b} L(w, b, \lambda) \Rightarrow \begin{cases} \lambda_1 = \lambda_2 \\ w = 3\lambda_1 - \lambda_2 = 2\lambda_1 \\ Q(\lambda) = Q(\lambda_1) = -2\lambda_1^2 + 2\lambda_1 \end{cases}$$

$$\max_{\lambda} Q(\lambda) \Rightarrow \lambda_1 = \lambda_2 = 1/2, w = 1, b = 2$$

Strong Duality

- Primal and dual space optimization:
 - Same result!



Dual Form

- H
 - Hessian matrix
 - Gram matrix
- Lambda
 - Support vector
 - Sparse

$$\max_{\lambda} Q(\lambda) = -0.5 \lambda^T H \lambda + f^T \lambda$$

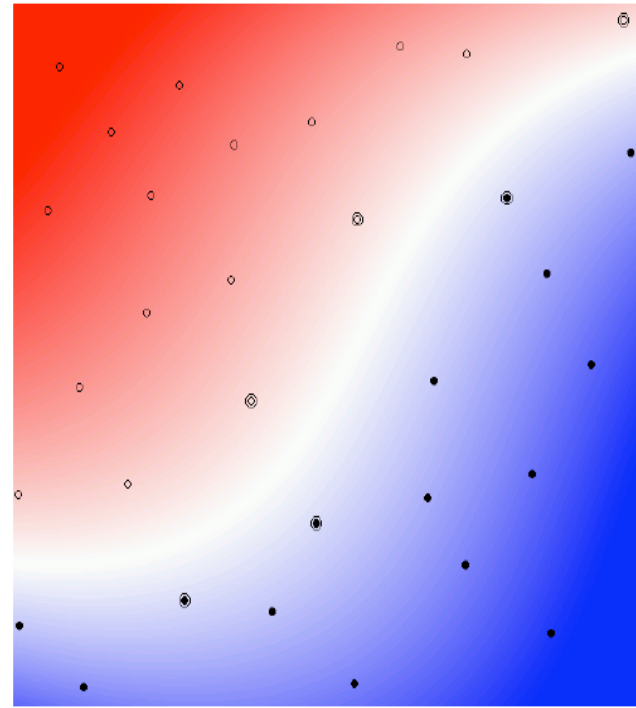
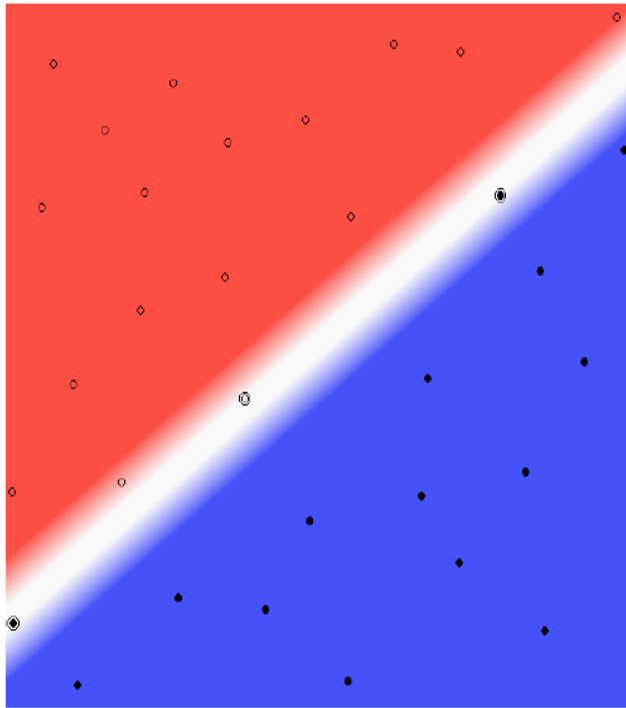
$$y^T \lambda = 0$$

$$\lambda \geq 0$$

$$\text{where, } H_{ij} = y_i y_j \langle \mathbf{x}_i^T \cdot \mathbf{x}_j \rangle$$

f is a unit vector

Non-linear separation of datasets

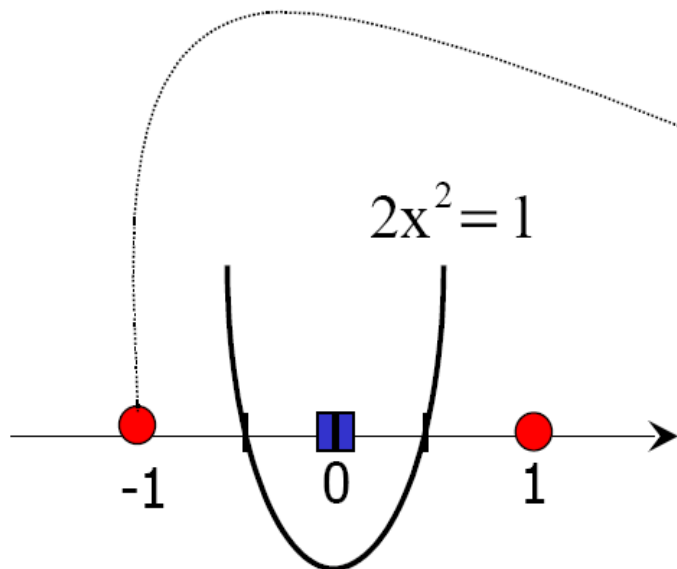


- **Non-linear separation is impossible in most problems**

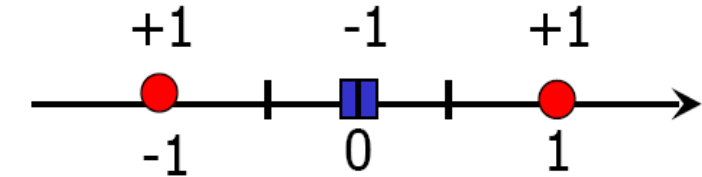
Non-separable datasets

- Solutions:

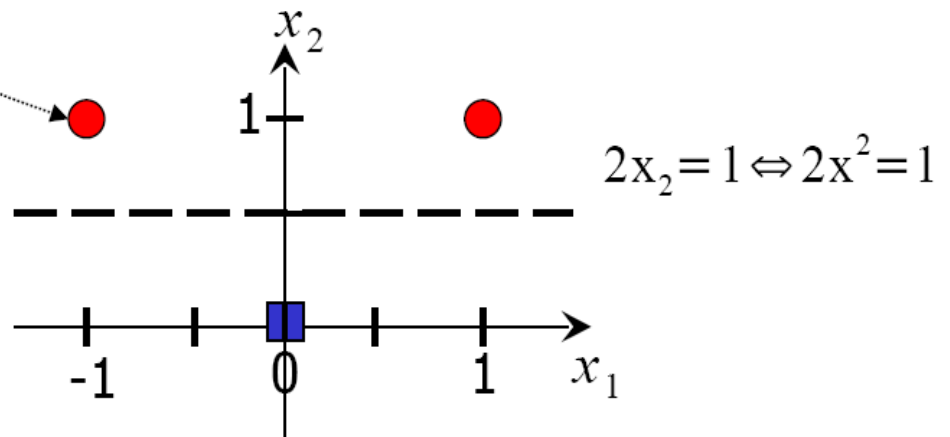
1) Nonlinear classifiers



2) Increase dimensionality of dataset and add a non-linear mapping Φ



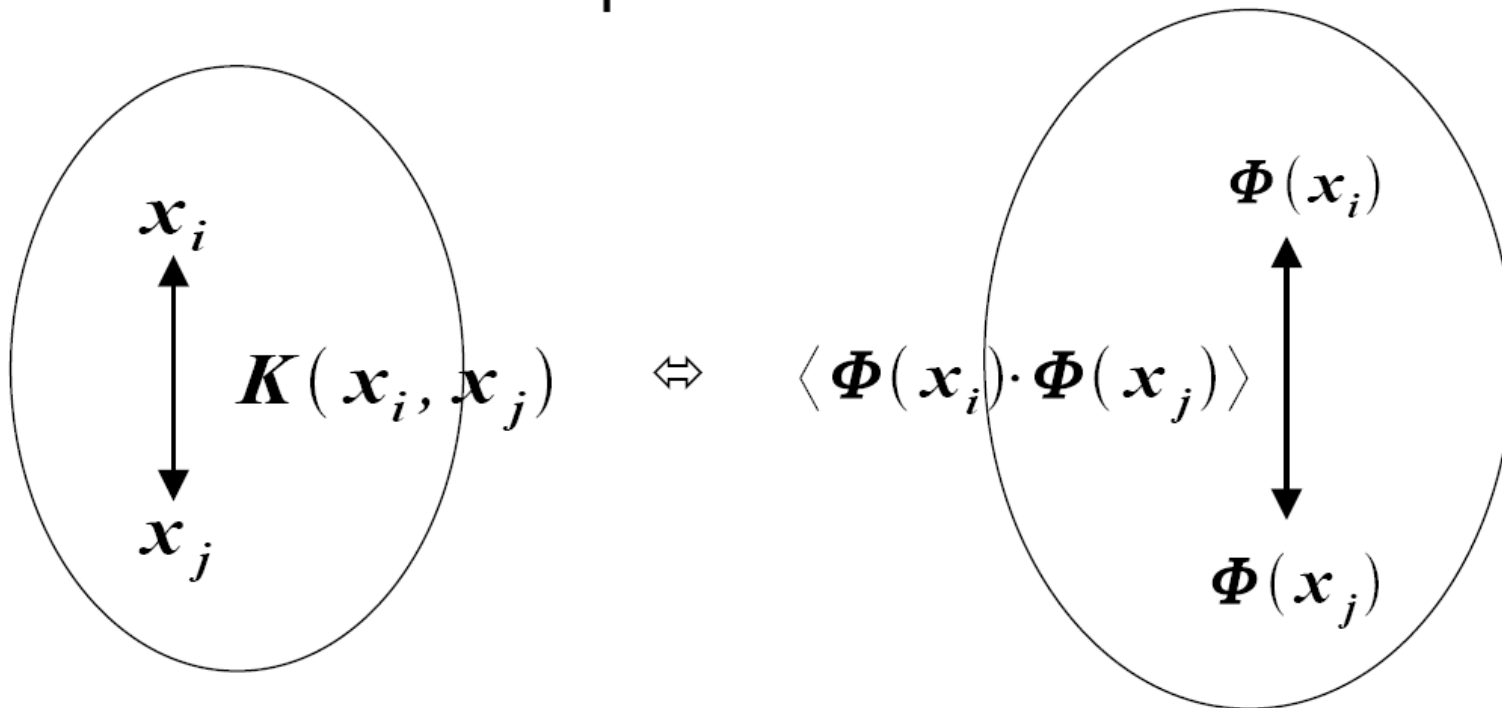
$$[x] \rightarrow \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} x \\ x^2 \end{bmatrix}$$



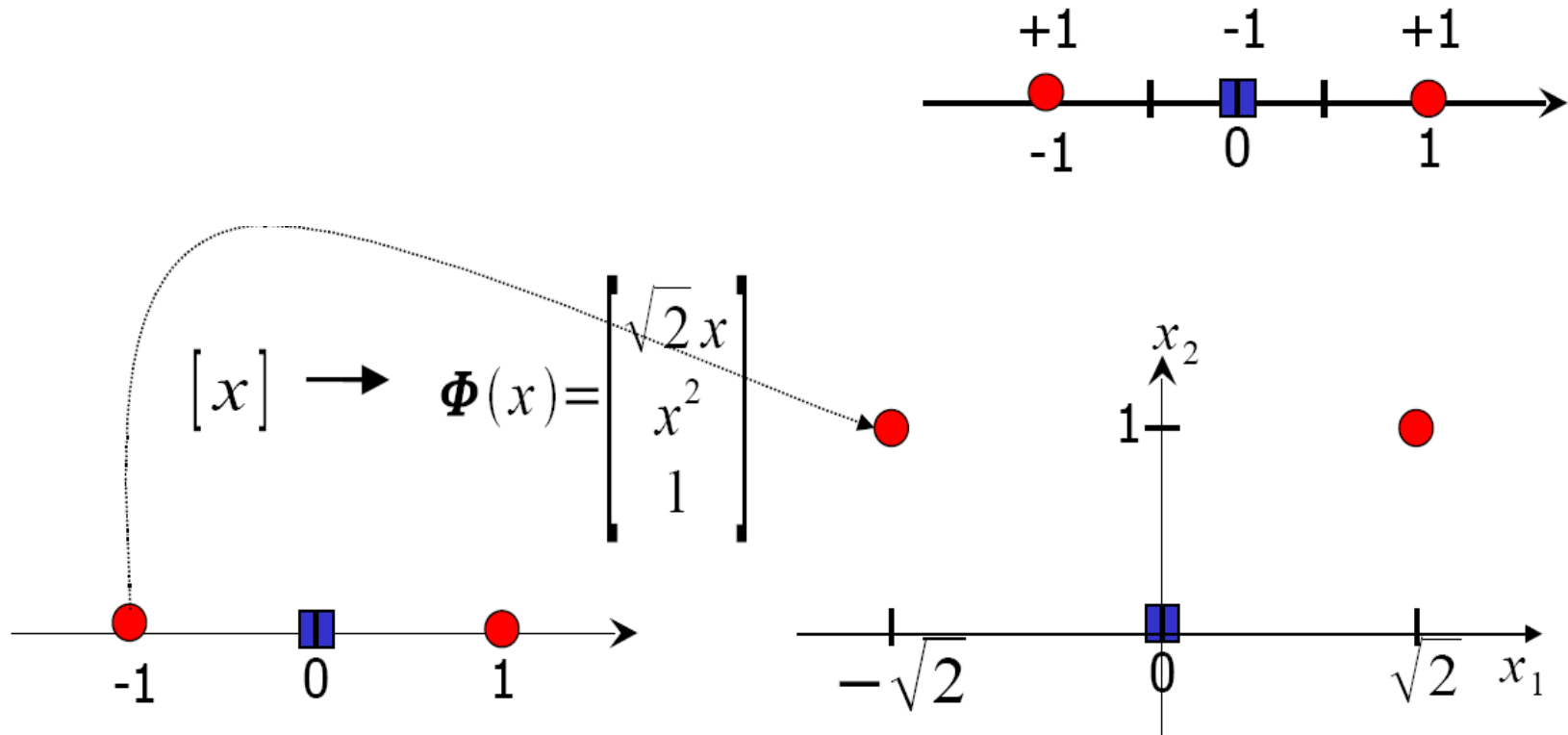
Kernel Trick

- Kernel function
 - in the original space
- Inner product
 - In the feature space with increased dimension

“similarity measure”
between 2 data samples



Kernel Trick Illustrated



$$K(x_i, x_j) = (x_i x_j + 1)^2$$

$$\langle \Phi(x_i) \cdot \Phi(x_j) \rangle = 2x_i x_j + x_i^2 x_j^2 + 1 = (x_i x_j + 1)^2 = K(x_i, x_j) \quad 23$$

Curse of Dimensionality Due to the Non-Linear Mapping

- Primal space

- Makes optimization much harder

$$\min_{w, b} \frac{1}{2} \langle \Phi^T(w) \cdot \Phi(w) \rangle$$
$$y_i (\langle \Phi^T(w) \cdot \Phi(x_i) \rangle + b) \geq 1$$

- Dual space

- Can be avoided

$$\max_{\lambda} Q(\lambda) = -0.5 \lambda^T H \lambda + f^T \lambda$$

$$y^T \lambda = 0$$

$$\lambda \geq 0$$

where, $H_{ij} = y_i y_j \mathbf{K}(x_i, x_j)$

f is a unit vector

Positive Semi-Definite (P.S.D.) Kernels (Mercer Condition)

- Dual form is convex

- H is P.S.D.
- Kernel must be P.S.D.

$$Q(\lambda) = -0.5 \lambda^T H \lambda + f^T \lambda$$

where, $H_{ij} = y_i y_j K(\mathbf{x}_i, \mathbf{x}_j)$

- Mercer kernels

- Polynomial
- Gaussian

$$K(\mathbf{x}, \mathbf{y}) = [\langle \mathbf{x}^T \mathbf{y} \rangle + 1]^p$$

$$K(\mathbf{x}, \mathbf{y}) = e^{-(\mathbf{x}-\mathbf{y})^T \Sigma^{-1} (\mathbf{x}-\mathbf{y})/2}$$

Advantages of SVM

- Work very well...
- Error bounds easy to obtain:
 - Generalization error **small** and **predictable**

$$E_{test} = E_{train} + E_{generalization} \quad \leftarrow \frac{|SV|}{N}$$

- Fool-proof method:
 - (Mostly) three kernels to choose from:
 - Gaussian
 - Linear and Polynomial
 - Sigmoid
 - Very small number of parameters to optimize

Limitations of SVM

→ Size limitation:

- Size of kernel matrix is quadratic with the number of training vectors

→ Speed limitations:

- 1) During **training**:
very large quadratic programming problem solved numerically
 - Solutions:
 - **Chunking**
 - **Sequential Minimal Optimization (SMO)**
breaks QP problem into many small QP problems solved analytically
 - **Hardware** implementations
- 2) During **testing**:
number of support vectors
 - Solution: **Online SVM**