CS570 Artificial Intelligence & Machine Learning

Support Vector Machines

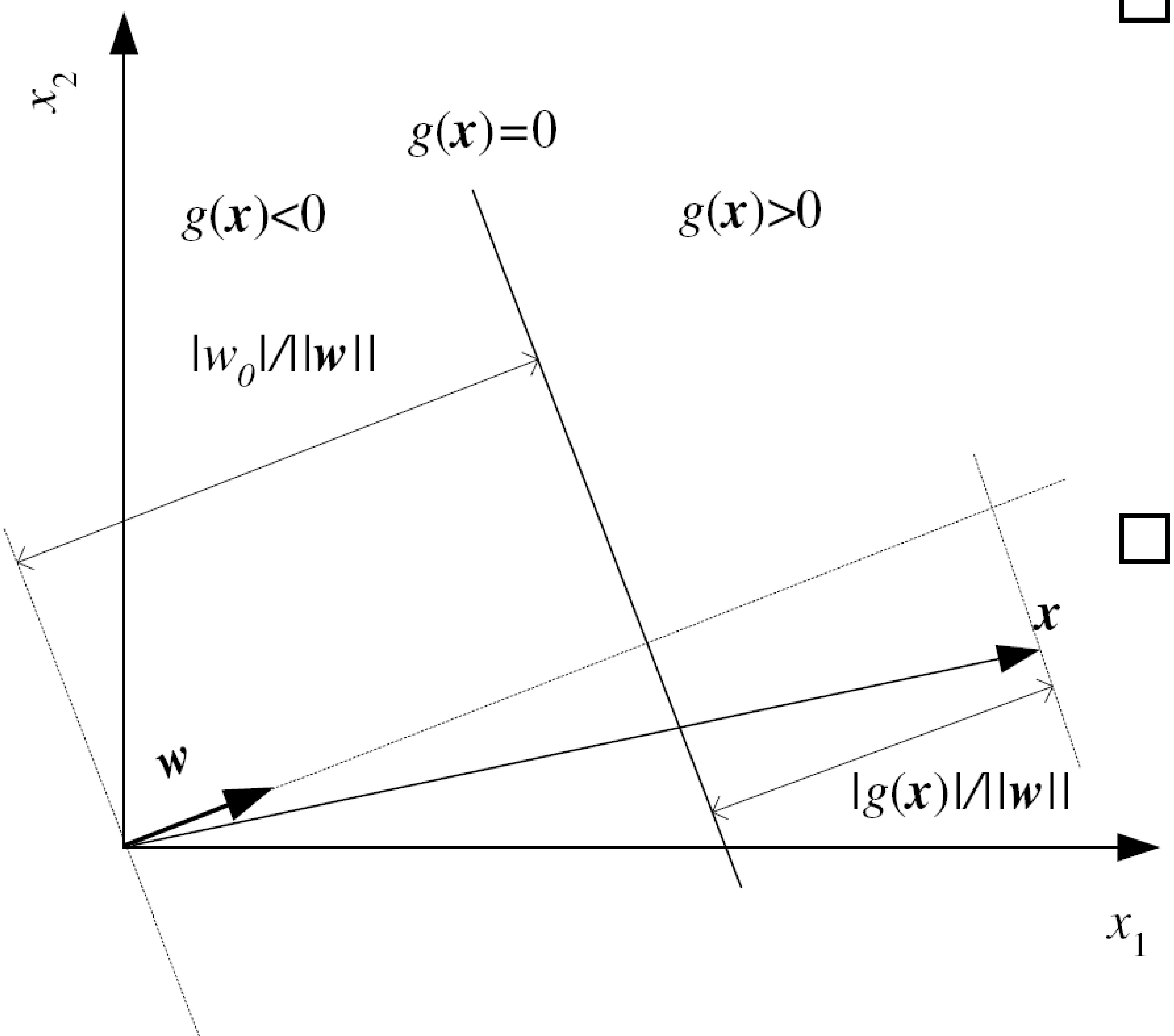
Kee-Eung Kim
Department of Computer Science
KAIST

Support Vector Machines

☐ Key idea: find the optimal separating hyperplane

- $\mathcal{X} = \{\mathbf{x}^t, r^t\}_t$ where $r^t = \begin{cases} +1 \text{ if } \mathbf{x}^t \in C_1 \\ -1 \text{ if } \mathbf{x}^t \in C_2 \end{cases}$
- Find **w** and w_0 such that
  $$\mathbf{w}^T\mathbf{x}^t + w_0 \geq +1 \text{ for } r^t = +1$$
  $$\mathbf{w}^T\mathbf{x}^t + w_0 \leq -1 \text{ for } r^t = -1$$
- Equivalently,
  $$r^t(\mathbf{w}^T\mathbf{x}^t + w_0) \geq +1$$

Geometric View



- ☐ Two points x_1 and x_2 on the decision surface:

  $$\mathbf{w}^T\mathbf{x}_1 + w_0 = \mathbf{w}^T\mathbf{x}_2 + w_0$$

  $$\mathbf{w}^T(\mathbf{x}_1 - \mathbf{x}_2) = 0$$

  - w is normal to any vector lying on the hyperplane

- ☐ Let $\mathbf{x} = \mathbf{x}_p + r\frac{\mathbf{w}}{||\mathbf{w}||}$
  - x_p is the normal projection of x onto the hyperplane
  - Since $g(x_p) = 0$, we have

    $$r = \frac{g(\mathbf{x})}{||\mathbf{w}||}$$

  - Position from the origin:

    $$r_0 = w_0/||\mathbf{w}||$$

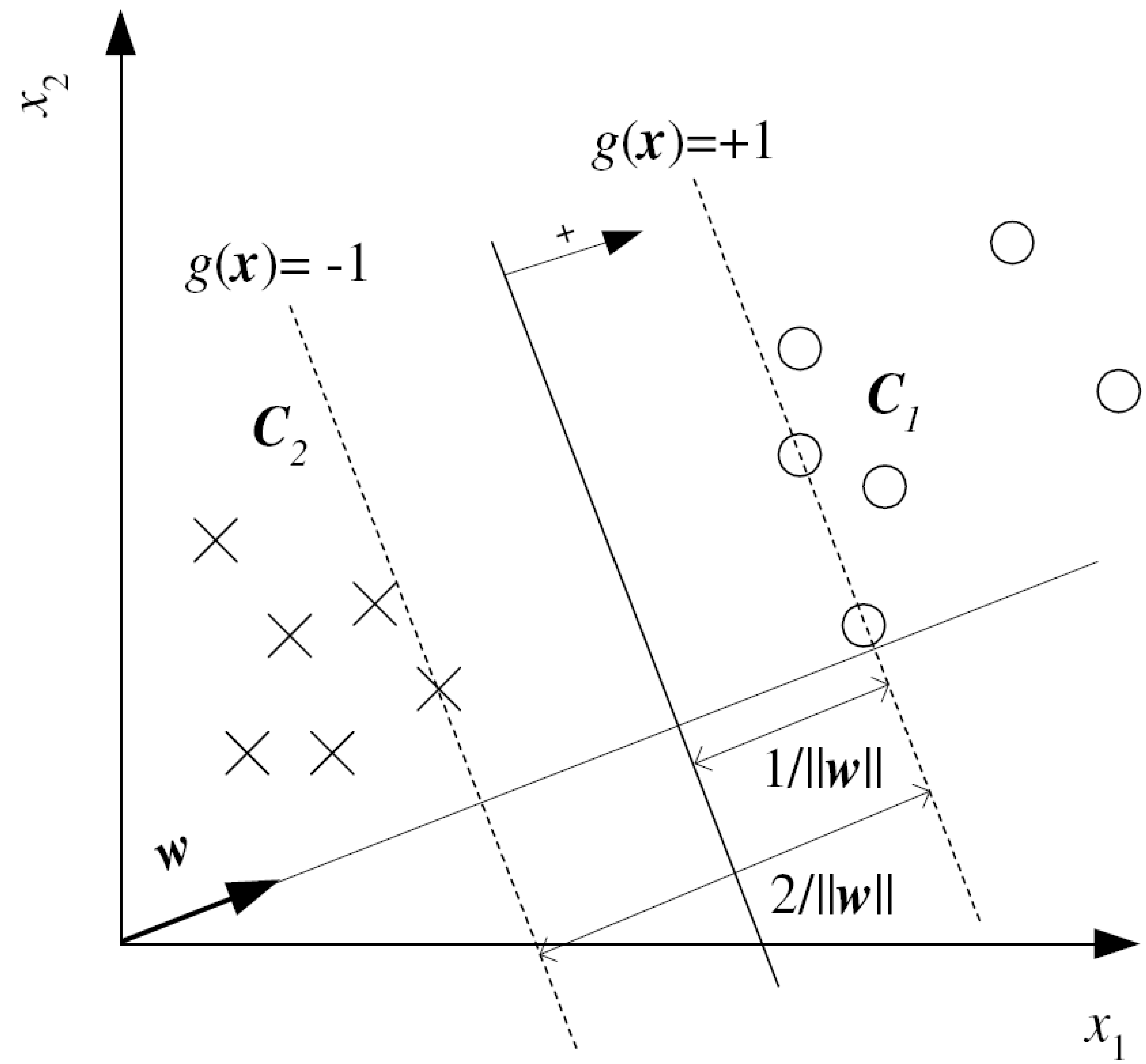
Margins

- Distance from the discriminant to the closest instance on either side

- Distance of **x** to the hyperplane:

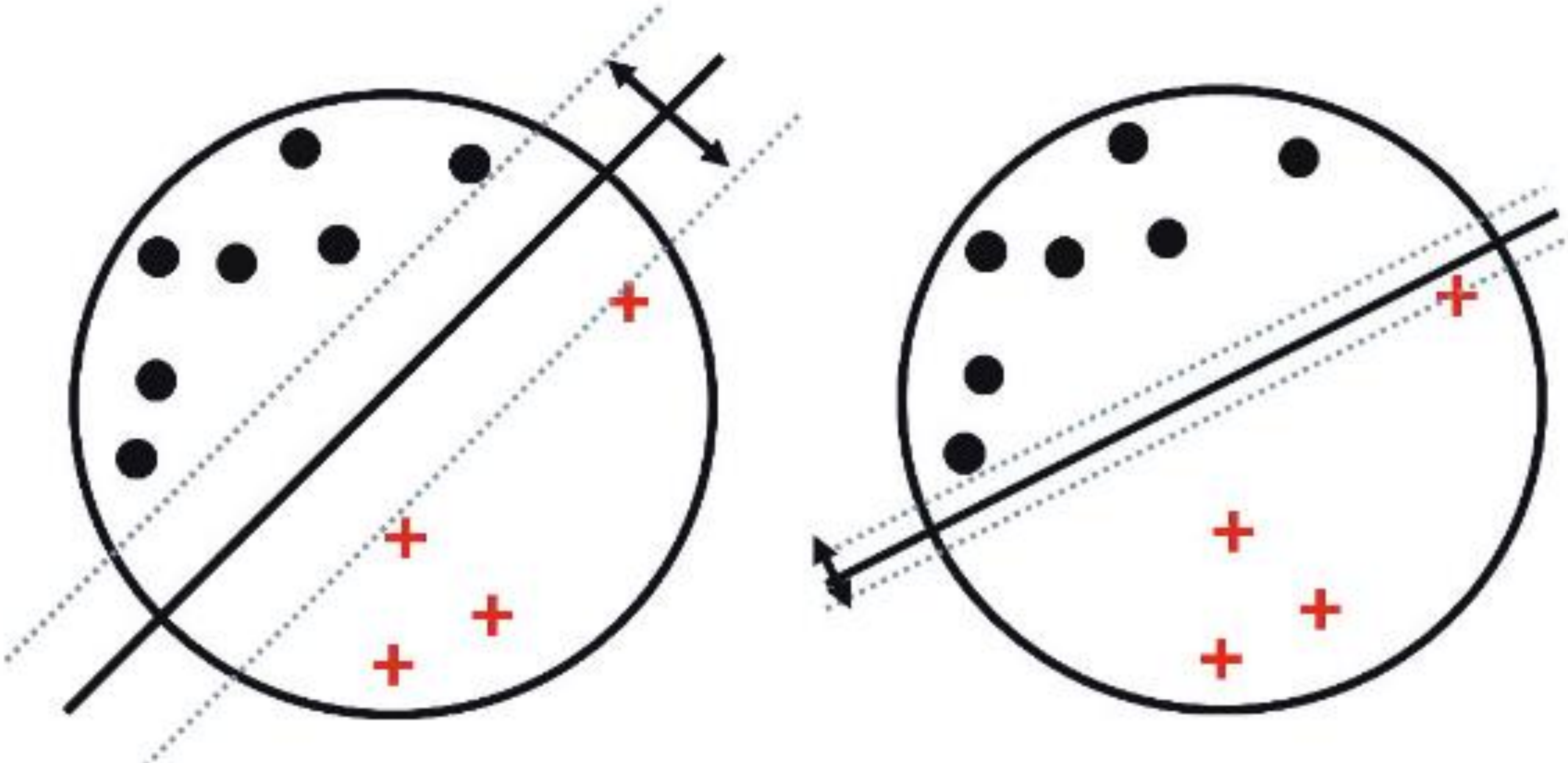
$$\frac{|\mathbf{w}^T\mathbf{x}^t + w_0|}{\|\mathbf{w}\|}$$

- Want: $\dfrac{r^t(\mathbf{w}^T\mathbf{x}^t + w_0)}{\|\mathbf{w}\|} \geq \rho, \forall t$



- For a unique solution, fix $\rho\|\mathbf{w}\| = 1$ and thus to maximize margin, minimize $\|\mathbf{w}\|$: $\min \frac{1}{2}\|\mathbf{w}\|^2$ subject to $r^t(\mathbf{w}^T\mathbf{x}^t + w_0) \geq +1, \forall t$
  - Quadratic programming problem!

Margins



Maximizing Margins

- $\min \frac{1}{2}\|\mathbf{w}\|^2$ subject to $r^t(\mathbf{w}^T\mathbf{x}^t + w_0) \geq +1, \forall t$
- $$L_p = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{t=1}^{N} \alpha^t[r^t(\mathbf{w}^T\mathbf{x}^t + w_0) - 1]$$
  $$= \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{t=1}^{N} \alpha^t r^t(\mathbf{w}^T\mathbf{x}^t + w_0) + \sum_{t=1}^{N} \alpha^t$$
- $\frac{\partial L_p}{\partial \mathbf{w}} = 0 \Rightarrow \mathbf{w} = \sum_{t=1}^{N} \alpha^t r^t \mathbf{x}^t \quad \frac{\partial L_p}{\partial w_0} = 0 \Rightarrow \sum_{t=1}^{N} \alpha^t r^t = 0$
- $$L_d = \tfrac{1}{2}\mathbf{w}^T\mathbf{w} - \mathbf{w}^T \sum_t \alpha^t r^t \mathbf{x}^t - w_0 \sum_t \alpha^t r^t + \sum_t \alpha^t$$
  $$= -\tfrac{1}{2}\mathbf{w}^T\mathbf{w} + \sum_t \alpha^t$$
  $$= -\tfrac{1}{2}\sum_t \sum_s \alpha^t \alpha^s r^t r^s (\mathbf{x}^t)^T \mathbf{x}^s + \sum_t \alpha^t$$
  subject to $\sum_t \alpha^t r^t = 0$ and $\alpha^t \geq 0, \forall t$
  - Most $\alpha^t = 0$ and only small number have $\alpha^t > 0$; x^t with $\alpha^t > 0$ are the support vectors

KAIST
Korea Advanced Institute of Science and Technology
한국과학기술원

Soft Margins

- If not linearly separable

$$r^t(\mathbf{w}^T\mathbf{x}^t + w_0) \geq 1 - \xi^t$$

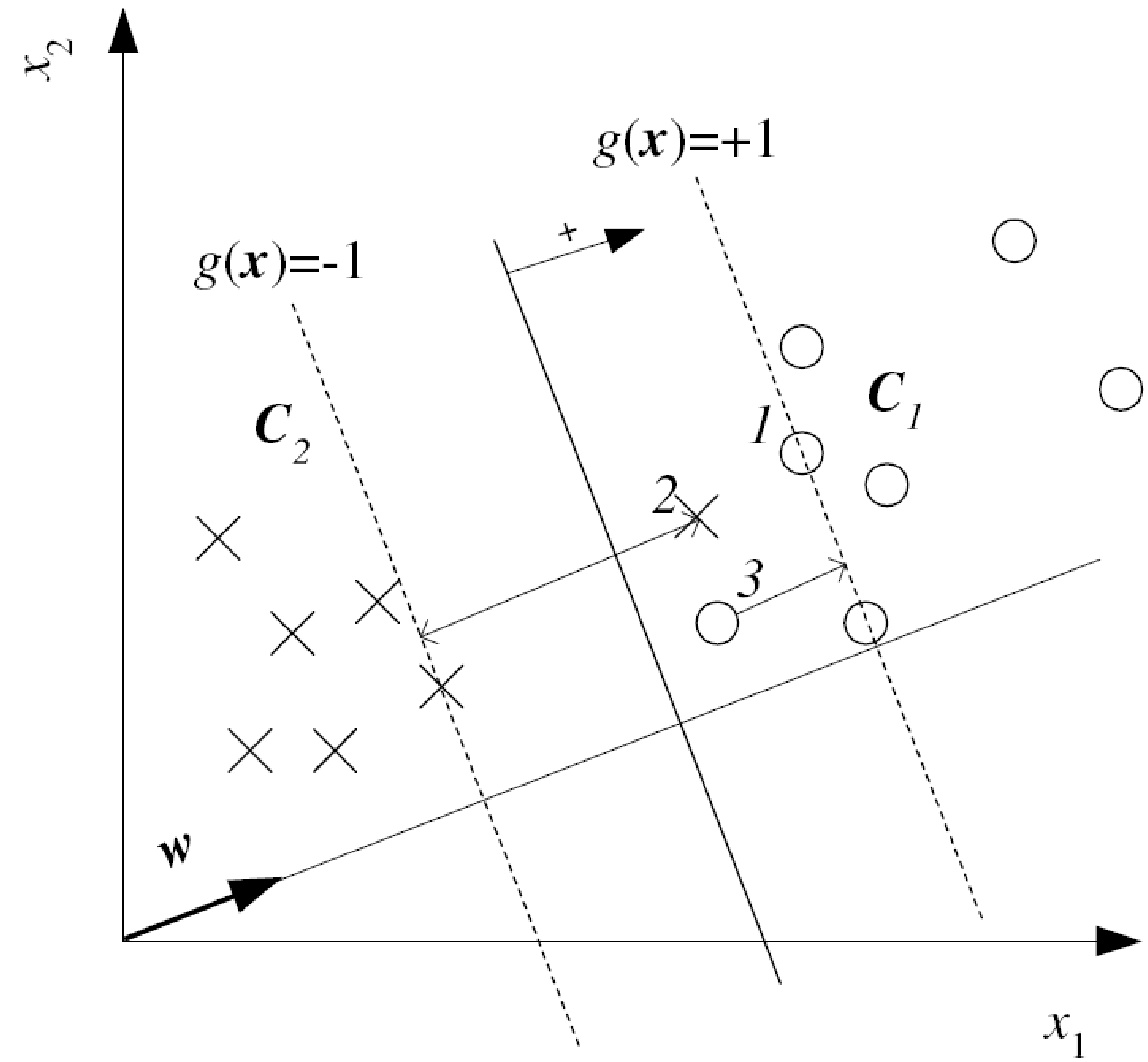
- Soft error $\sum_t \xi^t$
- New objective function:

$$\min \left[\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_t \xi^t\right]$$

subject to

$$r^t(\mathbf{w}^T\mathbf{x}^t + w_0) \geq 1 - \xi^t, \forall t$$

$$\xi^t \geq 0, \forall t$$



- New primal is

$$L_p = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_t \xi^t - \sum_{t=1}^{N} \alpha^t[r^t(\mathbf{w}^T\mathbf{x}^t + w_0) - 1 + \xi^t] - \sum_t \mu^t\xi^t$$

Kernel Machines

- Preprocess input x by basis functions
  - Suppose $\mathbf{z} = \varphi(\mathbf{x})$
  - Prepare transformed training set $\mathcal{Z} = \{\varphi(\mathbf{x}^t), r^t\}$
  - Linear model in space Z is nonlinear model in space X

    $g(\mathbf{z}) = \mathbf{w}^T\mathbf{z} \quad g(\mathbf{x}) = \mathbf{w}^T\varphi(\mathbf{x})$
- SVM on the transformed space Z
  - $\mathbf{w} = \sum_t \alpha^t r^t \mathbf{z}^t = \sum_t \alpha^t r^t \varphi(\mathbf{x}^t)$
  - $g(\mathbf{x}) = \mathbf{w}^T\varphi(\mathbf{x}) = \sum_t \alpha^t r^t \varphi(\mathbf{x}^t)^T \varphi(\mathbf{x})$
  - $g(\mathbf{x}) = \sum_t \alpha^t r^t K(\mathbf{x}^t, \mathbf{x})$
- Kernel functions K
  - Polynomials of degree q: $K(\mathbf{x}^t, \mathbf{x}) = (\mathbf{x}^T\mathbf{x}^t + 1)^q$
    - $K(\mathbf{x}, \mathbf{y}) = (\mathbf{x}^T\mathbf{y} + 1)^2 = (x_1y_1 + x_2y_2 + 1)^2$

      $= 1 + 2x_1y_1 + 2x_2y_2 + 2x_1x_2y_1y_2 + x_1^2y_1^2 + x_2^2y_2^2$

      $\varphi(\mathbf{x}) = [1, \sqrt{2}x_1, \sqrt{2}x_2, \sqrt{2}x_1x_2, x_1^2, x_2^2]^T$
  - Radial-basis functions: $K(\mathbf{x}^t, \mathbf{x}) = \exp[-\|\mathbf{x}^t - \mathbf{x}\|^2/\sigma^2]$
  - Sigmoid functions: $K(\mathbf{x}^t, \mathbf{x}) = \tanh(2\mathbf{x}^T\mathbf{x}^t + 1)$

Kernel - General Conditions

Definition

A function $K : X \times X \rightarrow \mathbb{R}$ is a positive definite kernel if for any n and any set $\{x_1, x_2, \ldots, x_n\} \subset X$, the matrix $A = (a_{ij} = K(x_i, x_j))$ is positive definite.

For any positive definite kernel, there exists a Hilbert space $\mathscr{H}$ and a *lifting map* $\Phi : X \rightarrow \mathscr{H}$ such that

$$K(x,y) = \langle \Phi(x), \Phi(y) \rangle_{\mathscr{H}}$$

A is called the Gram Matrix
A is positive definite
if $zAz^T > 0$ for nonzero $z \in R^n$

Theorem (Mercer)

If K is continuous and symmetric, then

$$K(x,y) = \sum_{0}^{\infty} \lambda_i v_i(x) v_i(y)$$

Kernel - General Conditions

Theorem (Mercer)

If K is continuous and symmetric, then

$$K(x,y) = \sum_{0}^{\infty} \lambda_i v_i(x) v_i(y)$$

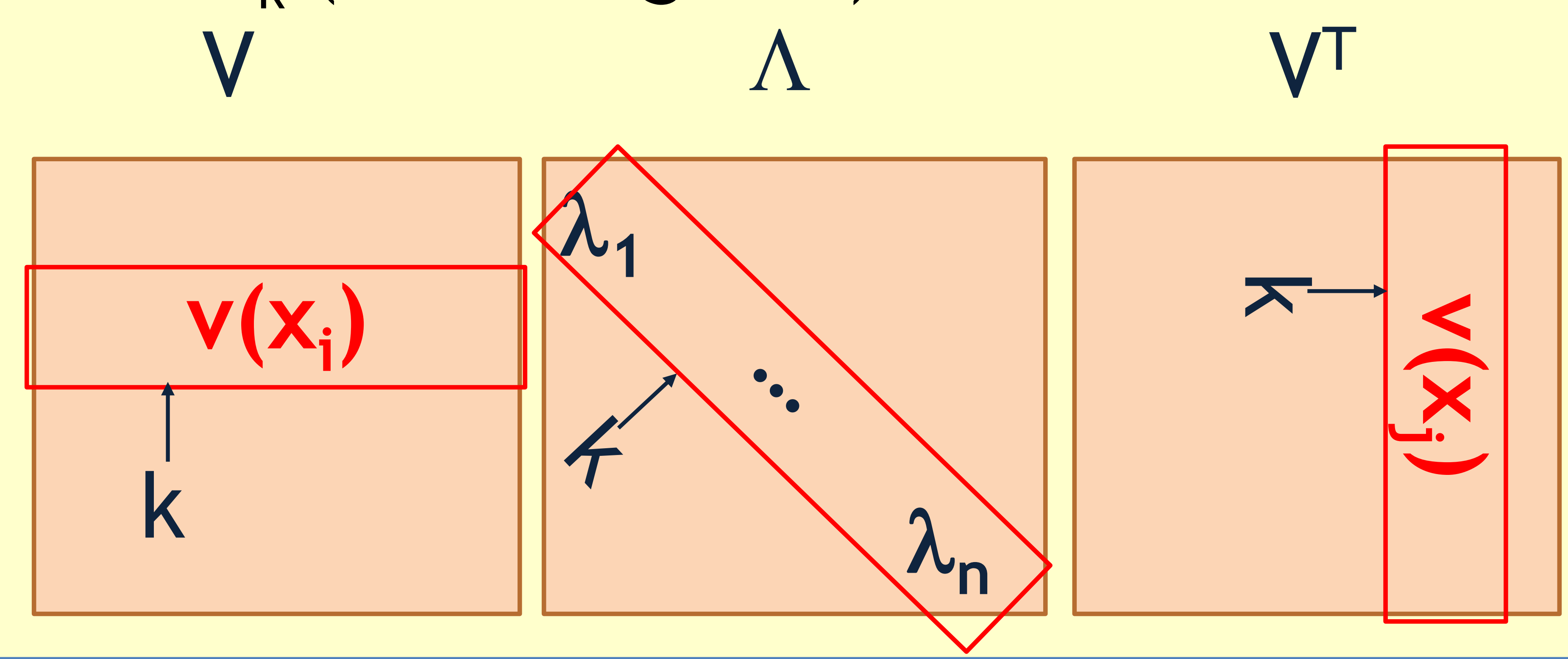
$$K(x_i, x_j) = \sum_{k=0}^{\infty} \lambda_k v_k(x_i) v_k(x_j)$$

Proofs:

Let $S \subset X$ be the set of all possible data points, the gram matrix (A, $a_{ij} = K(x_i, x_j)$) is positive semi-definite (by assumption).

If a matrix (A) is positive semi-definite, A can be factored as $A = V\Lambda V^T$ where Λ is a matrix with the non-negative eigenvalues λ_k (linear algebra).

Let $v(x_i)$ be the i'th row of V and $v_k(x_i)$ is the k'th value in the vector. Then, for any pair of x_i and x_j, $\sum_k \lambda_k v_k(x_i) v_k(x_j)$.



Examples:
https://docs.google.com/spreadsheet/ccc?key=0ArnnnlgFCwCBdDVGaDlfdXNoTGZzTG9NQzgzZHloaFE&usp=drive_web#gid=0

Kernel Machines

- Kernel functions K
  - Polynomials of degree q: $K(\mathbf{x}^t, \mathbf{x}) = (\mathbf{x}^T\mathbf{x}^t + 1)^q$
    - $K(\mathbf{x}, \mathbf{y}) = (\mathbf{x}^T\mathbf{y} + 1)^2 = (x_1y_1 + x_2y_2 + 1)^2$

$$= 1 + 2x_1y_1 + 2x_2y_2 + 2x_1x_2y_1y_2 + x_1^2y_1^2 + x_2^2y_2^2$$

$$\varphi(\mathbf{x}) = [1, \sqrt{2}x_1, \sqrt{2}x_2, \sqrt{2}x_1x_2, x_1^2, x_2^2]^T$$

  - Radial-basis functions: $K(\mathbf{x}^t, \mathbf{x}) = \exp[-\|\mathbf{x}^t - \mathbf{x}\|^2/\sigma^2]$
  - Sigmoid functions (not Mercer) : $K(\mathbf{x}^t, \mathbf{x}) = \tanh(2\mathbf{x}^T\mathbf{x}^t + 1)$
  - Cosine Similarity: similarity of two documents

$$\kappa(\mathbf{x}_i, \mathbf{x}_{i'}) = \frac{\mathbf{x}_i^T\mathbf{x}_{i'}}{||\mathbf{x}_i||_2||\mathbf{x}_{i'}||_2}$$

$$\text{tf}(x_{ij}) \triangleq \log(1 + x_{ij}) \qquad \text{tf-idf}(\mathbf{x}_i) \triangleq [\text{tf}(x_{ij}) \times \text{idf}(j)]_{j=1}^V$$

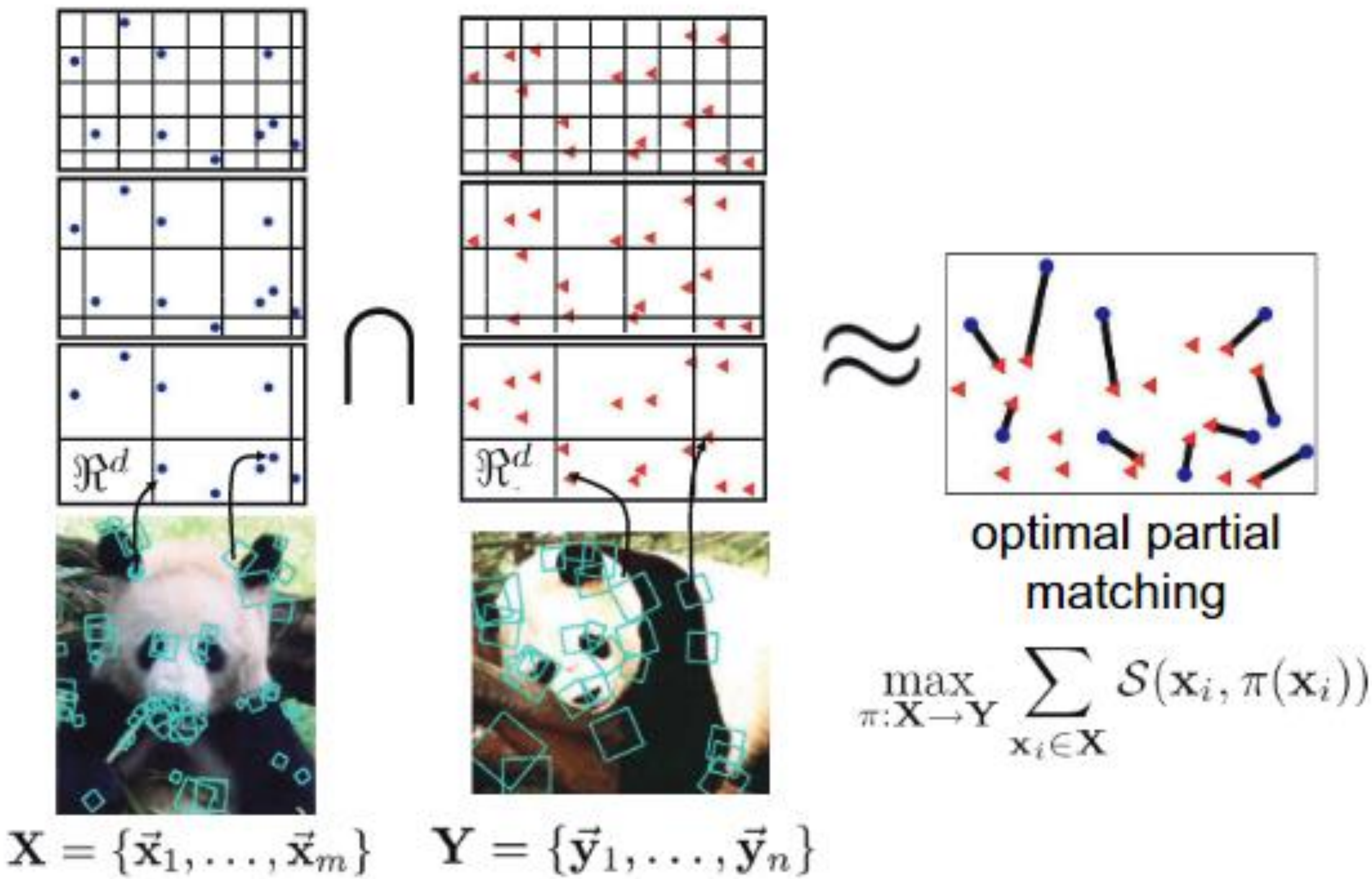
$$\text{idf}(j) \triangleq \log \frac{N}{1 + \sum_{i=1}^N \mathbb{I}(x_{ij} > 0)} \qquad \phi(\mathbf{x}) = \text{tf-idf}(\mathbf{x}).$$

$$\kappa(\mathbf{x}_i, \mathbf{x}_{i'}) = \frac{\phi(\mathbf{x}_i)^T\phi(\mathbf{x}_{i'})}{||\phi(\mathbf{x}_i)||_2||\phi(\mathbf{x}_{i'})||_2}$$

Kernel Machines

☐ Kernel functions K

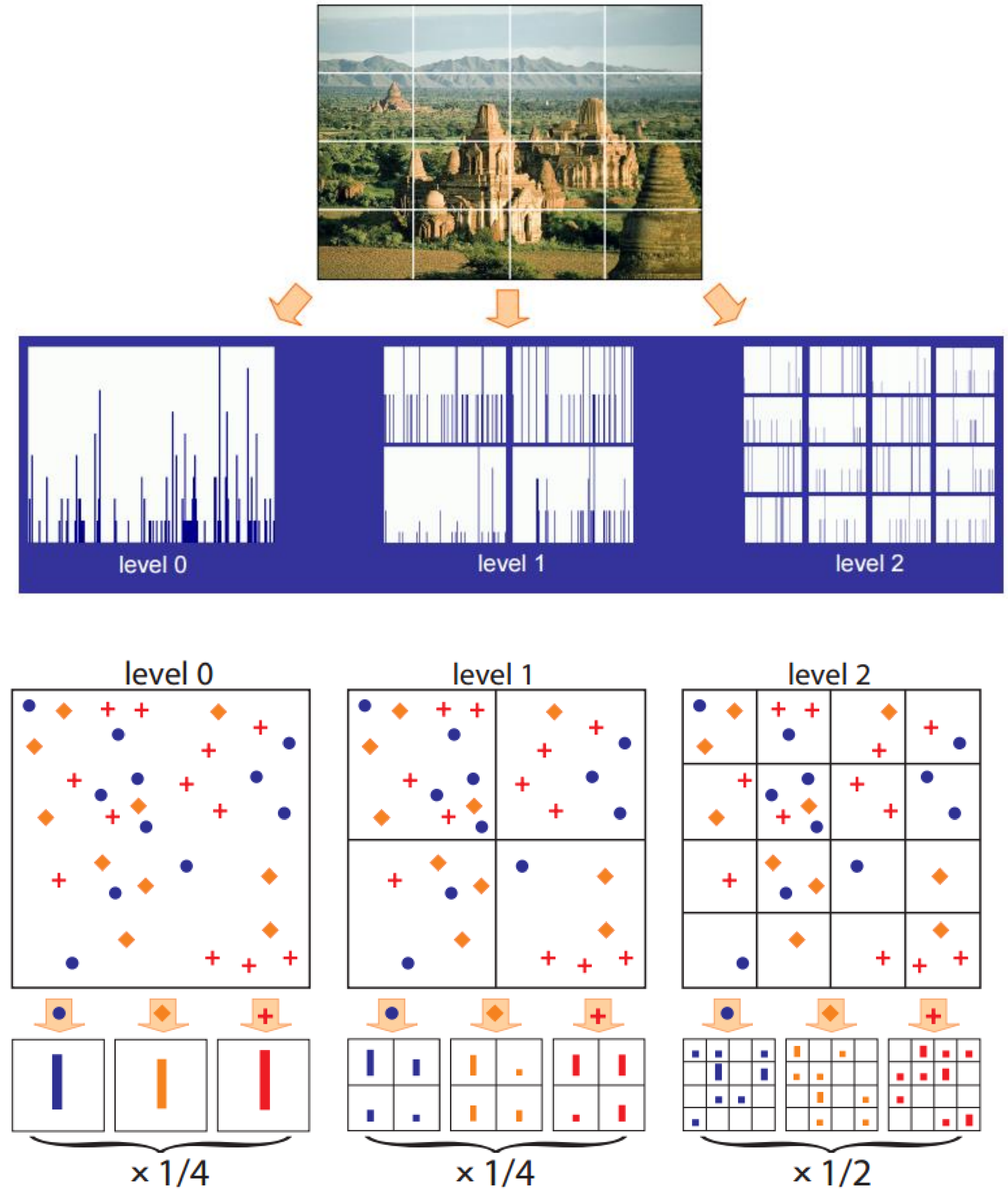
- Pyramid Matching Kernel



[Grauman and Darrell, 2006]

Kernel Machines

- Kernel functions K
  - Spatial Pyramid Matching



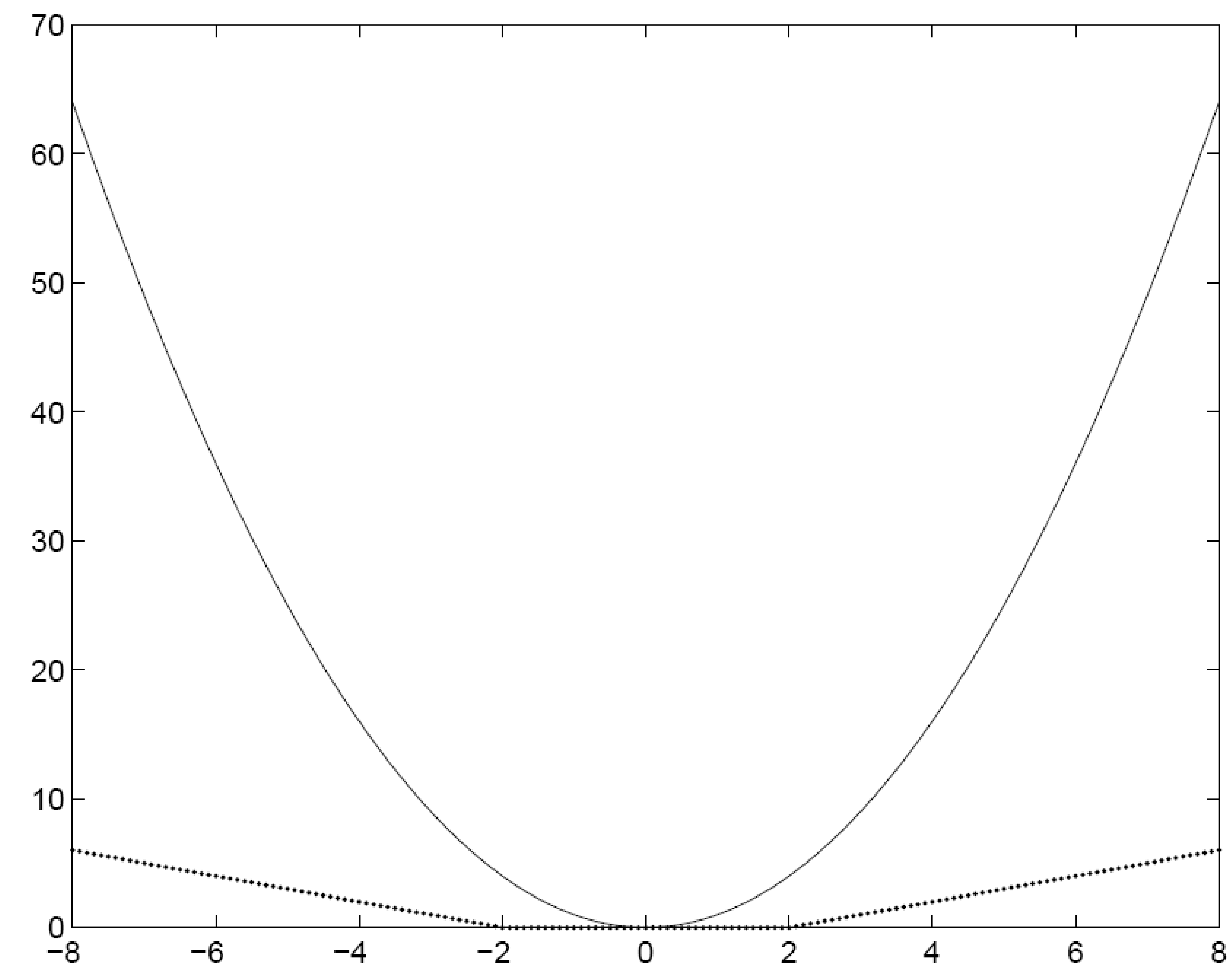
[Lazebnik, Schmid and Ponce, 2006]

SVM for Regression

- Assume a linear model (possibly kernelized)
  - $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0$

- Use ε-sensitive error function (instead of squared error function)

$$Err(r^t, f(\mathbf{x}^t)) = \begin{cases} 0 & \text{if } |r^t - f(\mathbf{x}^t)| < \epsilon \\ |r^t - f(\mathbf{x}^t)| - \epsilon & \text{otherwise} \end{cases}$$



- Problem formulation:

$$\min \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_t(\xi_+^t + \xi_-^t)$$

subject to

$$r^t - (\mathbf{w}^T\mathbf{x} + w_0) \le \epsilon + \xi_+^t$$
$$(\mathbf{w}^T\mathbf{x} + w_0) - r^t \le \epsilon + \xi_-^t$$
$$\xi_+^t, \xi_-^t \ge 0$$

SVM for Regression