

Introduction to Machine Learning - CS725  
Instructor: Prof. Ganesh Ramakrishnan  
Lecture 12 - Support Vector Regression and its Dual

# KKT conditions for the Constrained (Convex) Problem

- Here,  $\mathbf{w} \in \mathbb{R}^n$  and the domain is the intersection of all functions. Lagrangian is:

$$L(\mathbf{w}, \lambda, \mu) = f(\mathbf{w}) + \sum_{i=1}^m \lambda_i g_i(\mathbf{w}) + \sum_{j=1}^p \mu_j h_j(\mathbf{w})$$

- KKT **necessary** conditions for all differentiable functions (i.e.  $f, g_i, h_j$ ) with optimality points  $\hat{\mathbf{w}}$  and  $(\hat{\lambda}, \hat{\mu})$  are:
  - $\nabla f(\hat{\mathbf{w}}) + \sum_{i=1}^m \hat{\lambda}_i \nabla g_i(\hat{\mathbf{w}}) + \sum_{j=1}^p \hat{\mu}_j \nabla h_j(\hat{\mathbf{w}}) = 0$
  - $g_i(\hat{\mathbf{w}}) \leq 0; 1 \leq i \leq m$
  - $\hat{\lambda}_i \geq 0; 1 \leq i \leq m$
  - $\hat{\lambda}_i g_i(\hat{\mathbf{w}}) = 0; 1 \leq i \leq m$
  - $h_j(\hat{\mathbf{w}}) = 0; 1 \leq j \leq p$
- When  $f$  and  $g_i, \forall i \in [1, m]$  are convex and  $h_j, \forall j \in [1, p]$  are affine, KKT conditions are also **sufficient** for optimality at  $\hat{\mathbf{w}}$  and  $(\hat{\lambda}, \hat{\mu})$

KKT conditions for the Constrained (**Convex**) Problem  
Application 2: SVR and its Dual

# KKT conditions for SVR

$$L(\mathbf{w}, \alpha, \alpha^*, \mu, \mu^*) = \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_i (\xi_i + \xi_i^*) + \sum_{i=1}^m \alpha_i (y_i - \mathbf{w}^\top \phi(\mathbf{x}_i) - b - \epsilon - \xi_i) + \sum_{i=1}^m \alpha_i^* (b + \mathbf{w}^\top \phi(\mathbf{x}_i) - y_i - \epsilon - \xi_i^*) - \sum_{i=1}^m \mu_i \xi_i - \sum_{i=1}^m \mu_i^* \xi_i^*$$

- Differentiating the Lagrangian w.r.t.  $\mathbf{w}$ ,  
 $\mathbf{w} - \alpha_i \phi(\mathbf{x}_i) + \alpha_i^* \phi(\mathbf{x}_i) = 0$  i.e.,  $\mathbf{w} = \sum_{i=1}^m (\alpha_i - \alpha_i^*) \phi(\mathbf{x}_i)$
- Differentiating the Lagrangian w.r.t.  $\xi_i$ ,  
 $C - \alpha_i - \mu_i = 0$  i.e.,  $\alpha_i + \mu_i = C$
- Differentiating the Lagrangian w.r.t  $\xi_i^*$ ,  
 $\alpha_i^* + \mu_i^* = C$
- Differentiating the Lagrangian w.r.t  $b$ ,  
 $\sum_i (\alpha_i^* - \alpha_i) = 0$
- Complimentary slackness:  
 $\alpha_i (y_i - \mathbf{w}^\top \phi(\mathbf{x}_i) - b - \epsilon - \xi_i) = 0$  AND  $\mu_i \xi_i = 0$  AND  
 $\alpha_i^* (b + \mathbf{w}^\top \phi(\mathbf{x}_i) - y_i - \epsilon - \xi_i^*) = 0$  AND  $\mu_i^* \xi_i^* = 0$

# Support Vector Regression

## Dual Objective

# SVR Dual objective

- $L^*(\alpha, \alpha^*, \mu, \mu^*) = \min_{\mathbf{w}, b, \xi, \xi^*} L(\mathbf{w}, b, \xi, \xi^*, \alpha, \alpha^*, \mu, \mu^*)$
- Assume: In case of SVR, we have a strictly convex objective and linear constraints  
 $\Rightarrow$  KKT conditions are necessary and sufficient and strong duality holds (for  $\alpha, \alpha^* \geq 0$ ):

$$\min_{\mathbf{w}, b, \xi, \xi^*} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^m (\xi_i + \xi_i^*) = \max_{\alpha, \alpha^*, \mu, \mu^*} L^*(\alpha, \alpha^*, \mu, \mu^*)$$

s.t.  $y_i - \mathbf{w}^\top \phi(\mathbf{x}_i) - b \leq \epsilon - \xi_i$ , and  
 $\mathbf{w}^\top \phi(\mathbf{x}_i) + b - y_i \leq \epsilon - \xi_i^*$ , and  
 $\xi_i, \xi_i^* \geq 0, \forall i = 1, \dots, n$

- This value is precisely obtained at the  $\{\hat{\mathbf{w}}, \hat{b}, \hat{\xi}, \hat{\xi}^*, \hat{\alpha}, \hat{\alpha}^*, \hat{\mu}, \hat{\mu}^*\}$  that satisfies the necessary (and sufficient) KKT optimality conditions [**KKT Constraint Set**]
- Given strong duality, we can equivalently solve:  $\max_{\hat{\alpha}, \hat{\alpha}^*, \hat{\mu}, \hat{\mu}^*} L^*(\hat{\alpha}, \hat{\alpha}^*, \hat{\mu}, \hat{\mu}^*)$

- $$L(\hat{\alpha}, \hat{\alpha}^*, \hat{\mu}, \hat{\mu}^*) = \frac{1}{2} \|\hat{\mathbf{w}}\|^2 + C \sum_{i=1}^m (\hat{\xi}_i + \hat{\xi}_i^*) + \sum_{i=1}^m \left( \hat{\alpha}_i (y_i - \mathbf{w}^\top \phi(\mathbf{x}_i) - b - \epsilon - \hat{\xi}_i) + \hat{\alpha}_i^* (\mathbf{w}^\top \phi(\mathbf{x}_i) + b - y_i - \epsilon - \hat{\xi}_i^*) \right) - \sum_{i=1}^m (\hat{\mu}_i \hat{\xi}_i + \hat{\mu}_i^* \hat{\xi}_i^*)$$
- We obtain  $\hat{\mathbf{w}}, \hat{b}, \hat{\xi}_i, \hat{\xi}_i^*$  in terms of  $\hat{\alpha}, \hat{\alpha}^*, \hat{\mu}$  and  $\hat{\mu}^*$  by using the KKT conditions derived earlier as  $\hat{\mathbf{w}} = \sum_{i=1}^m (\hat{\alpha}_i - \hat{\alpha}_i^*) \phi(\mathbf{x}_i)$  and  $\sum_{i=1}^m (\hat{\alpha}_i - \hat{\alpha}_i^*) = 0$  and  $\hat{\alpha}_i + \hat{\mu}_i = C$  and  $\hat{\alpha}_i^* + \hat{\mu}_i^* = C$
- Thus, we get (after dropping the messy  $\hat{\cdot}$  notation):

- $$L(\hat{\alpha}, \hat{\alpha}^*, \hat{\mu}, \hat{\mu}^*) = \frac{1}{2} \|\hat{\mathbf{w}}\|^2 + C \sum_{i=1}^m (\hat{\xi}_i + \hat{\xi}_i^*) + \sum_{i=1}^m \left( \hat{\alpha}_i (y_i - \mathbf{w}^\top \phi(\mathbf{x}_i) - b - \epsilon - \hat{\xi}_i) + \hat{\alpha}_i^* (\mathbf{w}^\top \phi(\mathbf{x}_i) + b - y_i - \epsilon - \hat{\xi}_i^*) \right) - \sum_{i=1}^m (\hat{\mu}_i \hat{\xi}_i + \hat{\mu}_i^* \hat{\xi}_i^*)$$
- We obtain  $\hat{\mathbf{w}}, \hat{b}, \hat{\xi}_i, \hat{\xi}_i^*$  in terms of  $\hat{\alpha}, \hat{\alpha}^*, \hat{\mu}$  and  $\hat{\mu}^*$  by using the KKT conditions derived earlier as  $\hat{\mathbf{w}} = \sum_{i=1}^m (\hat{\alpha}_i - \hat{\alpha}_i^*) \phi(\mathbf{x}_i)$  and  $\sum_{i=1}^m (\hat{\alpha}_i - \hat{\alpha}_i^*) = 0$  and  $\hat{\alpha}_i + \hat{\mu}_i = C$  and  $\hat{\alpha}_i^* + \hat{\mu}_i^* = C$
- Thus, we get (after dropping the messy  $\hat{\cdot}$  notation):

$$L(\mathbf{w}, b, \xi, \xi^*, \alpha, \alpha^*, \mu, \mu^*) = \frac{1}{2} \sum_i \sum_j (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) \phi^\top(\mathbf{x}_i) \phi(\mathbf{x}_j) + \sum_i (\xi_i (C - \alpha_i - \mu_i) + \xi_i^* (C - \alpha_i^* - \mu_i^*)) - b \sum_i (\alpha_i - \alpha_i^*) - \epsilon \sum_i (\alpha_i + \alpha_i^*) + \sum_i y_i (\alpha_i - \alpha_i^*) - \sum_i \sum_j (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) \phi^\top(\mathbf{x}_i) \phi(\mathbf{x}_j)$$



- $$L(\hat{\alpha}, \hat{\alpha}^*, \hat{\mu}, \hat{\mu}^*) = \frac{1}{2} \|\hat{\mathbf{w}}\|^2 + C \sum_{i=1}^m (\hat{\xi}_i + \hat{\xi}_i^*) + \sum_{i=1}^m \left( \hat{\alpha}_i (y_i - \mathbf{w}^\top \phi(\mathbf{x}_i) - b - \epsilon - \hat{\xi}_i) + \hat{\alpha}_i^* (\mathbf{w}^\top \phi(\mathbf{x}_i) + b - y_i - \epsilon - \hat{\xi}_i^*) \right) - \sum_{i=1}^m (\hat{\mu}_i \hat{\xi}_i + \hat{\mu}_i^* \hat{\xi}_i^*)$$
- We obtain  $\hat{\mathbf{w}}, \hat{b}, \hat{\xi}_i, \hat{\xi}_i^*$  in terms of  $\hat{\alpha}, \hat{\alpha}^*, \hat{\mu}$  and  $\hat{\mu}^*$  by using the KKT conditions derived earlier as  $\hat{\mathbf{w}} = \sum_{i=1}^m (\hat{\alpha}_i - \hat{\alpha}_i^*) \phi(\mathbf{x}_i)$  and  $\sum_{i=1}^m (\hat{\alpha}_i - \hat{\alpha}_i^*) = 0$  and  $\hat{\alpha}_i + \hat{\mu}_i = C$  and  $\hat{\alpha}_i^* + \hat{\mu}_i^* = C$
- Thus, we get (after dropping the messy  $\hat{\cdot}$  notation):

$$\begin{aligned} L(\mathbf{w}, b, \xi, \xi^*, \alpha, \alpha^*, \mu, \mu^*) &= \frac{1}{2} \sum_i \sum_j (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) \phi^\top(\mathbf{x}_i) \phi(\mathbf{x}_j) + \sum_i (\xi_i (C - \alpha_i - \mu_i) + \xi_i^* (C - \alpha_i^* - \mu_i^*)) - b \sum_i (\alpha_i - \alpha_i^*) - \epsilon \sum_i (\alpha_i + \alpha_i^*) + \sum_i y_i (\alpha_i - \alpha_i^*) - \sum_i \sum_j (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) \phi^\top(\mathbf{x}_i) \phi(\mathbf{x}_j) \\ &= -\frac{1}{2} \sum_i \sum_j (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) \phi^\top(\mathbf{x}_i) \phi(\mathbf{x}_j) - \epsilon \sum_i (\alpha_i + \alpha_i^*) + \sum_i y_i (\alpha_i - \alpha_i^*) \end{aligned}$$

# SVR Dual Formulation using only dot products $\phi^T(\mathbf{x}_i)\phi(\mathbf{x}_j)$

- $\mathbf{w} = \sum_{i=1}^m (\alpha_i - \alpha_i^*) \phi(\mathbf{x}_i) \Rightarrow$  the final decision function  
$$f(\mathbf{x}) = \mathbf{w}^T \phi(\mathbf{x}) + b = \sum_{i=1}^m (\alpha_i - \alpha_i^*) \phi^T(\mathbf{x}_i) \phi(\mathbf{x}) + y_j - \sum_{i=1}^m (\alpha_i - \alpha_i^*) \phi^T(\mathbf{x}_i) \phi(\mathbf{x}_j) - \epsilon$$
 $\mathbf{x}_j$  is any point with  $\alpha_j \in (0, C)$ . **Tutorial 5: Derive kernelized expression for Ridge Regression**
- The dual optimization problem to compute the  $\alpha$ 's for SVR is:

# SVR Dual Formulation using only dot products $\phi^T(\mathbf{x}_i)\phi(\mathbf{x}_j)$

- $\mathbf{w} = \sum_{i=1}^m (\alpha_i - \alpha_i^*) \phi(\mathbf{x}_i) \Rightarrow$  the final decision function  
 $f(\mathbf{x}) = \mathbf{w}^T \phi(\mathbf{x}) + b = \sum_{i=1}^m (\alpha_i - \alpha_i^*) \phi^T(\mathbf{x}_i) \phi(\mathbf{x}) + y_j - \sum_{i=1}^m (\alpha_i - \alpha_i^*) \phi^T(\mathbf{x}_i) \phi(\mathbf{x}_j) - \epsilon$   
 $\mathbf{x}_j$  is any point with  $\alpha_j \in (0, C)$ . **Tutorial 5: Derive kernelized expression for Ridge Regression**
- The dual optimization problem to compute the  $\alpha$ 's for SVR is:

$$\begin{aligned} \max_{\alpha_i, \alpha_i^*} & -\frac{1}{2} \sum_i \sum_j (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) \phi^T(\mathbf{x}_i) \phi(\mathbf{x}_j) \\ & - \epsilon \sum_i (\alpha_i + \alpha_i^*) + \sum_i y_i (\alpha_i - \alpha_i^*) \end{aligned}$$

s.t.

- $\sum_i (\alpha_i - \alpha_i^*) = 0$
- $\alpha_i, \alpha_i^* \in [0, C]$
- We notice that the only way these three expressions involve  $\phi$  is through  $\phi^T(\mathbf{x}_i) \phi(\mathbf{x}_j) = K(\mathbf{x}_i, \mathbf{x}_j)$ , for some  $i, j$

## Support Vectors: Non-zero contribution $\alpha_i - \alpha_i^*$ outside $\epsilon$ -band

- **For any point  $(\mathbf{x}_i, y_i)$ , the product  $\alpha_i \alpha_i^* = 0$ .**

# Support Vectors: Non-zero contribution $\alpha_i - \alpha_i^*$ outside $\epsilon$ -band

- **For any point  $(\mathbf{x}_i, y_i)$ , the product  $\alpha_i \alpha_i^* = 0$ .**
  - Let  $\alpha_i > 0$  and  $\alpha_i^* > 0$ . This leads to a contradiction.
  - By Complimentary slackness,  $y_i - \mathbf{w}^\top \phi(\mathbf{x}_i) - b - \epsilon - \xi_i = 0$  AND  $b + \mathbf{w}^\top \phi(\mathbf{x}_i) - y_i - \epsilon - \xi_i^* = 0$ . Adding up the two equalities gives us:  $\xi_i + \xi_i^* = -2\epsilon$ .
  - Since only one of  $\xi_i$  and  $\xi_i^*$  can be non-zero,  $\implies$  the non-zero component is negative, which is a contradiction since  $\xi_i, \xi_i^* \geq 0$
  - Thus,  $\alpha_i - \alpha_i^* \propto \max\{\alpha_i, \alpha_i^*\}$
- **For points within the  $\epsilon$ -insensitive tube  $\alpha_i = 0$  and  $\alpha_i^* = 0$ :**

# Support Vectors: Non-zero contribution $\alpha_i - \alpha_i^*$ outside $\epsilon$ -band

- **For any point  $(\mathbf{x}_i, y_i)$ , the product  $\alpha_i \alpha_i^* = 0$ .**
  - Let  $\alpha_i > 0$  and  $\alpha_i^* > 0$ . This leads to a contradiction.
  - By Complimentary slackness,  $y_i - \mathbf{w}^\top \phi(\mathbf{x}_i) - b - \epsilon - \xi_i = 0$  AND  $b + \mathbf{w}^\top \phi(\mathbf{x}_i) - y_i - \epsilon - \xi_i^* = 0$ . Adding up the two equalities gives us:  $\xi_i + \xi_i^* = -2\epsilon$ .
  - Since only one of  $\xi_i$  and  $\xi_i^*$  can be non-zero,  $\implies$  the non-zero component is negative, which is a contradiction since  $\xi_i, \xi_i^* \geq 0$
  - Thus,  $\alpha_i - \alpha_i^* \propto \max\{\alpha_i, \alpha_i^*\}$
- **For points within the  $\epsilon$ -insensitive tube  $\alpha_i = 0$  and  $\alpha_i^* = 0$ :**
  - If  $y_i - \mathbf{w}^\top \phi(\mathbf{x}_i) - b - \epsilon - \xi_i < 0$ , then  $\alpha_i = 0$ ,  $\mu_i = C$  and  $\xi_i = 0$ . Similarly,  $b + \mathbf{w}^\top \phi(\mathbf{x}_i) - y_i - \epsilon < 0$  leading to  $\alpha_i^* = 0$ .
- **$\alpha_i = C$  and  $\alpha_i^* = C$  correspond to points lying either outside or on the  $\epsilon$ -tube:**

# Support Vectors: Non-zero contribution $\alpha_i - \alpha_i^*$ outside $\epsilon$ -band

- **For any point  $(\mathbf{x}_i, y_i)$ , the product  $\alpha_i \alpha_i^* = 0$ .**
  - Let  $\alpha_i > 0$  and  $\alpha_i^* > 0$ . This leads to a contradiction.
  - By Complimentary slackness,  $y_i - \mathbf{w}^\top \phi(\mathbf{x}_i) - b - \epsilon - \xi_i = 0$  AND  $b + \mathbf{w}^\top \phi(\mathbf{x}_i) - y_i - \epsilon - \xi_i^* = 0$ . Adding up the two equalities gives us:  $\xi_i + \xi_i^* = -2\epsilon$ .
  - Since only one of  $\xi_i$  and  $\xi_i^*$  can be non-zero,  $\implies$  the non-zero component is negative, which is a contradiction since  $\xi_i, \xi_i^* \geq 0$
  - Thus,  $\alpha_i - \alpha_i^* \propto \max\{\alpha_i, \alpha_i^*\}$
- **For points within the  $\epsilon$ -insensitive tube  $\alpha_i = 0$  and  $\alpha_i^* = 0$ :**
  - If  $y_i - \mathbf{w}^\top \phi(\mathbf{x}_i) - b - \epsilon - \xi_i < 0$ , then  $\alpha_i = 0$ ,  $\mu_i = C$  and  $\xi_i = 0$ . Similarly,  $b + \mathbf{w}^\top \phi(\mathbf{x}_i) - y_i - \epsilon < 0$  leading to  $\alpha_i^* = 0$ .
- **$\alpha_i = C$  and  $\alpha_i^* = C$  correspond to points lying either outside or on the  $\epsilon$ -tube:**
  - If  $\alpha_i = C$ , then  $\mu_i = 0$  and  $y_i - \mathbf{w}^\top \phi(\mathbf{x}_i) - b - \epsilon = \xi_i \geq 0$ .
  - Similarly,  $\alpha_i^* = C$  corresponds to points lying below (or beyond) the lower  $\epsilon$ -band.
- **For points on boundary of the  $\epsilon$ -insensitive tube  $\alpha_i \in [0, C]$ :**

# Support Vectors: Non-zero contribution $\alpha_i - \alpha_i^*$ outside $\epsilon$ -band

- **For any point  $(\mathbf{x}_i, y_i)$ , the product  $\alpha_i \alpha_i^* = 0$ .**
  - Let  $\alpha_i > 0$  and  $\alpha_i^* > 0$ . This leads to a contradiction.
  - By Complimentary slackness,  $y_i - \mathbf{w}^\top \phi(\mathbf{x}_i) - b - \epsilon - \xi_i = 0$  AND  $b + \mathbf{w}^\top \phi(\mathbf{x}_i) - y_i - \epsilon - \xi_i^* = 0$ . Adding up the two equalities gives us:  $\xi_i + \xi_i^* = -2\epsilon$ .
  - Since only one of  $\xi_i$  and  $\xi_i^*$  can be non-zero,  $\implies$  the non-zero component is negative, which is a contradiction since  $\xi_i, \xi_i^* \geq 0$
  - Thus,  $\alpha_i - \alpha_i^* \propto \max\{\alpha_i, \alpha_i^*\}$
- **For points within the  $\epsilon$ -insensitive tube  $\alpha_i = 0$  and  $\alpha_i^* = 0$ :**
  - If  $y_i - \mathbf{w}^\top \phi(\mathbf{x}_i) - b - \epsilon - \xi_i < 0$ , then  $\alpha_i = 0$ ,  $\mu_i = C$  and  $\xi_i = 0$ . Similarly,  $b + \mathbf{w}^\top \phi(\mathbf{x}_i) - y_i - \epsilon < 0$  leading to  $\alpha_i^* = 0$ .
- **$\alpha_i = C$  and  $\alpha_i^* = C$  correspond to points lying either outside or on the  $\epsilon$ -tube:**
  - If  $\alpha_i = C$ , then  $\mu_i = 0$  and  $y_i - \mathbf{w}^\top \phi(\mathbf{x}_i) - b - \epsilon = \xi_i \geq 0$ .
  - Similarly,  $\alpha_i^* = C$  corresponds to points lying below (or beyond) the lower  $\epsilon$ -band.
- **For points on boundary of the  $\epsilon$ -insensitive tube  $\alpha_i \in [0, C]$ :**
  - For any point on the upper margin,  $y_i - \mathbf{w}^\top \phi(\mathbf{x}_i) - b - \epsilon = 0$  and  $\xi_i = 0 \implies \mu_i \geq 0 \implies \alpha_i \in [0, C]$ . Similarly,  $\alpha_i^* \in [0, C]$  for points lying on the margin of the lower  $\epsilon$ -band.



# Kernel function: $K(\mathbf{x}_i, \mathbf{x}_j) = \phi^\top(\mathbf{x}_i)\phi(\mathbf{x}_j)$

- We call  $\phi^\top(\mathbf{x}_i)\phi(\mathbf{x}_j)$  a **kernel function**:  
 $K(\mathbf{x}_i, \mathbf{x}_j) = \phi^\top(\mathbf{x}_i)\phi(\mathbf{x}_j)$
- The Kernel Trick: For some important choices of  $\phi$ , compute  $K(\mathbf{x}_i, \mathbf{x}_j)$  directly and more efficiently than having to explicitly compute/enumerate  $\phi(\mathbf{x}_i)$  and  $\phi(\mathbf{x}_j)$
- The expression for decision function becomes  $f(x) = \sum_{i=1}^m \alpha_i K(\mathbf{x}, \mathbf{x}_i)$
- Computation of  $\alpha_i$  is specific to the objective function being minimized: Closed form exists for Ridge regression but NOT for SVR

# The Kernelized version of SVR

- The kernelized dual problem:

$$\begin{aligned} \max_{\alpha_i, \alpha_i^*} & -\frac{1}{2} \sum_i \sum_j (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) K(\mathbf{x}_i, \mathbf{x}_j) \\ & -\epsilon \sum_i (\alpha_i + \alpha_i^*) + \sum_i y_i (\alpha_i - \alpha_i^*) \end{aligned}$$

s.t.

- $\sum_i (\alpha_i - \alpha_i^*) = 0$
- $\alpha_i, \alpha_i^* \in [0, C]$
- The kernelized decision function:  
 $f(\mathbf{x}) = \sum_i (\alpha_i - \alpha_i^*) K(\mathbf{x}_i, \mathbf{x}) + b$
- Using any point  $\mathbf{x}_j$  with  $\alpha_j \in (0, C)$ :  
 $b = y_j - \sum_i (\alpha_i - \alpha_i^*) K(\mathbf{x}_i, \mathbf{x}_j)$
- Computing  $K(\mathbf{x}_1, \mathbf{x}_2)$  often does not even require computing  $\phi(\mathbf{x}_1)$  or  $\phi(\mathbf{x}_2)$  explicitly

# Basis function expansion and the Kernel trick

- We started off with the functional form<sup>1</sup>

$$f(\mathbf{x}) = \sum_{j=1}^p w_j \phi_j(\mathbf{x})$$

Each  $\phi_j$  is called a *basis function* and this representation is called *basis function expansion*<sup>2</sup>

- And we landed up with an equivalent

$$f(\mathbf{x}) = \sum_{i=1}^m \alpha_i K(\mathbf{x}, \mathbf{x}_i)$$

for Ridge regression and Support Vector Regression

- Aside: For  $p \in [0, \infty)$ , with what  $K$ , kind of regularizers, loss functions, etc., will these dual representations hold?<sup>3</sup>

<sup>1</sup>The additional  $b$  term can be either absorbed in  $\phi$  or kept separate as discussed on several occasions.

<sup>2</sup>Section 2.8.3 of Tibshi

<sup>3</sup>Section 5.9.1 of Tibshi

# Tutorial 5: Kernelizing Ridge Regression

- Given  $w = (\Phi^T \Phi + \lambda I)^{-1} \Phi^T y$  and using the identity  $(P^{-1} + B^T R^{-1} B)^{-1} B^T R^{-1} = P B^T (B P B^T + R)^{-1}$ 
  - $\Rightarrow w = \Phi^T (\Phi \Phi^T + \lambda I)^{-1} y = \sum_{i=1}^m \alpha_i \phi(x_i)$  where  $\alpha_i = ((\Phi \Phi^T + \lambda I)^{-1} y)_i$
  - $\Rightarrow$  the final decision function  $f(\mathbf{x}) = \phi^T(\mathbf{x}) \mathbf{w} = \sum_{i=1}^m \alpha_i \phi^T(\mathbf{x}) \phi(\mathbf{x}_i)$
- Again, **We notice that the only way the decision function  $f(\mathbf{x})$  involves  $\phi$  is through  $\phi^T(\mathbf{x}_i) \phi(\mathbf{x}_j)$ , for some  $i, j$**

# The Representer Theorem & Reproducing Kernel Hilbert Space (RKHS)

- ① The solution  $f^* \in \mathcal{H}$  (Hilbert space) to the following problem

$$f^* = \arg \min_{f \in \mathcal{H}} \sum_{i=1}^m \mathbf{E} \left( f \left( \mathbf{x}^{(i)} \right), y^{(i)} \right) + \Omega(\|f\|_K)$$

can be always written as  $f^*(\mathbf{x}) = \sum_{i=1}^m \alpha_i K(\mathbf{x}, \mathbf{x}^{(i)})$ , provided  $\Omega(\|f\|_K)$  is a ....

- ② More specifically, if  $f(\mathbf{x}) = \mathbf{w}^T \phi(\mathbf{x}) + b$  and  $K(\mathbf{x}', \mathbf{x}) = \phi^T(\mathbf{x})\phi(\mathbf{x}')$  then the solution  $\mathbf{w}^* \in \Re^n$  to the following problem

$$(\mathbf{w}^*, b^*) = \arg \min_{\mathbf{w}, b} \sum_{i=1}^m \mathbf{E} \left( f \left( \mathbf{x}^{(i)} \right), y^{(i)} \right) + \Omega(\|\mathbf{w}\|_2)$$

can be always written as  $\phi^T(\mathbf{x})\mathbf{w}^* + b = \sum_{i=1}^m \alpha_i K(\mathbf{x}, \mathbf{x}^{(i)})$ , provided  $\Omega(\|\mathbf{w}\|_2)$  is a monotonically increasing function of  $\|\mathbf{w}\|_2$ .  $\Re^n$  is the Hilbert space and  $K(., \mathbf{x}) : \mathcal{X} \rightarrow \Re$  is the **Reproducing (RKHS) Kernel**

# The Representer Theorem and SVR

## 1 The SVR Objective

$$(\mathbf{w}^*, b^*, \xi_i^*) = \arg \min_{\mathbf{w}, b, \xi_i} C \sum_{i=1}^m (\xi_i + \xi_i^*) + \frac{1}{2} \|\mathbf{w}\|_2^2$$

s.t.  $y_i - \mathbf{w}^\top \phi(\mathbf{x}_i) - b \leq \epsilon - \xi_i$ , and

$\mathbf{w}^\top \phi(\mathbf{x}_i) + b - y_i \leq \epsilon - \xi_i^*$ , and

$\xi_i, \xi_i^* \geq 0, \forall i = 1, \dots, m$

## 2 Can be rewritten as

# The Representer Theorem and SVR

## 1 The SVR Objective

$$(\mathbf{w}^*, b^*, \xi_i^*) = \arg \min_{\mathbf{w}, b, \xi_i} C \sum_{i=1}^m (\xi_i + \xi_i^*) + \frac{1}{2} \|\mathbf{w}\|_2^2$$

s.t.  $y_i - \mathbf{w}^\top \phi(\mathbf{x}_i) - b \leq \epsilon - \xi_i$ , and

$\mathbf{w}^\top \phi(\mathbf{x}_i) + b - y_i \leq \epsilon - \xi_i^*$ , and

$\xi_i, \xi_i^* \geq 0, \forall i = 1, \dots, n$

## 2 Can be rewritten as

$$(\mathbf{w}^*, b^*, \xi_i^*) = \arg \min_{\mathbf{w}, b, \xi_i} C \sum_{i=1}^m \max \left\{ \epsilon \pm \left( y_i - \mathbf{w}^\top \phi(\mathbf{x}_i) - b \right), 0 \right\} + \frac{1}{2} \|\mathbf{w}\|_2^2$$

# The Representer Theorem and SVR (contd.)

- ① If  $f(\mathbf{x}) = \mathbf{w}^T \phi(\mathbf{x}) + b$  and  $K(\mathbf{x}', \mathbf{x}) = \phi^T(\mathbf{x})\phi(\mathbf{x}')$  and given the SVR objective

$$(\mathbf{w}^*, b^*, \xi_i^*) = \arg \min_{\mathbf{w}, b, \xi_i} C \sum_{i=1}^m \max \left\{ \epsilon \pm (y_i - \mathbf{w}^T \phi(\mathbf{x}_i) - b), 0 \right\} + \frac{1}{2} \|\mathbf{w}\|_2^2$$

- ② Setting  $\mathbf{E}(f(\mathbf{x}^{(i)}), y^{(i)}) = C \max \{ \epsilon \pm (y_i - \mathbf{w}^T \phi(\mathbf{x}_i) - b), 0 \}$  and  $\Omega(\|\mathbf{w}\|_2) = \frac{1}{2} \|\mathbf{w}\|_2^2$ , we can apply the Representer theorem to SVR, so that  $\phi^T(\mathbf{x})\mathbf{w}^* + b = \sum_{i=1}^m \alpha_i K(\mathbf{x}, \mathbf{x}^{(i)})$



# An Example Kernel

Assume:  $x_1, x_2 \in \mathbb{R}^2$

- Let  $K(x_1, x_2) = (1 + x_1^\top x_2)^2$
- Which value of  $\phi(x)$  will yield  $\phi^\top(x_1)\phi(x_2) = K(x_1, x_2) = (1 + x_1^\top x_2)^2$
- Is such a  $\phi$  guaranteed to exist?
- Is there a unique  $\phi$  for given  $K$ ?

$$K(x_1, x_2) = (1 + x_{11}x_{21} + x_{12}x_{22})^2 = (1 + x_{11}^2x_{21}^2 + x_{12}^2x_{22}^2 + 2x_{11}x_{21}x_{12}x_{22} + 2x_{11}x_{21} + 2x_{12}x_{22})$$

$$\phi(x_i) = [1, x_{i1}^2, x_{i2}^2, \sqrt{2}x_{i1}x_{i2}, \sqrt{2}x_{i1}, \sqrt{2}x_{i2}]$$

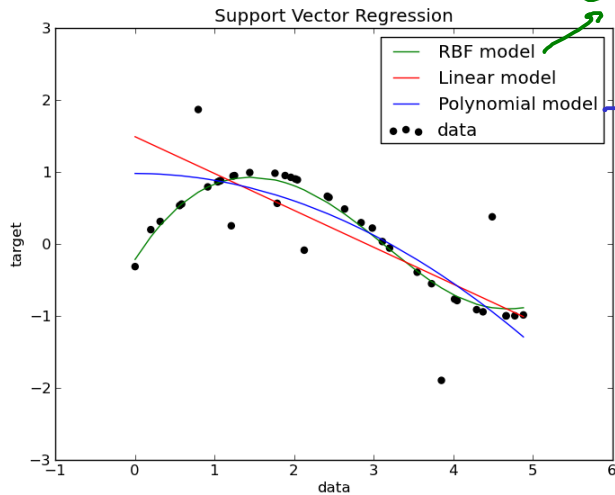
# An Example Kernel

- We can prove that such a  $\phi$  exists
- For example, for a 2-dimensional  $\mathbf{x}_i$ :

$$\phi(\mathbf{x}_i) = \begin{bmatrix} 1 \\ x_{i1}\sqrt{2} \\ x_{i2}\sqrt{2} \\ x_{i1}x_{i2}\sqrt{2} \\ x_{i1}^2 \\ x_{i2}^2 \end{bmatrix}$$

$$\left. \begin{array}{l} 1 \\ x_{i1}\sqrt{2} \\ x_{i2}\sqrt{2} \\ x_{i1}x_{i2}\sqrt{2} \\ x_{i1}^2 \\ x_{i2}^2 \end{array} \right\} (1 + \mathbf{x}_i^\top \mathbf{x}_i)^2$$

- $\phi(\mathbf{x}_i)$  exists in a 6-dimensional space
- But, to compute  $K(\mathbf{x}_1, \mathbf{x}_2)$ , all we need is  $\mathbf{x}_1^\top \mathbf{x}_2$  without having to enumerate  $\phi(\mathbf{x}_i)$



$$e^{-\frac{\|x_1 - x_2\|^2}{2\sigma^2}}$$

$$(1 + x_1^T x_2)^d$$

# More on the Kernel Trick

If  $K$  is indeed a "valid" kernel Gram matrix MUST be PSD

**Kernels** operate in a *high-dimensional, implicit* feature space without necessarily computing the coordinates of the data in that space, but rather by simply computing the Kernel function

- This approach is called the "kernel trick" and will subsequently talk about *valid kernels*
- This operation is often computationally cheaper than the explicit computation of the coordinates
- Claim: If  $\mathcal{K}_{ij} = K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$  are entries of an  $n \times n$  **Gram Matrix**  $\mathcal{K}$  then

- $\mathcal{K}$  must be positive semi-definite

- Proof:  $\mathbf{b}^T \mathcal{K} \mathbf{b} = \sum_{i,j} b_i \mathcal{K}_{ij} b_j = \sum_{i,j} b_i b_j \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle$

$$= \langle \sum_i b_i \phi(\mathbf{x}_i), \sum_j b_j \phi(\mathbf{x}_j) \rangle = \left\| \sum_i b_i \phi(\mathbf{x}_i) \right\|_2^2 \geq 0$$

$$\mathcal{K} = \begin{bmatrix} K(x_1, x_1) & K(x_1, x_2) & \dots \\ \vdots & \ddots & \ddots \\ \dots & \dots & K(x_m, x_m) \end{bmatrix}$$

Input =  $\phi(\cdot)$   
Q: Does  $\exists \phi$  s.t.  $K(x_1, x_2) = \langle \phi(x_1), \phi(x_2) \rangle$

# Existence of basis expansion $\phi$ for symmetric $K$ ?

- *Positive-definite kernel*: For any dataset  $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m\}$  and for any  $m$ , the Gram matrix  $\mathcal{K}$  must be positive definite

*Eigenvalue  
decomposition  
symmetric  
psd*

$$\mathcal{K} = \begin{bmatrix} K(\mathbf{x}_1, \mathbf{x}_1) & \dots & K(\mathbf{x}_1, \mathbf{x}_n) \\ \dots & K(\mathbf{x}_i, \mathbf{x}_j) & \dots \\ K(\mathbf{x}_m, \mathbf{x}_1) & \dots & K(\mathbf{x}_m, \mathbf{x}_m) \end{bmatrix}$$

so that  $\mathcal{K} = \underline{U\Sigma U^T} = (U\Sigma^{\frac{1}{2}})(U\Sigma^{\frac{1}{2}})^T = RR^T$  where rows of  $U$  are linearly independent and  $\Sigma$  is a positive diagonal matrix

*$\phi\phi^T$  with  $R=\phi$*

*Good intuition but impractical*

NEXT CLASS: We extend EIGENVALUE  
DECOMPOSITION TO EIGENFUNCTION

DECOMPOSITION (Mercer's Theorem)

<sup>4</sup>Eigen-decomposition wrt linear operators. See  
[https://en.wikipedia.org/wiki/Mercer%27s\\_theorem](https://en.wikipedia.org/wiki/Mercer%27s_theorem)

<sup>5</sup>That is, if every Cauchy sequence is convergent.