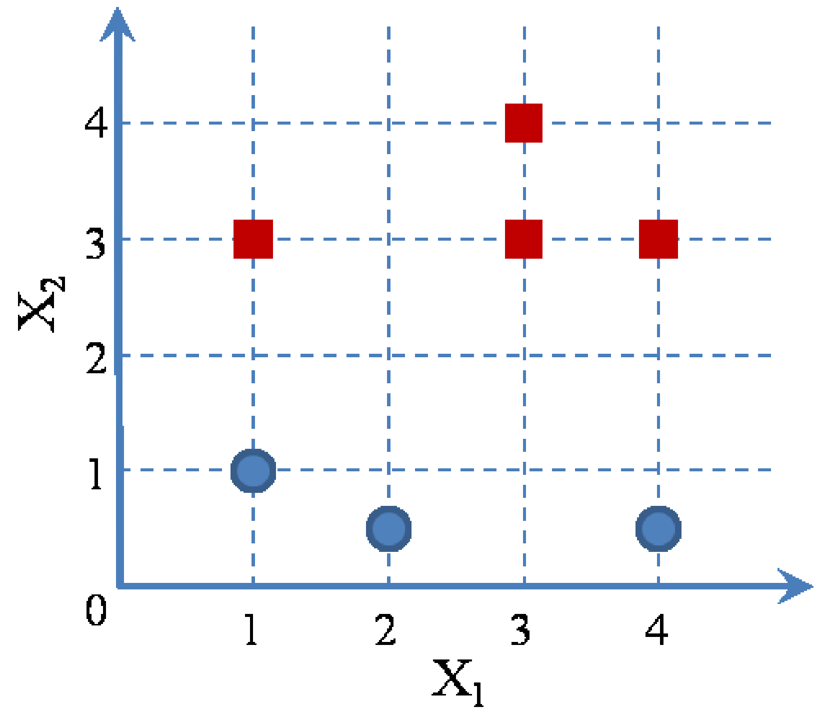
CZ4041/CE4041: Machine Learning

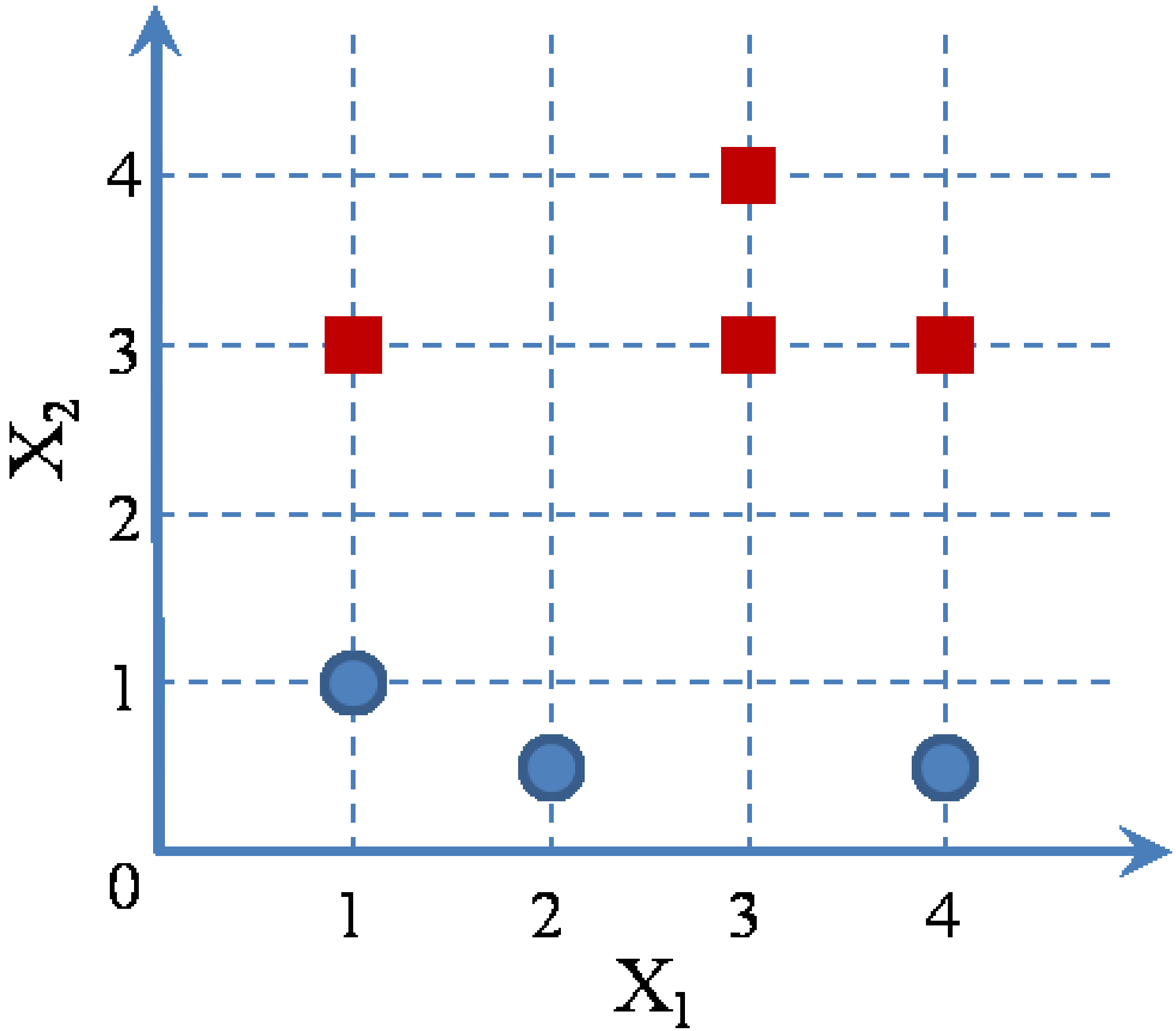
Week 8: Support Vector Machines and Regularized Linear Regression

Question 1

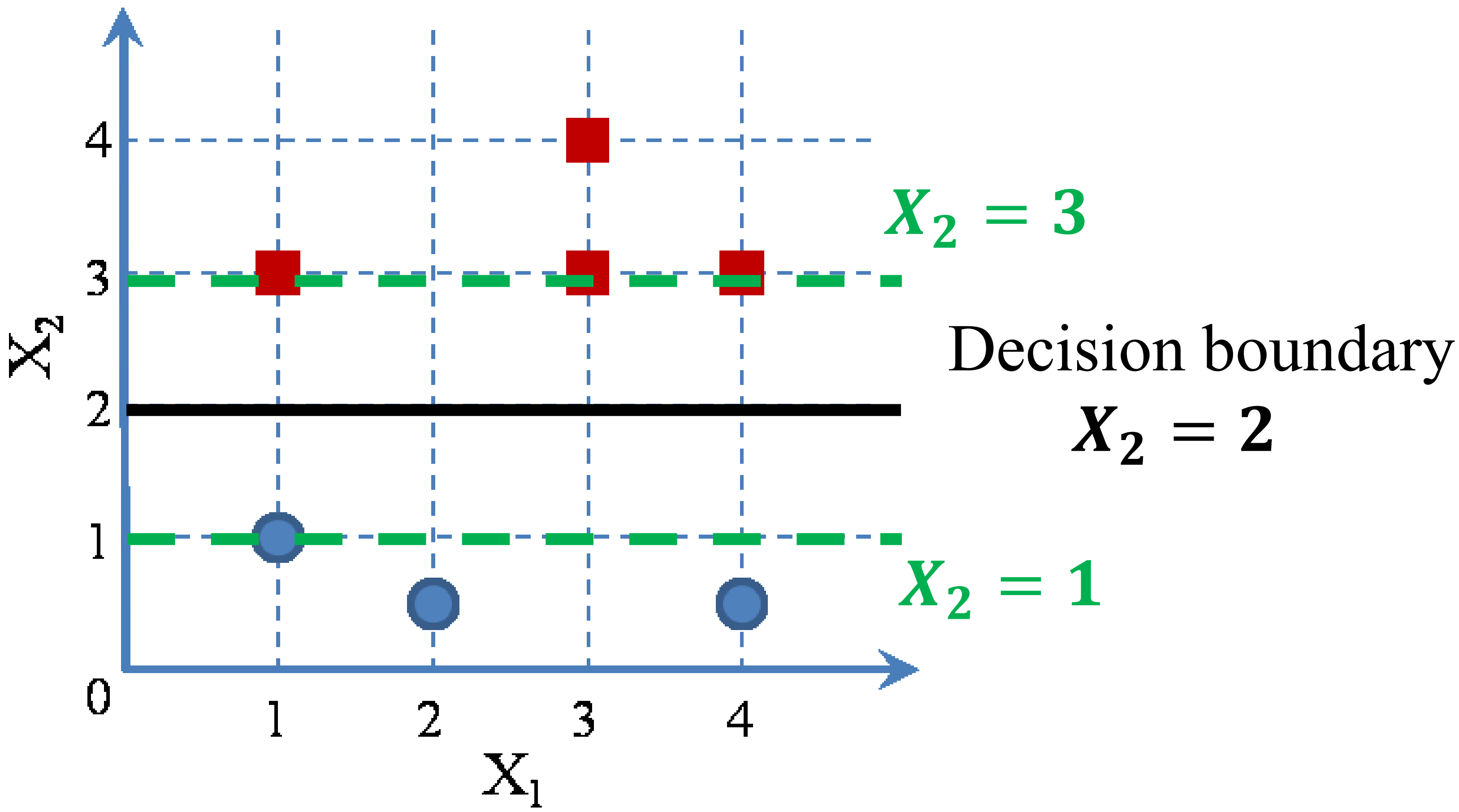
Consider a 2-dimensional dataset for two-class classification by SVM, as shown in Figure 1, where the red “square” and blue “circle” denote the positive and negative classes respectively. Is this dataset separable by a linear SVM classifier? If no, why? If yes, what is the decision boundary of the linear SVM? And what are the pair of parallel hyperplanes associated with the decision boundary? (No need to provide proofs)



Question 1



Question 1



Question 2

The two parallel hyperplanes passing the closest circle(s) and square(s) can be written as

$$\boldsymbol{w} \cdot \boldsymbol{x} + b = k, \text{where } k > 0$$

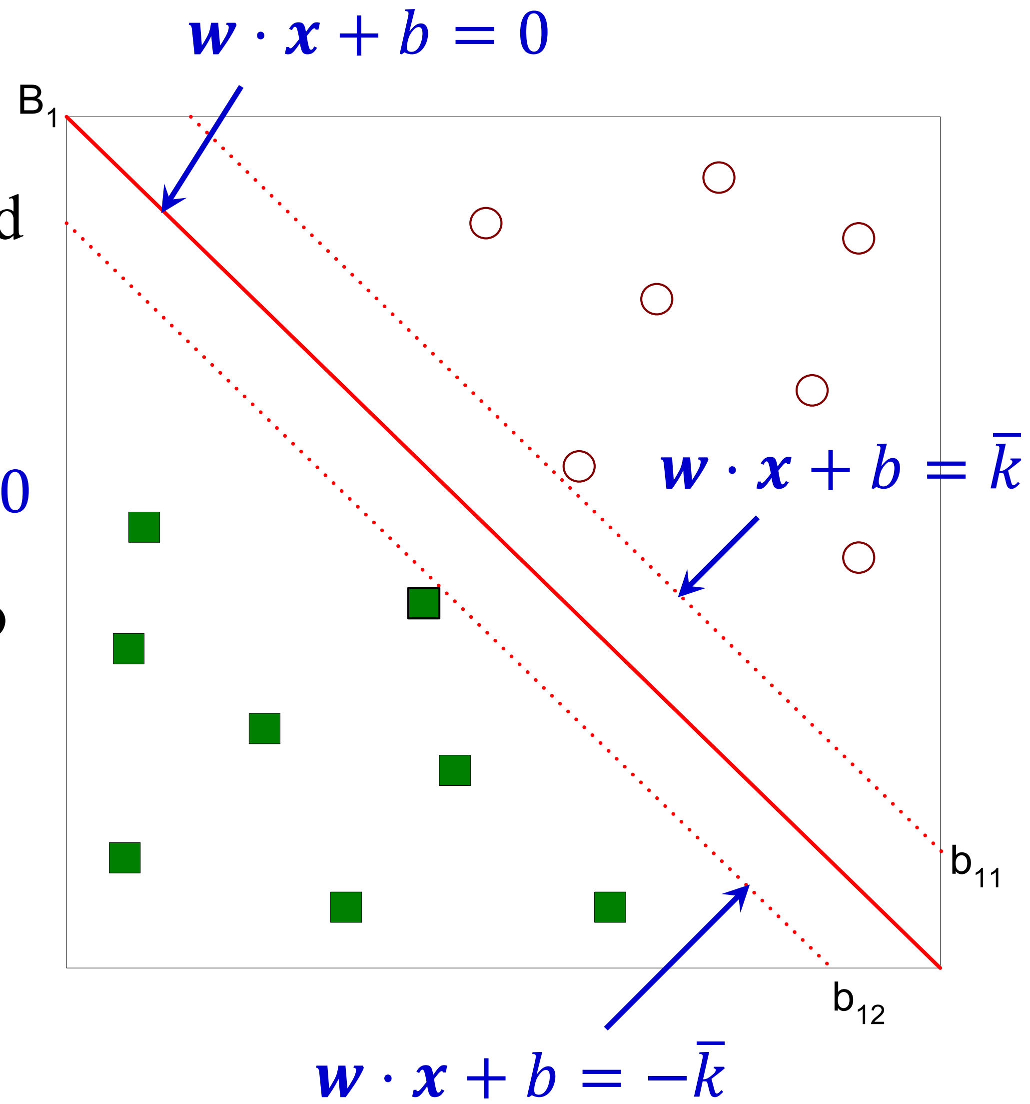
$$\boldsymbol{w} \cdot \boldsymbol{x} + b = k', \text{where } k' < 0$$

It can be shown that, these two parallel hyperplanes can be further rewritten as

$$\boldsymbol{w} \cdot \boldsymbol{x} + b = \bar{k}$$

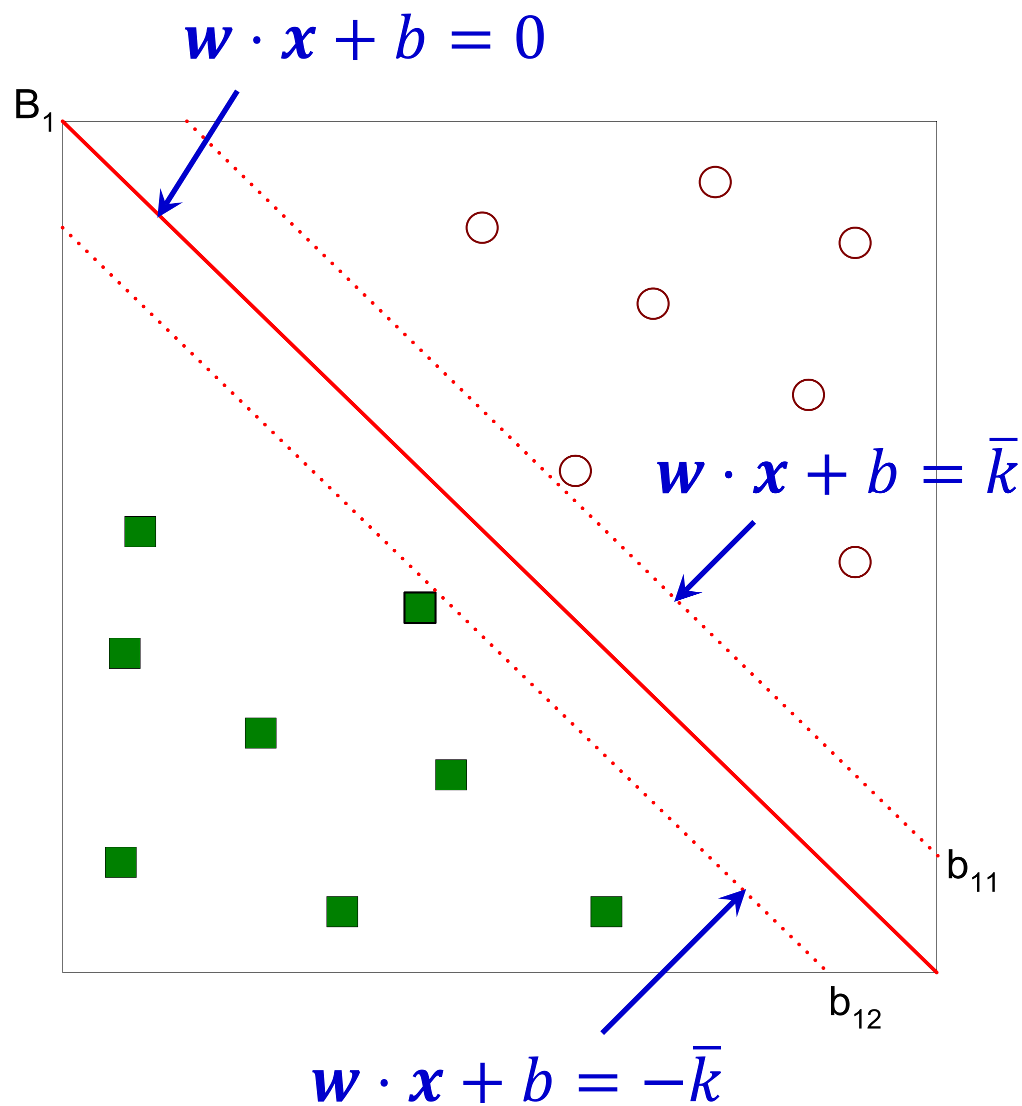
$$\boldsymbol{w} \cdot \boldsymbol{x} + b = -\bar{k}$$

where $\bar{k} > 0$



Question 2

- $\boldsymbol{w}$ determines the orientation (slope) of the decision boundary.
- The support vectors determine how the decision boundary moves in parallel motion.
- Together, they determine b.

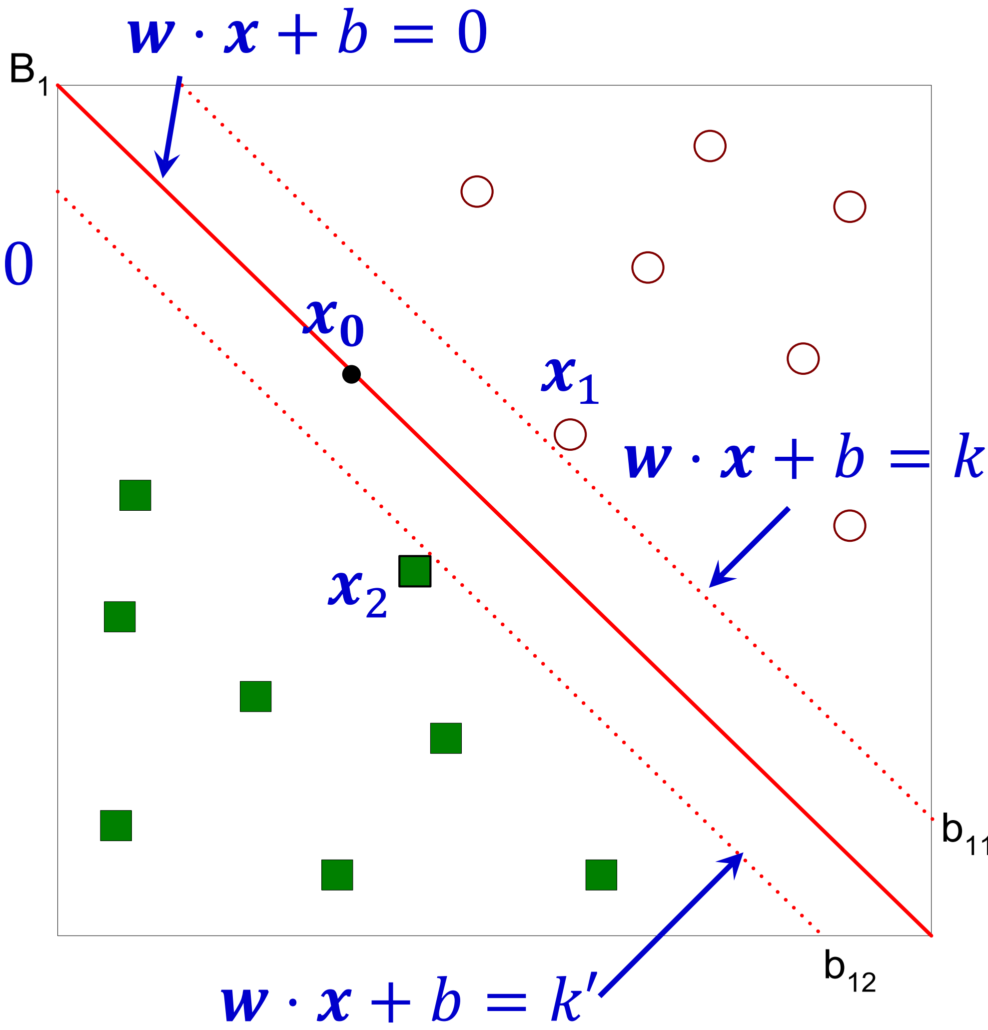


Question 2 (cont.)

b_{11}: $\boldsymbol{w} \cdot \boldsymbol{x}_1 + b = k$, where $k > 0$

b_{12}: $\boldsymbol{w} \cdot \boldsymbol{x}_2 + b = k'$, where $k' < 0$

Given two support vectors (or two points on b_{11} and b_{12} repsectively), I can choose b such that $k = -k'$

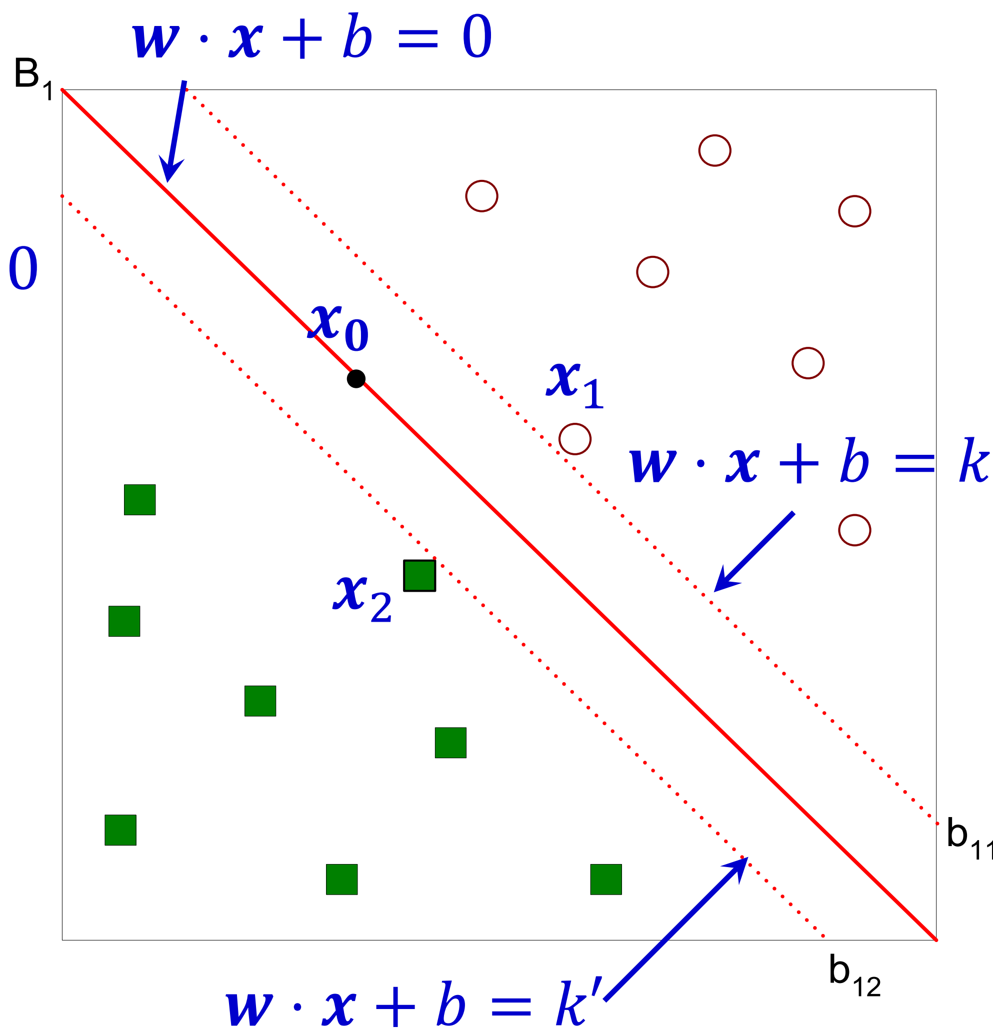


Question 2 (cont.)

b_{11}: $\boldsymbol{w} \cdot \boldsymbol{x}_1 + b = k$, where $k > 0$

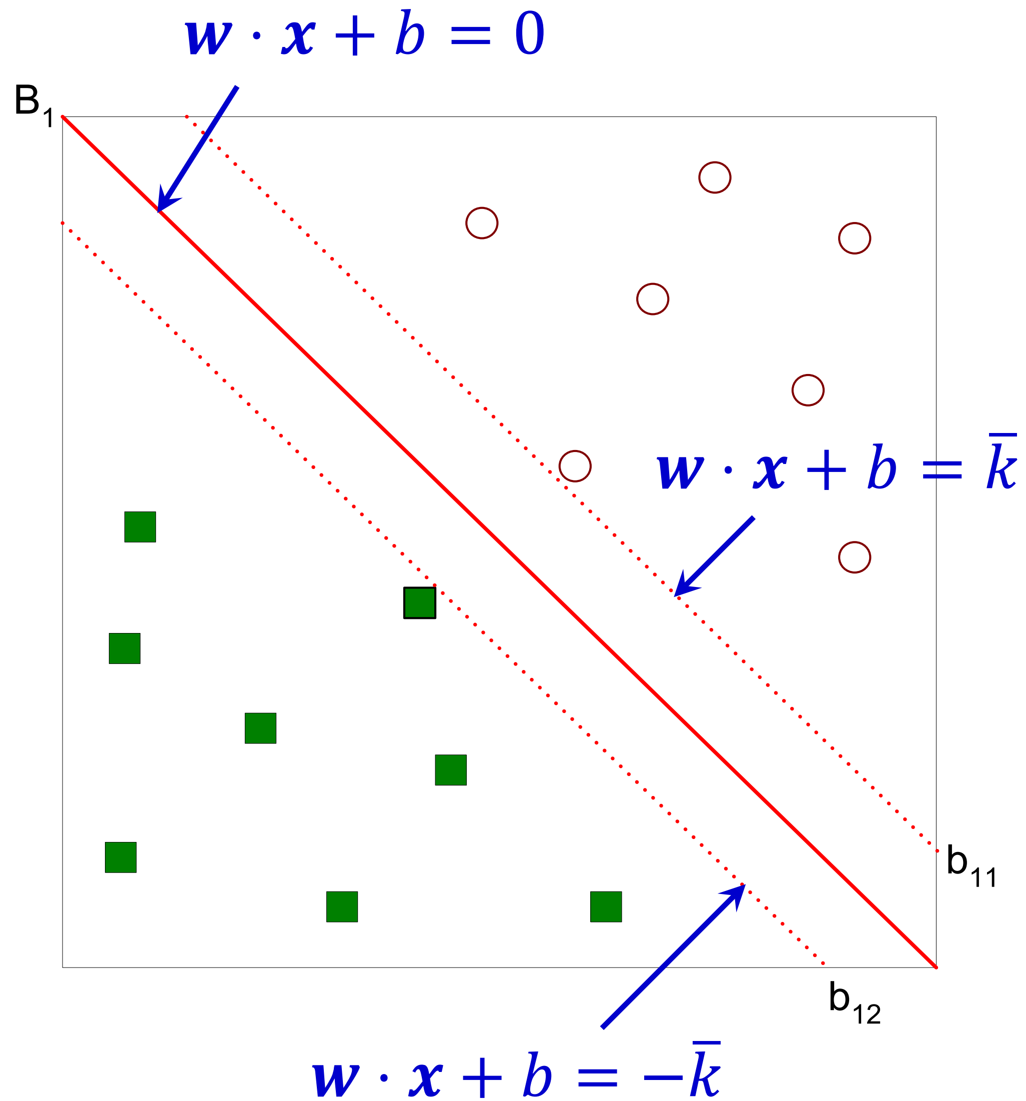
b_{12}: $\boldsymbol{w} \cdot \boldsymbol{x}_2 + b = k'$, where $k' < 0$

$$\boldsymbol{w} \cdot \boldsymbol{x}_1 + b = -(\boldsymbol{w} \cdot \boldsymbol{x}_2 + b)$$
$$2b = -\boldsymbol{w} \cdot \boldsymbol{x}_1 - \boldsymbol{w} \cdot \boldsymbol{x}_2$$
$$b = -\frac{1}{2}\boldsymbol{w} \cdot (\boldsymbol{x}_1 + \boldsymbol{x}_2)$$



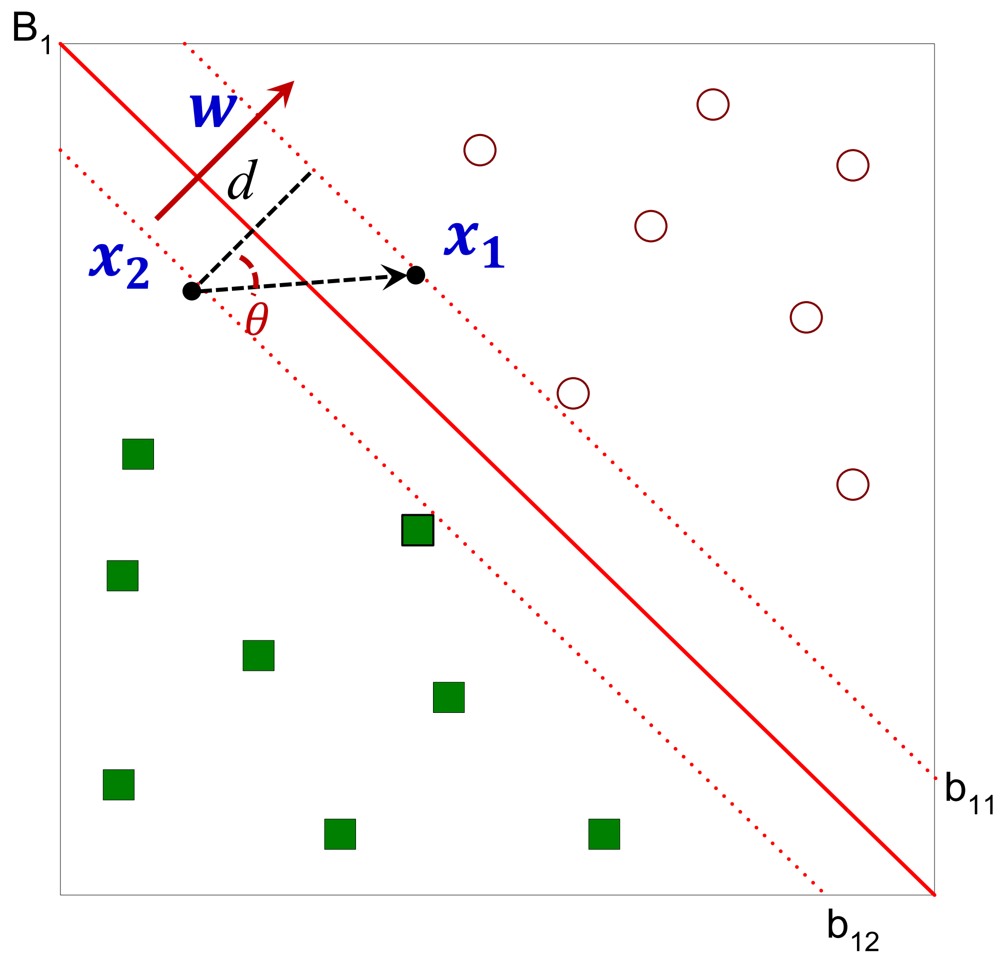
Question 3

- Two support vectors: (2, 3) and (-1, 4) from the two classes, respectively.
- Decision boundary: $\boldsymbol{w} \cdot \boldsymbol{x} + b = 3x_1 + x_2 + b = 0$
- What is the margin?



Question 3

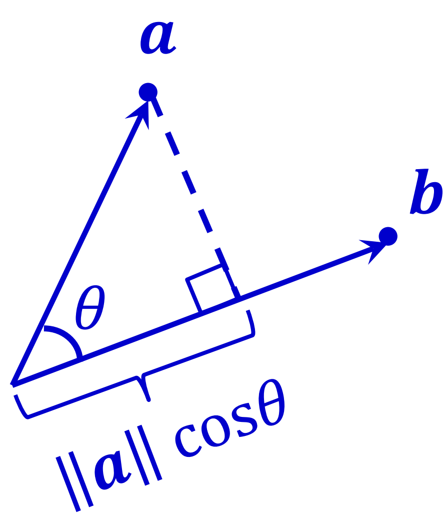
- From the lecture, $\boldsymbol{w}$ is orthogonal to the decision boundary.
- All we need is to find the length of projection of the vector $(\boldsymbol{x}_1 - \boldsymbol{x}_2)$ onto the direction of $\boldsymbol{w}$.



Review: Geometry of Inner Products

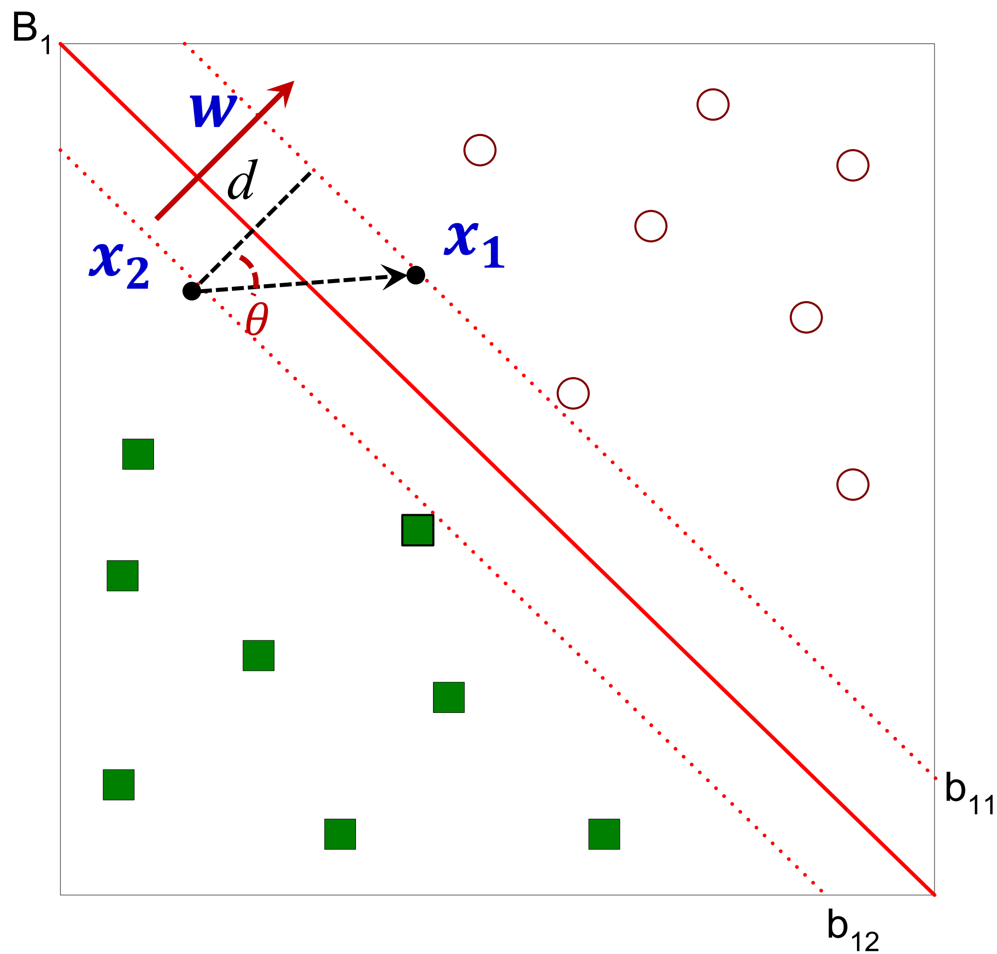
$\boldsymbol{a} \cdot \boldsymbol{b} = \|\boldsymbol{a}\| \|\boldsymbol{b}\| \cos\theta$, where θ is the angle between $\boldsymbol{a}$ and $\boldsymbol{b}$

$\frac{\boldsymbol{a} \cdot \boldsymbol{b}}{\|\boldsymbol{b}\|} = \|\boldsymbol{a}\| \cos\theta$ is the length of the projection of $\boldsymbol{a}$ onto $\boldsymbol{b}$



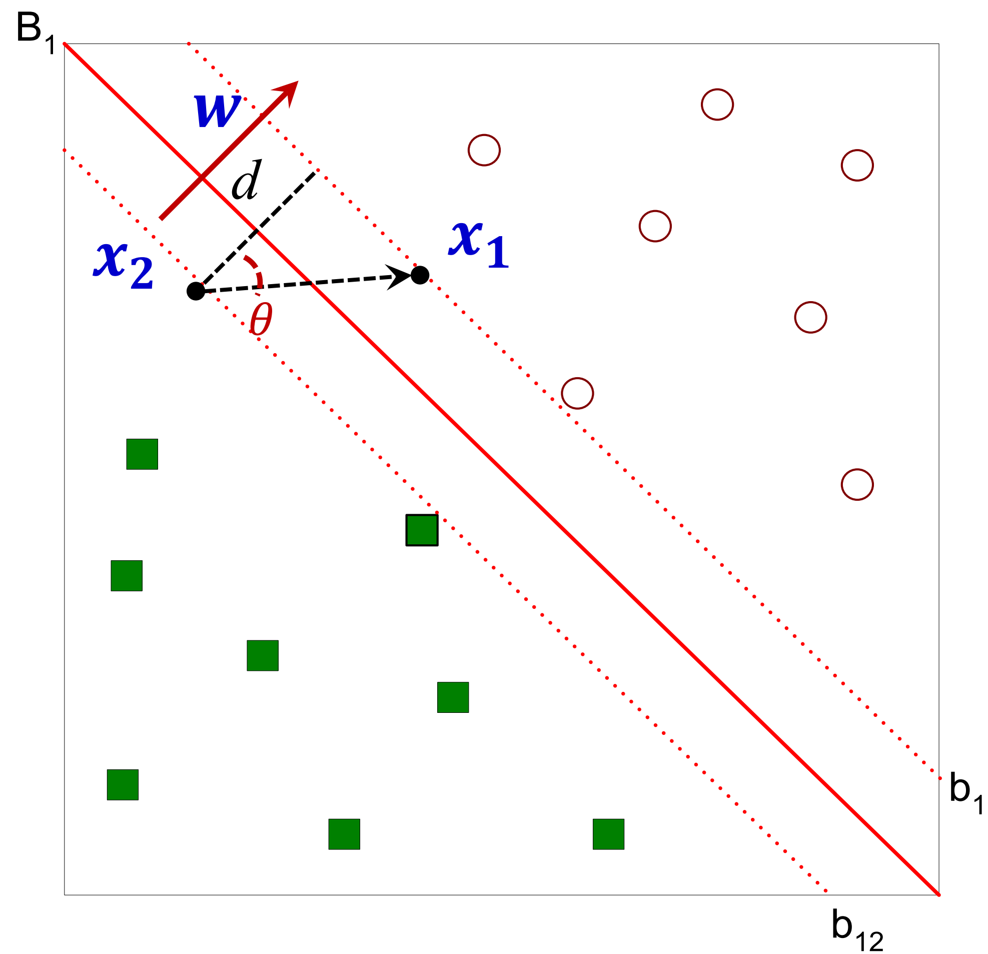
Question 3

- We know that
  $$\boldsymbol{w} \cdot (\boldsymbol{x}_1 - \boldsymbol{x}_2)$$
  $$= \|\boldsymbol{w}\|\ \|\boldsymbol{x}_1 - \boldsymbol{x}_2\| \cos(\theta)$$
  $$= \|\boldsymbol{w}\|\ d \ \underline{\text{or}}\ -\|\boldsymbol{w}\|\ d$$
- Thus, $d = \left|\frac{\boldsymbol{w}\cdot(\boldsymbol{x}_1 - \boldsymbol{x}_2)}{\|\boldsymbol{w}\|}\right|$
- $\boldsymbol{x}_1 - \boldsymbol{x}_2 = (2,3) - (-1,4)$
  $$= (3, -1)$$



Question 3

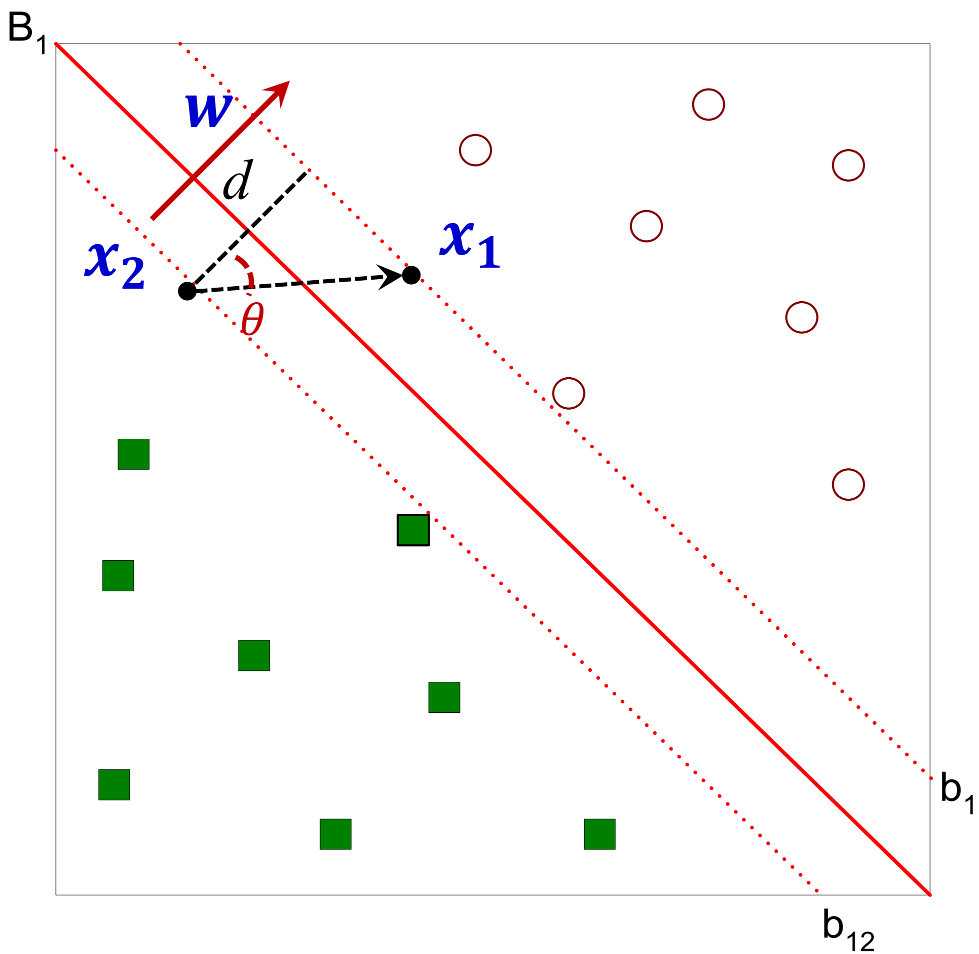
- Thus, $d = \left|\frac{\boldsymbol{w}\cdot(\boldsymbol{x}_1 - \boldsymbol{x}_2)}{\|\boldsymbol{w}\|}\right|$
- $\boldsymbol{x}_1 - \boldsymbol{x}_2 = (2,3) - (-1,4)$
  $= (3, -1)$
- We know $\boldsymbol{w} = (3,1)$
- $|\boldsymbol{w} \cdot (\boldsymbol{x}_1 - \boldsymbol{x}_2)| = |3 \times 3 - 1 \times 1| = 8$
- $\|\boldsymbol{w}\| = \sqrt{10}$
- $d =$



Question 3

- Why can't we directly use the equation $||\boldsymbol{w}||\ d = 2$?
- That equation is only valid when $\boldsymbol{w}$ is properly rescaled.
- That is, when

$$\boldsymbol{w} \cdot \boldsymbol{x}_1 + b = 1$$
$$\boldsymbol{w} \cdot \boldsymbol{x}_2 + b = -1$$



Q4: Regularized Linear Regression

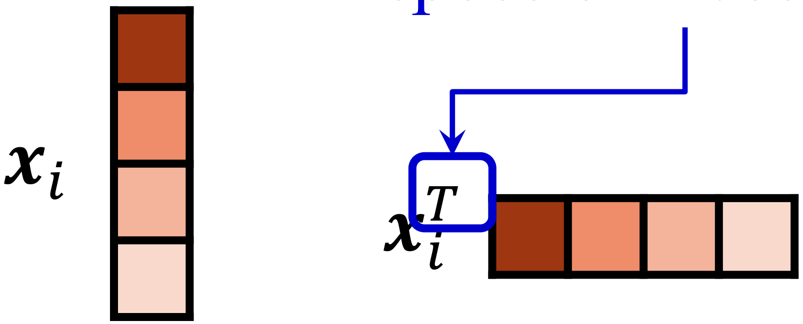
- To solve the unconstrained minimization problem, we can set the derivative of $\mathcal{L}(\boldsymbol{w}) + \frac{\lambda}{2}\|\boldsymbol{w}\|_2^2$ w.r.t. $\boldsymbol{w}$ to **zero**

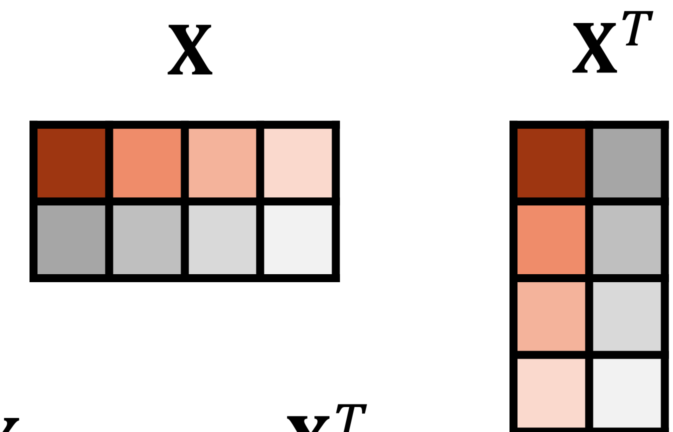
$$\frac{\partial\left(\mathcal{L}(\boldsymbol{w}) + \frac{\lambda}{2}\|\boldsymbol{w}\|_2^2\right)}{\partial \boldsymbol{w}} = \frac{\partial\left(\frac{1}{2}\sum_{i=1}^{N}(\boldsymbol{w}\cdot\boldsymbol{x}_i - y_i)^2 + \frac{\lambda}{2}\boldsymbol{w}\cdot\boldsymbol{w}\right)}{\partial \boldsymbol{w}} = \mathbf{0}$$

- We can obtain a closed-form solution

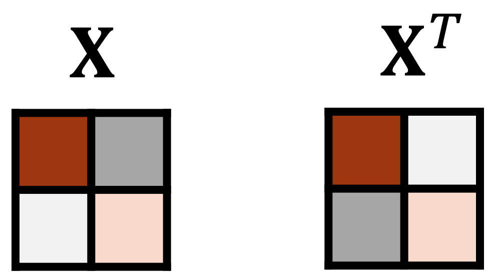
Some Concepts: Review

Transpose of a vector/matrix

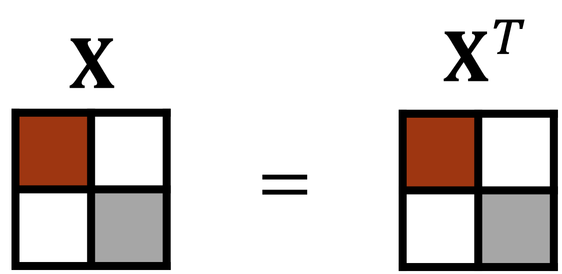




If $\mathbf{X}$ is a square matrix, then its numbers of rows and columns are the same



If $\mathbf{X}$ is a symmetric matrix, then it is square and $\mathbf{X}^T = \mathbf{X}$

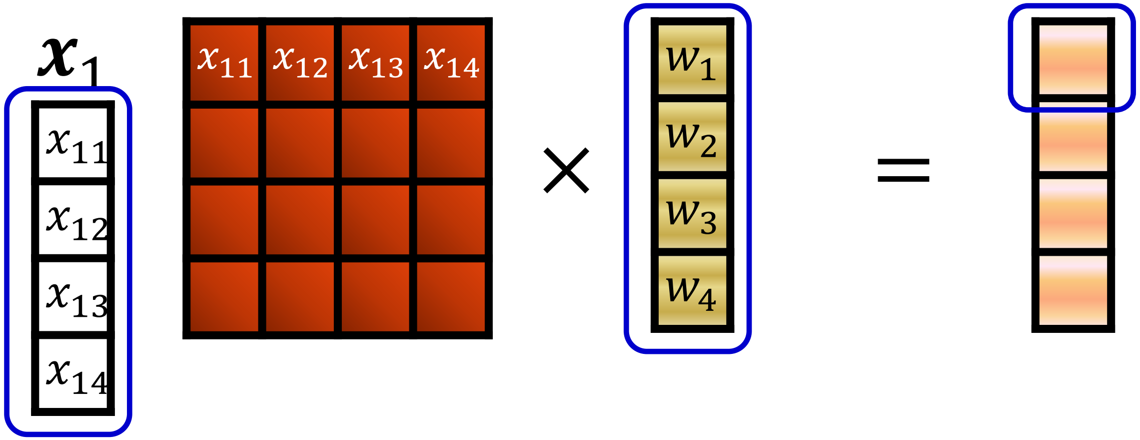


Some Concepts: Review

- The transpose of $\mathbf{XY}$:
$$(\mathbf{XY})^T = \mathbf{Y}^T\mathbf{X}^T$$
- The transpose of $\mathbf{X}\boldsymbol{w}$
$$(\mathbf{X}\boldsymbol{w})^T = \boldsymbol{w}^T\mathbf{X}^T$$
- The transpose of $\boldsymbol{x}^T\boldsymbol{w}$
$$(\boldsymbol{x}^T\boldsymbol{w})^T = \boldsymbol{w}^T\boldsymbol{x}$$
- The transpose of a scalar is the scalar itself
$$x^T = x$$

Some Concepts: Review (cont.)

$\mathbf{X} \in \mathbb{R}^{d\times d}$ $\boldsymbol{w} \in \mathbb{R}^{d\times 1}$ $\boldsymbol{z} = \mathbf{X}\boldsymbol{w} \in \mathbb{R}^{d\times 1}$



$$\boldsymbol{x}_1{}^T\boldsymbol{w} = \sum_{j=1}^{d} x_{1j}w_j = z_1$$

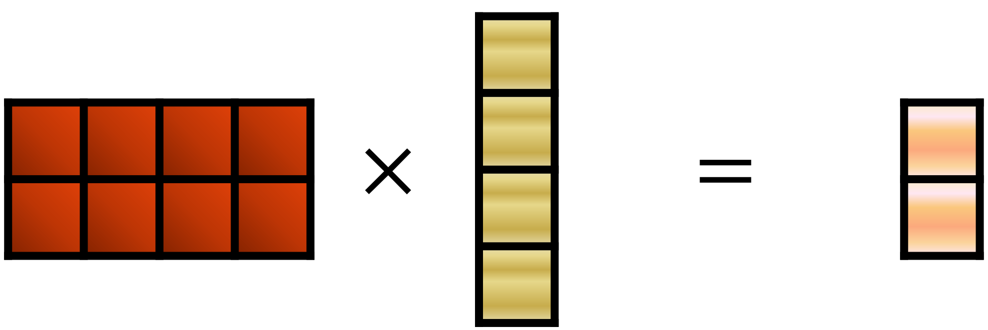
$$\|$$

$$\boldsymbol{w}^T\boldsymbol{x}_1$$

$$\|$$

$$\boldsymbol{w}\cdot\boldsymbol{x}_1 = \boldsymbol{x}_1\cdot\boldsymbol{w}$$

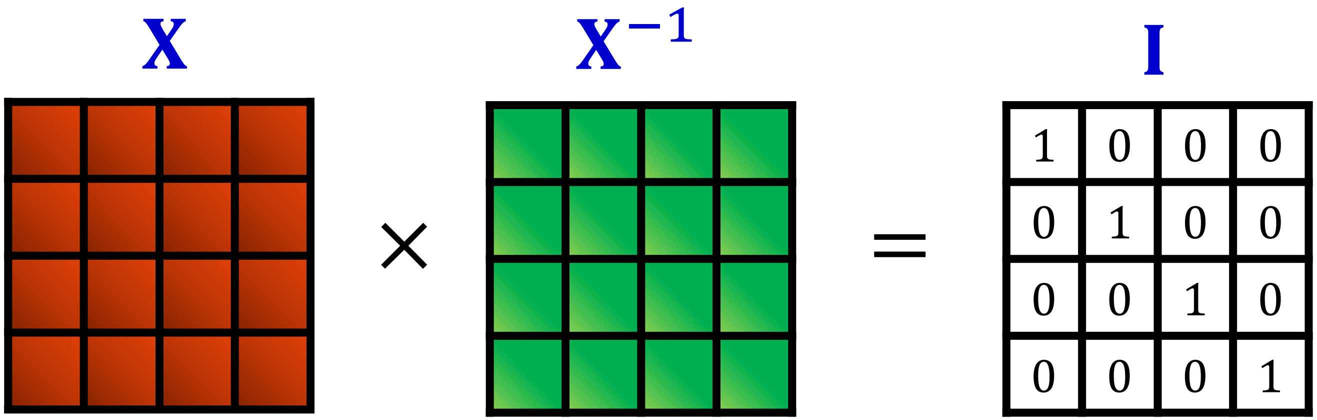
$\mathbf{X} \in \mathbb{R}^{d\times k}$ $\boldsymbol{w} \in \mathbb{R}^{k\times 1}$ $\boldsymbol{z} = \mathbf{X}\boldsymbol{w} \in \mathbb{R}^{d\times 1}$



Some Concepts: Review (cont.)

- For a square matrix **X**, if **X** is invertible, then

$$\mathbf{X}\mathbf{X}^{-1} = \boxed{\mathbf{I}} \longleftarrow \text{Identity matrix}$$

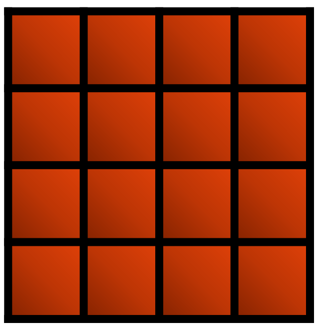


Some Concepts: Review (cont.)

- Any vector (or matrix) $\boldsymbol{x}$ (or $\mathbf{X}$) times identity matrix $\mathbf{I}$ equals to the vector (or matrix) itself

$$\mathbf{I}\boldsymbol{x} = \boldsymbol{x} \ (\boldsymbol{x}^T\mathbf{I} = \boldsymbol{x}^T) \quad \text{OR} \quad \mathbf{XI} = \mathbf{X} \ (\mathbf{IX} = \mathbf{X})$$

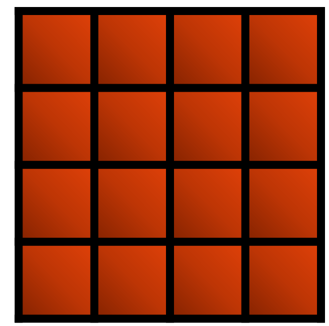
X



$\times$ **I**

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

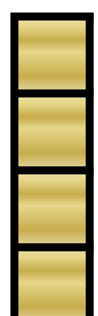
$=$ **X**



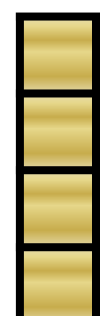
I

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$\times$ $\boldsymbol{x}$



$=$ $\boldsymbol{x}$



Question 4

- Denote by $\mathbf{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_N)^T$

$$\mathbf{X} = \begin{array}{c} \boldsymbol{x}_1 \qquad\qquad\qquad \\ \begin{pmatrix} x_{10} & \cdots & x_{N0} \\ \vdots & \ddots & \vdots \\ x_{1d} & \cdots & x_{Nd} \end{pmatrix}^T \end{array} = \begin{array}{cc} \boldsymbol{x}_1 & \\ & \begin{pmatrix} x_{10} & \cdots & x_{1d} \\ \vdots & \ddots & \vdots \\ x_{N0} & \cdots & x_{Nd} \end{pmatrix} \end{array}$$

- And by $\boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}$

How to get this closed-form solution?

- The closed-form solution for $\boldsymbol{w}$:

$$\boldsymbol{w} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\boldsymbol{y}$$

$$\frac{\partial\left(\frac{1}{2}\sum_{i=1}^{N}(\boldsymbol{w}\cdot\boldsymbol{x}_i - y_i)^2 + \frac{\lambda}{2}\boldsymbol{w}\cdot\boldsymbol{w}\right)}{\partial\boldsymbol{w}} = \mathbf{0}$$

$$\frac{\partial\,\frac{1}{2}\sum_{i=1}^{N}(\boldsymbol{w}\cdot\boldsymbol{x}_i - y_i)^2}{\partial\boldsymbol{w}}$$

$$= \frac{1}{2}\sum_{i=1}^{N} 2(\boldsymbol{w}\cdot\boldsymbol{x}_i - y_i)\frac{\partial(\boldsymbol{w}\cdot\boldsymbol{x}_i - y_i)}{\partial\boldsymbol{w}}$$

$$= \sum_{i=1}^{N}(\boldsymbol{w}\cdot\boldsymbol{x}_i - y_i)\,\boldsymbol{x}_i$$

$$\frac{\partial\left(\frac{\lambda}{2}\boldsymbol{w}\cdot\boldsymbol{w}\right)}{\partial\boldsymbol{w}}$$

$$= \frac{\lambda}{2}\times 2\boldsymbol{w} = \lambda\boldsymbol{w}$$

$$\sum_{i=1}^{N}(\boldsymbol{w}\cdot\boldsymbol{x}_i - y_i)\boldsymbol{x}_i + \lambda\boldsymbol{w} = \mathbf{0}$$

$$\sum_{i=1}^{N}(\boldsymbol{w}\cdot\boldsymbol{x}_i)\boldsymbol{x}_i - \sum_{i=1}^{N} y_i\boldsymbol{x}_i + \lambda\boldsymbol{w} = \mathbf{0}$$

$$\sum_{i=1}^{N} \boxed{(\boldsymbol{w} \cdot \boldsymbol{x}_i)\boldsymbol{x}_i} - \sum_{i=1}^{N} y_i \boldsymbol{x}_i + \lambda \boldsymbol{w} = \boldsymbol{0}$$

$$\boldsymbol{x}_i(\boldsymbol{w} \cdot \boldsymbol{x}_i) = \boldsymbol{x}_i(\boldsymbol{x}_i^T \boldsymbol{w})$$

$$= (\boldsymbol{x}_i \boldsymbol{x}_i^T)\boldsymbol{w}$$

Identity matrix

$$\left(\sum_{i=1}^{N} \boxed{(\boldsymbol{x}_i {\boldsymbol{x}_i}^T)}\right) \boldsymbol{w} - \sum_{i=1}^{N} y_i \boldsymbol{x}_i + \lambda \boxed{\mathbf{I}}\, \boldsymbol{w} = \boldsymbol{0}$$

A matrix of $(d+1)$ by $(d+1)$ size, where $\boldsymbol{x}_i$ is a column vector of $(d+1)$ dimensions

$$\boldsymbol{x}_i \qquad\qquad \boldsymbol{x}_i^T$$

$$\begin{bmatrix} x_{i0} \\ x_{i1} \\ \cdots \\ x_{id} \end{bmatrix} \qquad \begin{bmatrix} x_{i0} & x_{i1} & \dots & x_{id} \end{bmatrix}$$

$$\downarrow$$

$$\boldsymbol{x}_i \boldsymbol{x}_i{}^T = \begin{pmatrix} x_{i0} \times x_{i0} & \cdots & x_{i0} \times x_{id} \\ \vdots & \ddots & \vdots \\ x_{id} \times x_{i0} & \cdots & x_{id} \times x_{id} \end{pmatrix}$$

$$\boldsymbol{x}_i \boldsymbol{x}_i{}^T = \begin{pmatrix} x_{i0} \times x_{i0} & \cdots & x_{i0} \times x_{id} \\ \vdots & \ddots & \vdots \\ x_{id} \times x_{i0} & \cdots & x_{id} \times x_{id} \end{pmatrix}$$

$$\Downarrow$$

$$\sum_{i=1}^{N} (\boldsymbol{x}_i \boldsymbol{x}_i{}^T) = \begin{pmatrix} \sum_{i=1}^{N} x_{i0} \times x_{i0} & \cdots & \sum_{i=1}^{N} x_{i0} \times x_{id} \\ \vdots & \ddots & \vdots \\ \sum_{i=1}^{N} x_{id} \times x_{i0} & \cdots & \sum_{i=1}^{N} x_{id} \times x_{id} \end{pmatrix}$$

$$\sum_{i=1}^{N}(\boldsymbol{x}_i\boldsymbol{x}_i{}^T) = \begin{pmatrix} \sum_{i=1}^{N} x_{i0} \times x_{i0} & \cdots & \sum_{i=1}^{N} x_{i0} \times x_{id} \\ \vdots & \ddots & \vdots \\ \sum_{i=1}^{N} x_{id} \times x_{i0} & \cdots & \sum_{i=1}^{N} x_{id} \times x_{id} \end{pmatrix}$$

$$\mathbf{X} = \begin{array}{c} \begin{array}{lll} \boldsymbol{x}_1 & & \phantom{x_{N0}} \end{array} \\ \begin{pmatrix} x_{10} & \cdots & x_{N0} \\ \vdots & \ddots & \vdots \\ x_{1d} & \cdots & x_{Nd} \end{pmatrix}^T \end{array} = \begin{array}{cc} \boldsymbol{x}_1 & \\ & \begin{pmatrix} x_{10} & \cdots & x_{1d} \\ \vdots & \ddots & \vdots \\ x_{N0} & \cdots & x_{Nd} \end{pmatrix} \end{array}$$

$$\sum_{i=1}^{N}(\boldsymbol{x}_i\boldsymbol{x}_i{}^T) = \mathbf{X}^T\mathbf{X}$$

Therefore $$\left(\sum_{i=1}^{N}(\boldsymbol{x}_i\boldsymbol{x}_i{}^T)\right)\boldsymbol{w} - \sum_{i=1}^{N} y_i\boldsymbol{x}_i + \lambda\mathbf{I}\,\boldsymbol{w} = \mathbf{0}$$

$$(\mathbf{X}^T\mathbf{X})\boldsymbol{w} - \mathbf{X}^T\boldsymbol{y} + \lambda\mathbf{I}\,\boldsymbol{w} = \mathbf{0}$$

Always invertible as long as λ is positive

$$\boxed{(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})}\boldsymbol{w} = \mathbf{X}^T\boldsymbol{y}$$

$$(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})\boldsymbol{w} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\boldsymbol{y}$$

$\mathbf{I}\boldsymbol{w} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\boldsymbol{y}$

$$\boldsymbol{w} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\boldsymbol{y}$$

Thank you!