香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

COMP 5212
Machine Learning
Lecture 5

Support Vector Machine

Junxian He
Feb 26, 2026

Attendance Quiz APP Download

HKUST iLearn

The Hong Kong University of Science and Technology

10K+ Downloads | Everyone

Install | Share | Add to wishlist

Canvas
This will open the 'Canvas Student' app which provides an easy access to the online content of your courses at HKUST - watch videos, post to discussions, submit quizzes, etc.

SFQ
Allows you to complete the Student Feedback Questionnaire for all your courses at HKUST on the move.

iPRS
Enables you to quickly respond to questions or polls created by your instructor in class.

Recap: Kernel Trick

$$\theta = \sum_{i=1}^{n} \beta_i \phi(x^{(i)}) \qquad \beta_i \in R$$

Recap: Kernel Trick

parameter → feature map

$$\theta = \sum_{i=1}^{n} \beta_i \phi(x^{(i)}) \qquad \beta_i \in R$$

$$\beta_i := \beta_i + \alpha \left(y^{(i)} - \sum_{j=1}^{n} \beta_j \phi(x^{(j)})^T \phi(x^{(i)}) \right)$$

Recap: Kernel Trick

$$\theta = \sum_{i=1}^{n} \beta_i \phi(x^{(i)}) \qquad \beta_i \in R$$

$\langle \phi(x), \phi(z) \rangle$

$$\beta_i := \beta_i + \alpha \left(y^{(i)} - \sum_{j=1}^{n} \beta_j \phi(x^{(j)})^T \phi(x^{(i)}) \right)$$

Kernel $K(x, z) \quad \mathcal{X} \times \mathcal{X} \to \mathbb{R} \quad \mathcal{X}$ is the space of the input

$$K(x, z) \triangleq \langle \phi(x), \phi(z) \rangle$$

Recap: Kernel Trick

Recap: Kernel Trick

- Compute $K(\phi(x^{(i)}), \phi(x^{(j)})) = < \phi(x^{(i)}), \phi(x^{(j)}) >$ for all i, j

Recap: Kernel Trick

- Compute $K(\phi(x^{(i)}), \phi(x^{(j)})) = < \phi(x^{(i)}), \phi(x^{(j)}) >$ for all i, j
- Loop $$\beta_i := \beta_i + \alpha \left(y^{(i)} - \sum_{j=1}^{n} \beta_j K(x^{(i)}, x^{(j)}) \right) \quad \forall i \in \{1, \ldots, n\}$$

$\theta^T \phi(x)$

$O(p)$

Recap: Kernel Trick

- Compute $K(\phi(x^{(i)}), \phi(x^{(j)})) = < \phi(x^{(i)}), \phi(x^{(j)}) >$ for all i, j
- Loop $\beta_i := \beta_i + \alpha \left(y^{(i)} - \sum_{j=1}^{n} \beta_j K(x^{(i)}, x^{(j)}) \right) \quad \forall i \in \{1, \ldots, n\}$

Recall that n is the number of data samples

Recap: Kernel Trick

- Compute $K(\phi(x^{(i)}), \phi(x^{(j)})) = <\phi(x^{(i)}), \phi(x^{(j)})>$ for all i, j
- Loop $$\beta_i := \beta_i + \alpha\left(y^{(i)} - \sum_{j=1}^{n} \beta_j K(x^{(i)}, x^{(j)})\right) \quad \forall i \in \{1, \ldots, n\}$$

Recall that n is the number of data samples

- Inference: $$\theta^T \phi(x) = \sum_{i=1}^{n} \beta_i \phi(x^{(i)})^T \phi(x) = \sum_{i=1}^{n} \beta_i K(x^{(i)}, x)$$

Recap: Kernel Trick

- Compute $K(\phi(x^{(i)}), \phi(x^{(j)})) = < \phi(x^{(i)}), \phi(x^{(j)}) >$ for all i, j
- Loop $$\beta_i := \beta_i + \alpha \left(y^{(i)} - \sum_{j=1}^{n} \beta_j K(x^{(i)}, x^{(j)}) \right) \quad \forall i \in \{1, \ldots, n\}$$

Recall that n is the number of data samples

- Inference: $$\theta^T \phi(x) = \sum_{i=1}^{n} \beta_i \phi(x^{(i)})^T \phi(x) = \sum_{i=1}^{n} \beta_i K(x^{(i)}, x)$$

The Kernel function is all we need for training and inference!

Recap: Implicit Feature Map

- Explicit Feature Map: first define feature map $\phi(x)$, then compute the Kernel according to $\phi(x)$

- Implicit Feature Map: first define the Kernel Function K(), without knowing what the feature map is

Recap: Implicit Feature Map (Example)

$$K(x, z) = (x^T z)^2 \qquad x, z \in \mathbb{R}^d$$

Recap: Implicit Feature Map (Example)

$$K(x, z) = (x^T z)^2 \qquad x, z \in \mathbb{R}^d$$

What is the feature map to make K a valid kernel function?

Recap: Implicit Feature Map (Example)

$\phi(x)^T \phi(z)$

$$K(x, z) = (x^T z)^2 \qquad x, z \in \mathbb{R}^d$$

What is the feature map to make K a valid kernel function?

$$\begin{aligned} K(x,z) &= \left(\sum_{i=1}^{d} x_i z_i\right)\left(\sum_{j=1}^{d} x_j z_j\right) \\ &= \sum_{i=1}^{d}\sum_{j=1}^{d} x_i x_j z_i z_j \\ &= \sum_{i,j=1}^{d} (x_i x_j)(z_i z_j) \end{aligned}$$

Recap: Implicit Feature Map (Example)

$$K(x, z) = (x^T z)^2 \qquad x, z \in \mathbb{R}^d$$

What is the feature map to make K a valid kernel function?

$$\begin{aligned} K(x,z) &= \left(\sum_{i=1}^{d} x_i z_i\right)\left(\sum_{j=1}^{d} x_j z_j\right) \\ &= \sum_{i=1}^{d}\sum_{j=1}^{d} x_i x_j z_i z_j \\ &= \sum_{i,j=1}^{d} (x_i x_j)(z_i z_j) \end{aligned}$$

$$\phi(x) = \begin{bmatrix} x_1 x_1 \\ x_1 x_2 \\ x_1 x_3 \\ x_2 x_1 \\ x_2 x_2 \\ x_2 x_3 \\ x_3 x_1 \\ x_3 x_2 \\ x_3 x_3 \end{bmatrix}$$

Recap: Implicit Feature Map (Example)

$$K(x, z) = (x^T z)^2 \qquad x, z \in \mathbb{R}^d$$

What is the feature map to make K a valid kernel function?

$$\begin{aligned} K(x,z) &= \left(\sum_{i=1}^{d} x_i z_i\right)\left(\sum_{j=1}^{d} x_j z_j\right) \\ &= \sum_{i=1}^{d}\sum_{j=1}^{d} x_i x_j z_i z_j \\ &= \sum_{i,j=1}^{d} (x_i x_j)(z_i z_j) \end{aligned}$$

$$\phi(x) = \begin{bmatrix} x_1x_1 \\ x_1x_2 \\ x_1x_3 \\ x_2x_1 \\ x_2x_2 \\ x_2x_3 \\ x_3x_1 \\ x_3x_2 \\ x_3x_3 \end{bmatrix}$$

Requires O(d^2) compute for feature mapping

Recap: Implicit Feature Map (Example)

$$K(x, z) = (x^T z)^2 \qquad x, z \in \mathbb{R}^d$$

What is the feature map to make K a valid kernel function?

$$\begin{aligned} K(x,z) &= \left(\sum_{i=1}^{d} x_i z_i\right)\left(\sum_{j=1}^{d} x_j z_j\right) \\ &= \sum_{i=1}^{d}\sum_{j=1}^{d} x_i x_j z_i z_j \\ &= \sum_{i,j=1}^{d} (x_i x_j)(z_i z_j) \end{aligned}$$

$$\phi(x) = \begin{bmatrix} x_1x_1 \\ x_1x_2 \\ x_1x_3 \\ x_2x_1 \\ x_2x_2 \\ x_2x_3 \\ x_3x_1 \\ x_3x_2 \\ x_3x_3 \end{bmatrix}$$

Requires O(d^2) compute for feature mapping

Requires O(d) compute for Kernel function

What Makes a Valid Kernel Function: Necessary Condition

What Makes a Valid Kernel Function: Necessary Condition

- Kernel Matrix $K_{ij} = K(x^{(i)}, x^{(j)}) = \phi(x^{(i)})^T \phi(x^{(j)})$

What Makes a Valid Kernel Function: Necessary Condition

- Kernel Matrix $K_{ij} = K(x^{(i)}, x^{(j)}) = \phi(x^{(i)})^T \phi(x^{(j)})$

- K is symmetric

$$K(x, z) = K(z, x)$$

What Makes a Valid Kernel Function: Necessary Condition

- Kernel Matrix $K_{ij} = K(x^{(i)}, x^{(j)}) = \phi(x^{(i)})^T \phi(x^{(j)})$

- K is symmetric

$$
\begin{aligned}
z^T K z &= \sum_i \sum_j z_i K_{ij} z_j \\
&= \sum_i \sum_j z_i \phi(x^{(i)})^T \phi(x^{(j)}) z_j \\
&= \sum_i \sum_j z_i \sum_k \phi_k(x^{(i)}) \phi_k(x^{(j)}) z_j \\
&= \sum_k \sum_i \sum_j z_i \phi_k(x^{(i)}) \phi_k(x^{(j)}) z_j \\
&= \sum_k \left(\sum_i z_i \phi_k(x^{(i)}) \right)^2 \\
&\geq 0.
\end{aligned}
$$

What Makes a Valid Kernel Function: Necessary Condition

- Kernel Matrix $K_{ij} = K(x^{(i)}, x^{(j)}) = \phi(x^{(i)})^T \phi(x^{(j)})$

- K is symmetric

- K is positive semidefinite

$$
\begin{aligned}
z^T K z &= \sum_i \sum_j z_i K_{ij} z_j \\
&= \sum_i \sum_j z_i \phi(x^{(i)})^T \phi(x^{(j)}) z_j \\
&= \sum_i \sum_j z_i \sum_k \phi_k(x^{(i)}) \phi_k(x^{(j)}) z_j \\
&= \sum_k \sum_i \sum_j z_i \phi_k(x^{(i)}) \phi_k(x^{(j)}) z_j \\
&= \sum_k \left(\sum_i z_i \phi_k(x^{(i)}) \right)^2 \\
&\geq 0.
\end{aligned}
$$

$$K(x, z) = (x^T z)^2$$

$$z^T K z = \sum_i \sum_j z_i K_{ij} z_j$$

$$= \sum_i \sum_j z_i (x_i^T x_j)^2 z_j$$

$$\geq 0$$

What Makes a Valid Kernel Function: Necessary and Sufficient Condition

Theorem (Mercer). Let $K : \mathbb{R}^d \times \mathbb{R}^d \mapsto \mathbb{R}$ be given. Then for K to be a valid (Mercer) kernel, it is necessary and sufficient that for any $\{x^{(1)}, \ldots, x^{(n)}\}$, $(n < \infty)$, the corresponding kernel matrix is symmetric positive semi-definite.

$\langle x_i, x_j \rangle$

↑ replace

$K(x_i, x_j)$

Recap: Application of Kernel Methods

Recap: Application of Kernel Methods

- In generalized linear models (which we have shown)

Recap: Application of Kernel Methods

- In generalized linear models (which we have shown)
- In support vector machines (which we will show next)

Recap: Application of Kernel Methods

- In generalized linear models (which we have shown)
- In support vector machines (which we will show next)
- Any learning algorithm that you can write in terms of only $\langle x, z \rangle$

replace

$K(x, z)$

Recap: Application of Kernel Methods

- In generalized linear models (which we have shown)

- In support vector machines (which we will show next)

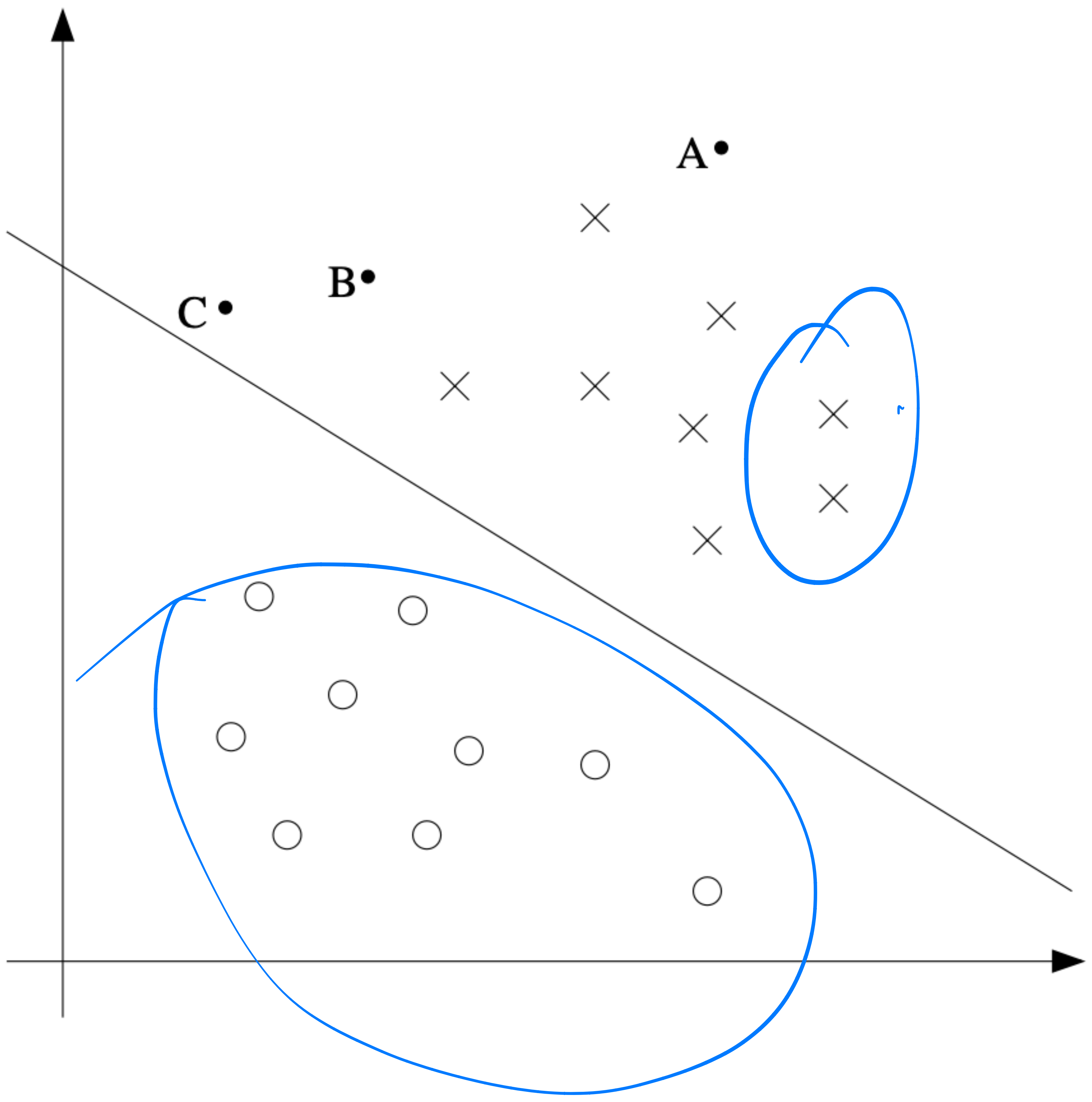
- Any learning algorithm that you can write in terms of only <x, z>

Just replace <x, z> with K(x, z), you magically transform the algorithm to work efficiently in the *implicit* high dimensional feature space

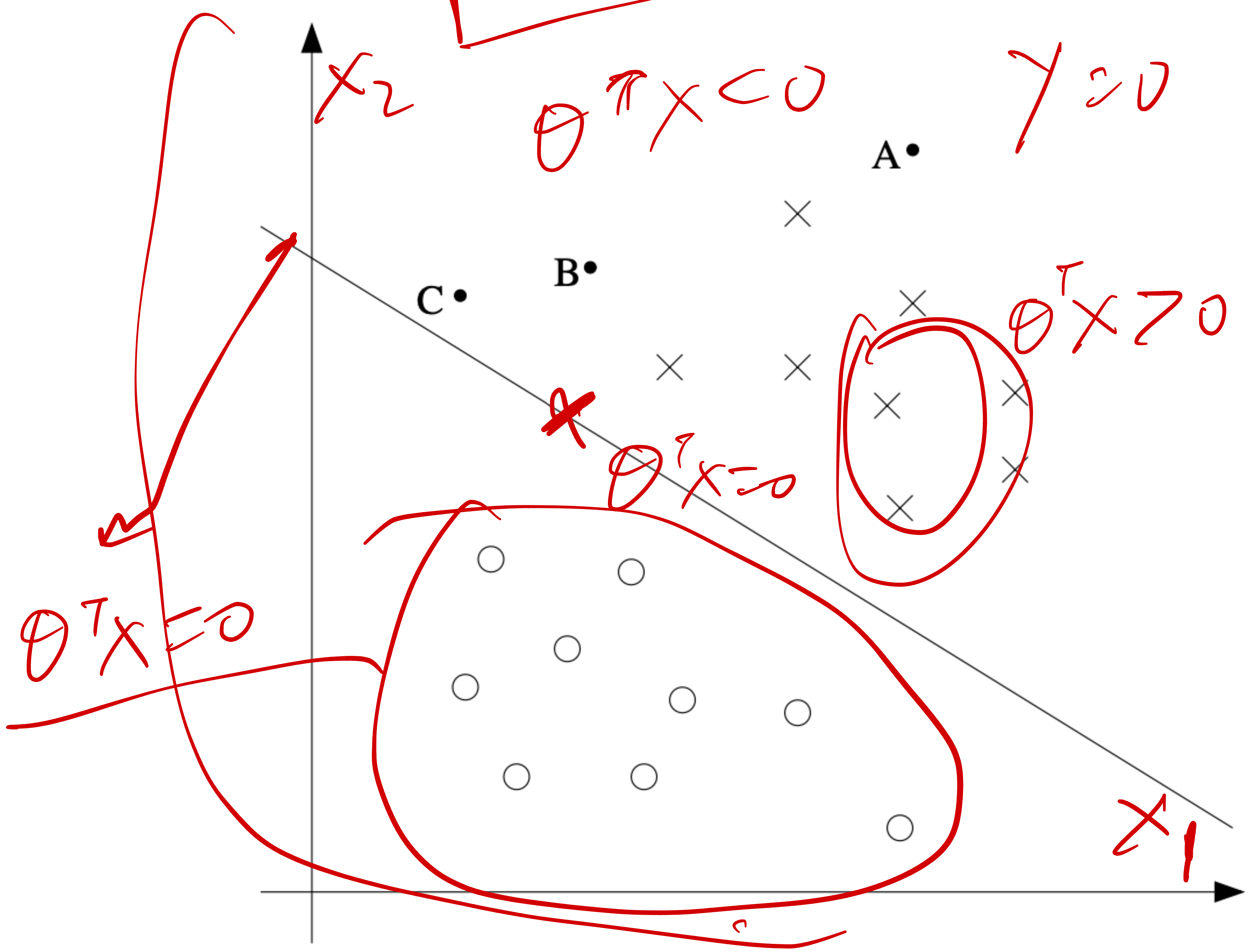
Support Vector Machines

Confidence in Logistic Regression

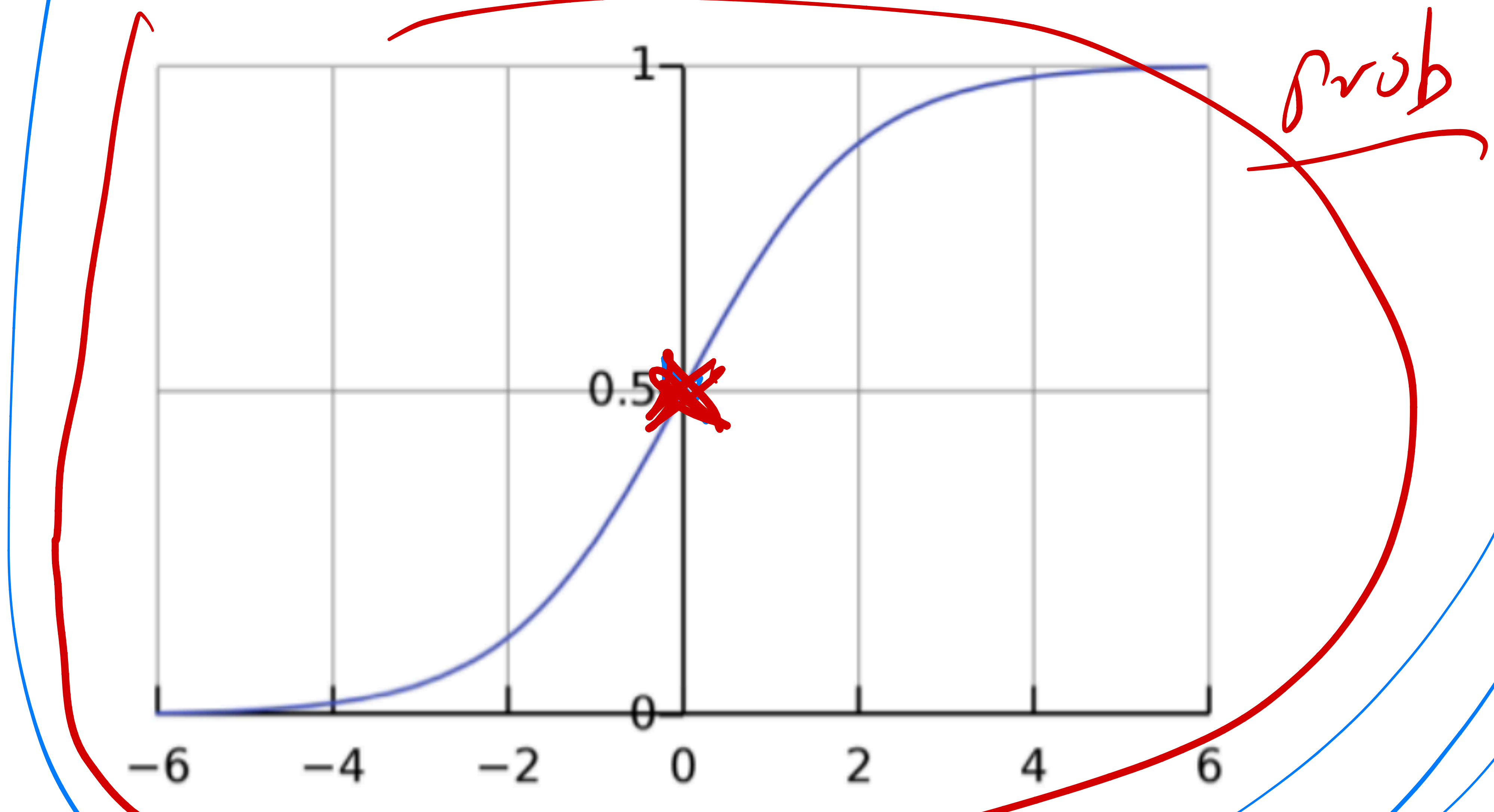
Confidence in Logistic Regression



Confidence in Logistic Regression

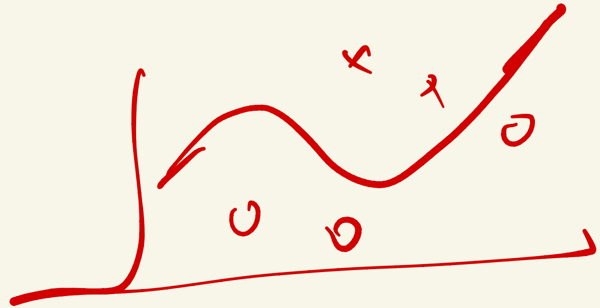


$$p(y) = \frac{1}{1 + e^{-\theta^T x}}$$



$\theta^T x \geq 0, \quad y=1$

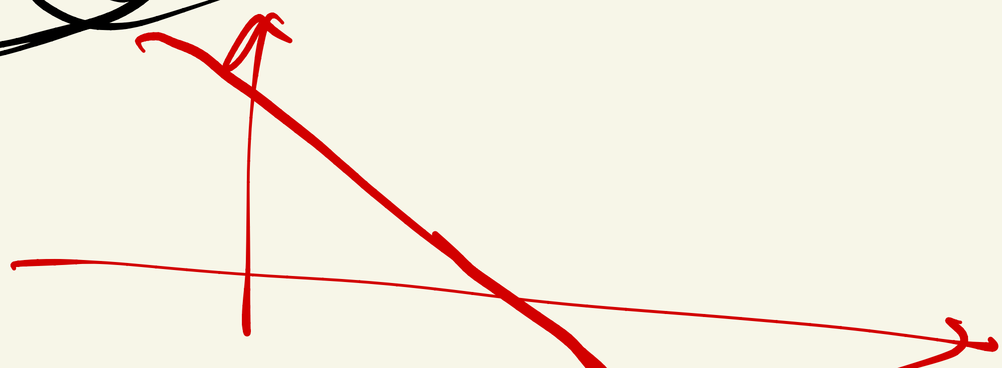
$\theta^T x < 0, \quad y=1$



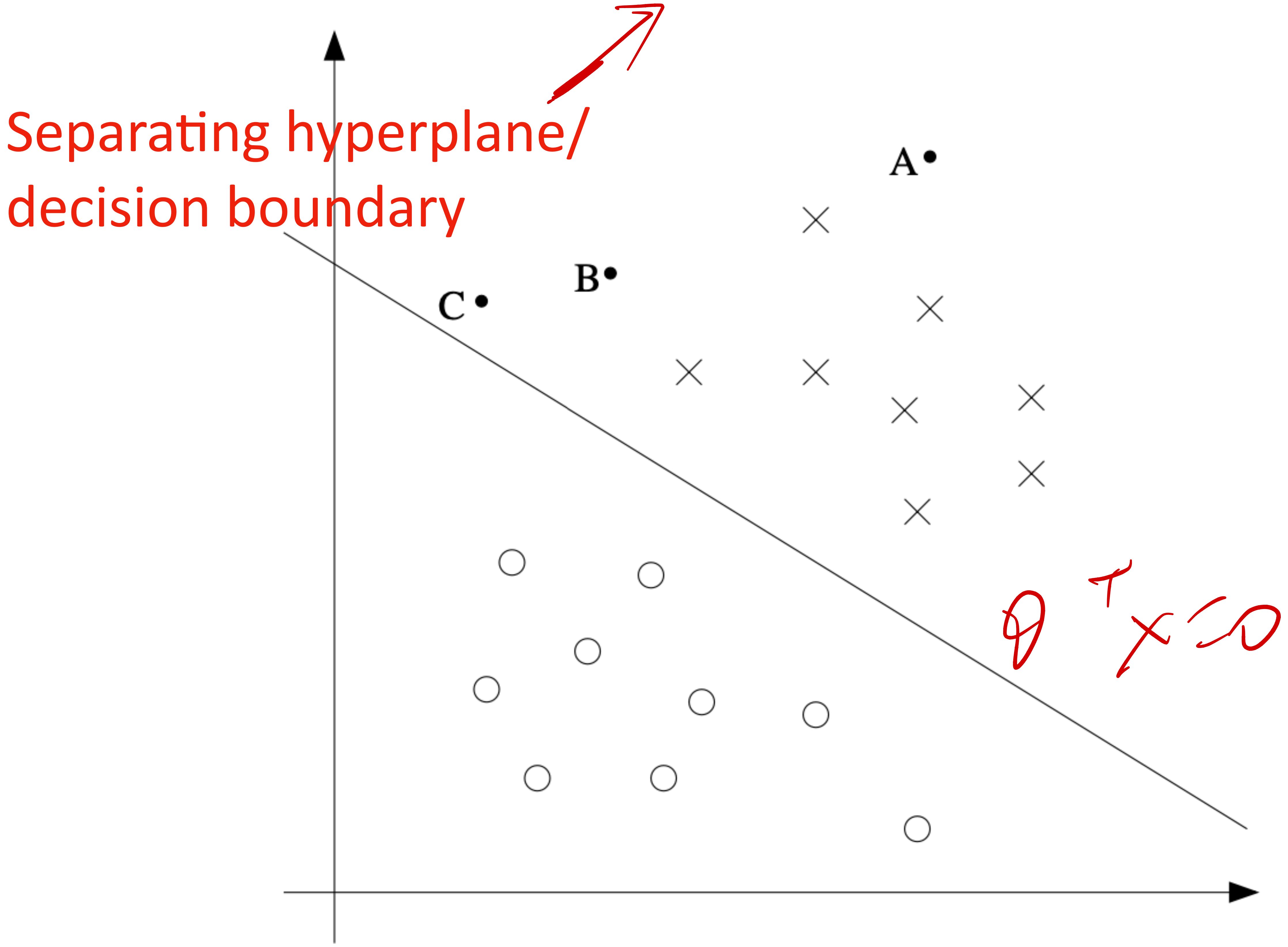
$x_0 = 1$

$\theta^T x = 0$ → linear

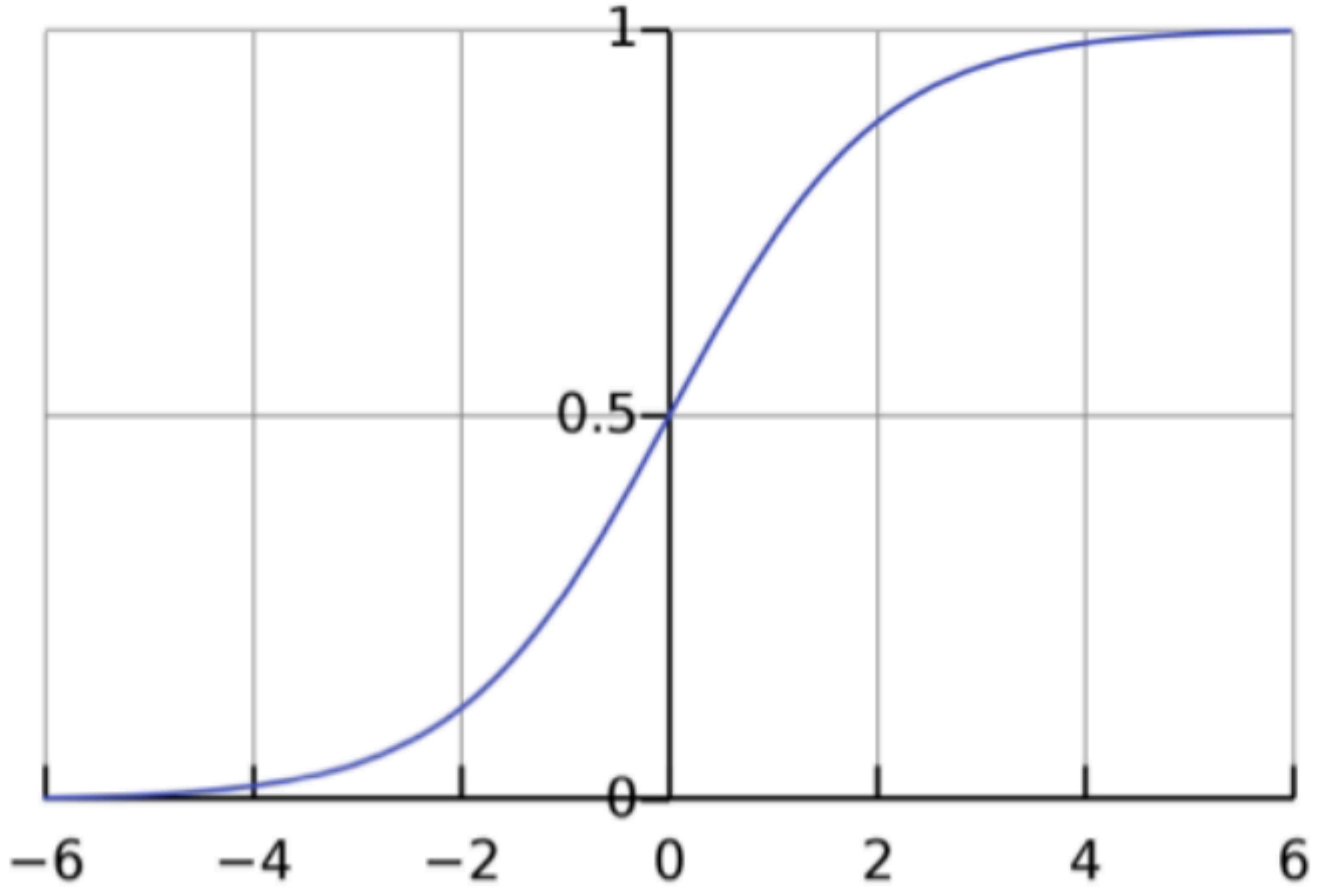
→ $\theta_1 x_1 + \theta_2 x_2 + \theta_0 = 0$



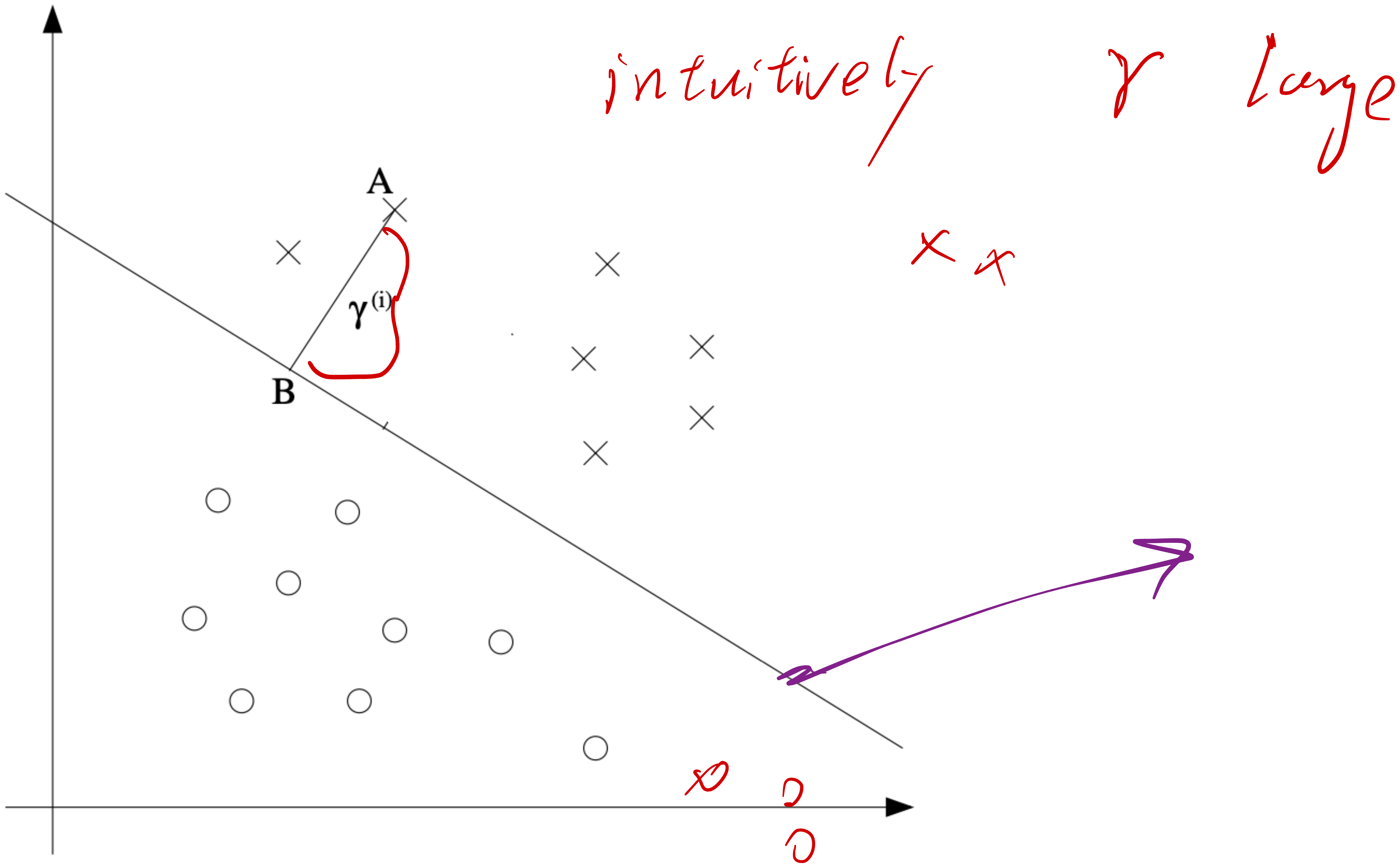
Confidence in Logistic Regression

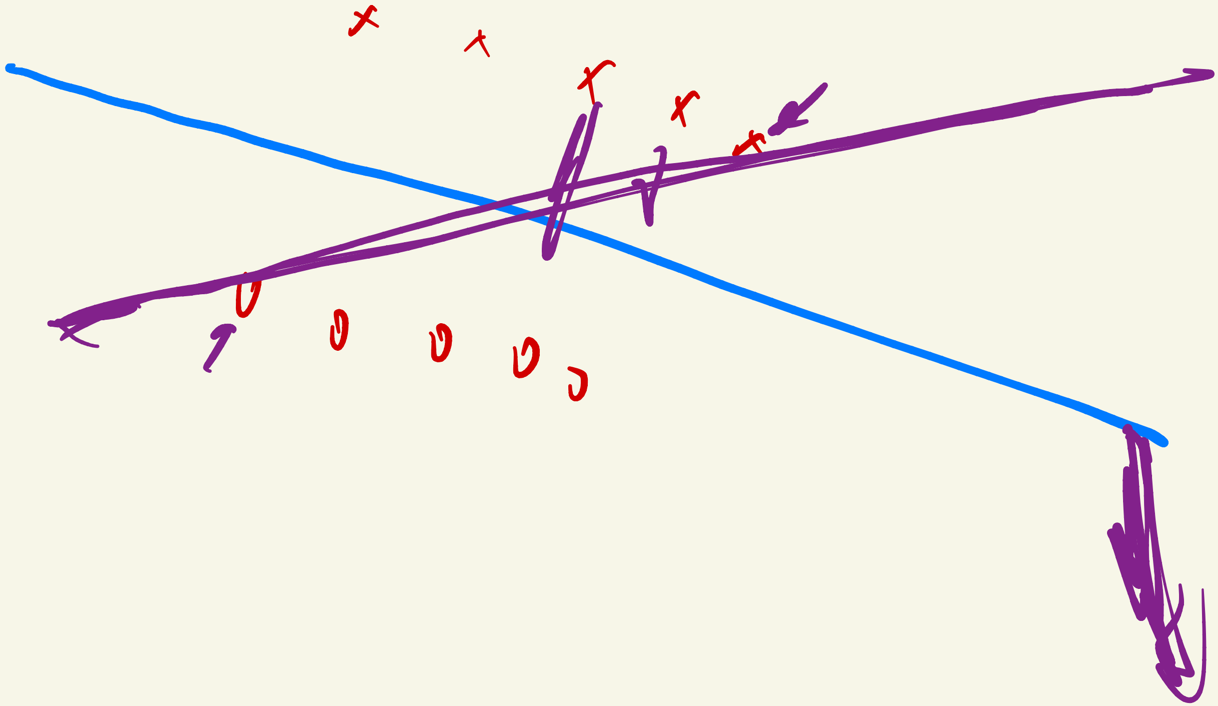


max $$p(y) = \frac{1}{1 + e^{-\theta^T x}}$$



Margin





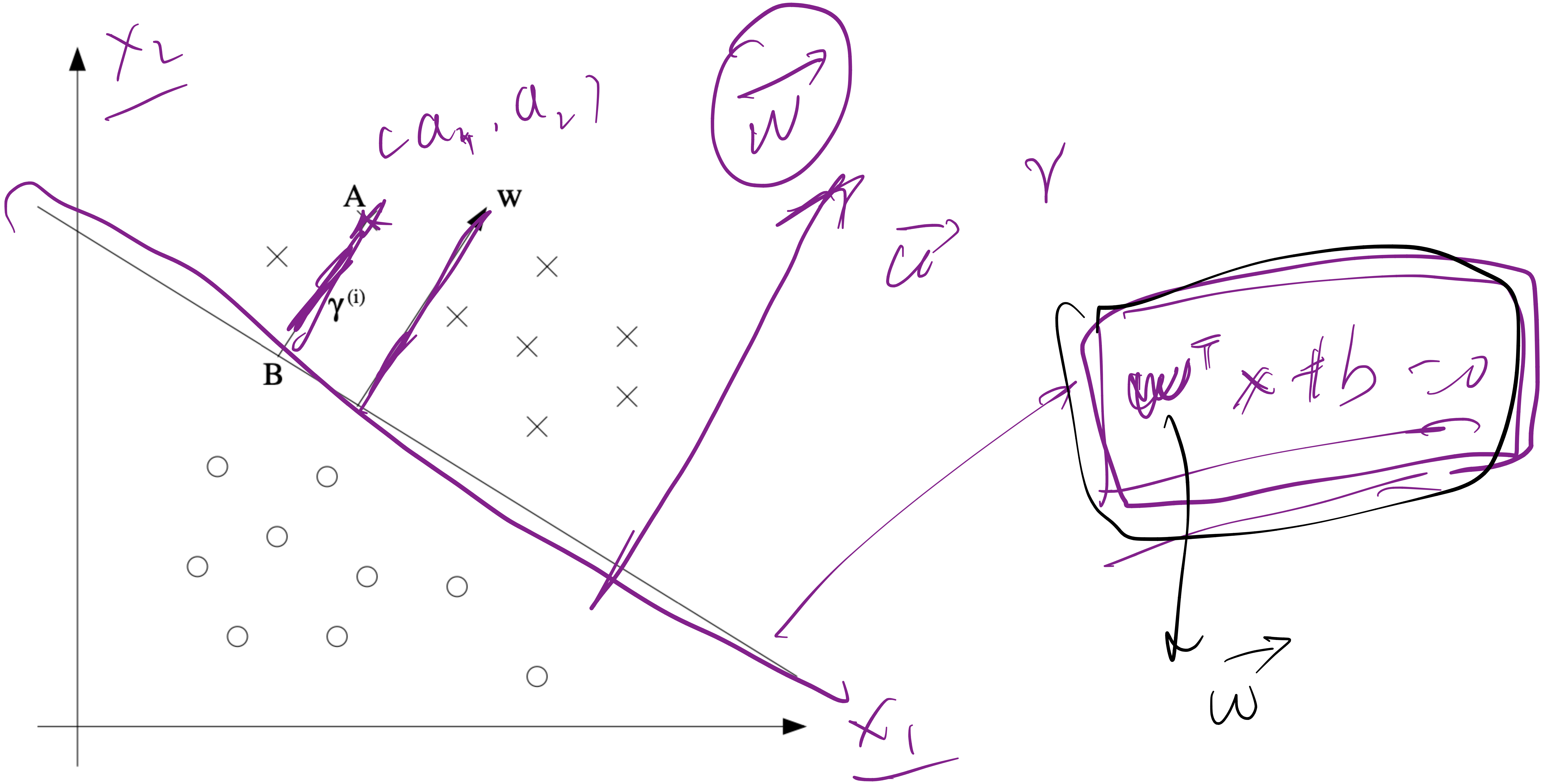
New Notations

Consider a binary classification problem, with the input feature x and $y \in \{-1, 1\}$ (instead of $\{0, 1\}$), the classifier is:

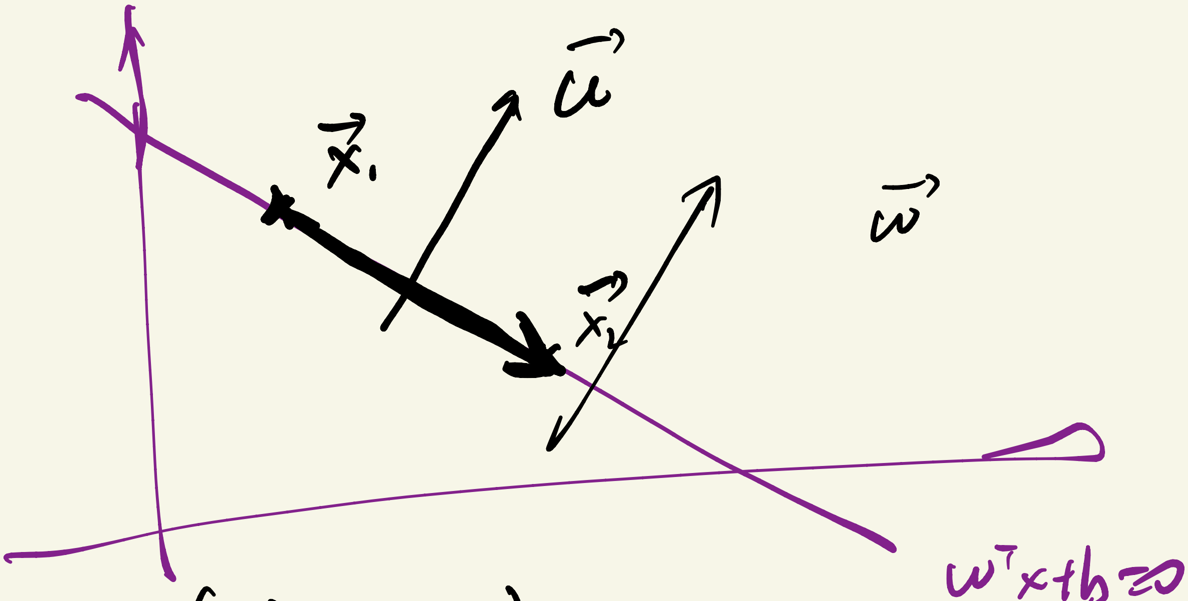
$$h_{w,b}(x) = g(w^T x + b).$$

$$g(z) = 1 \text{ if } z \geq 0, \text{ and } g(z) = -1$$

Geometric Margin



What is the geometric margin?

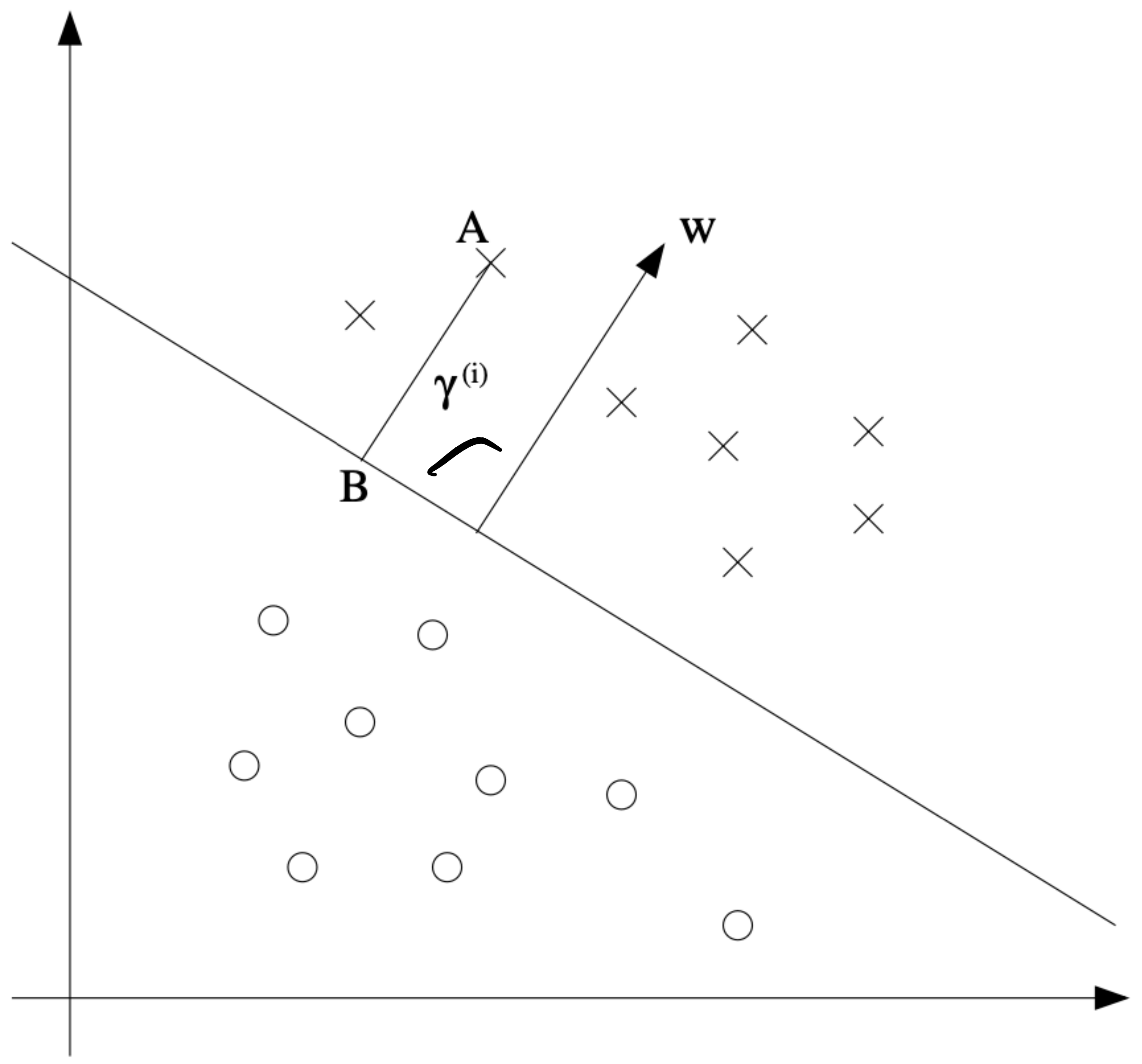


$$\vec{u} \cdot (\vec{x}_2 - \vec{x}_1) = 0$$

$$\left.\begin{array}{l} w^T \vec{x}_1 + b = 0 \\ w^T \vec{x}_2 + b = 0 \end{array}\right\} \Rightarrow w^T (\vec{x}_1 - \vec{x}_2) = 0$$

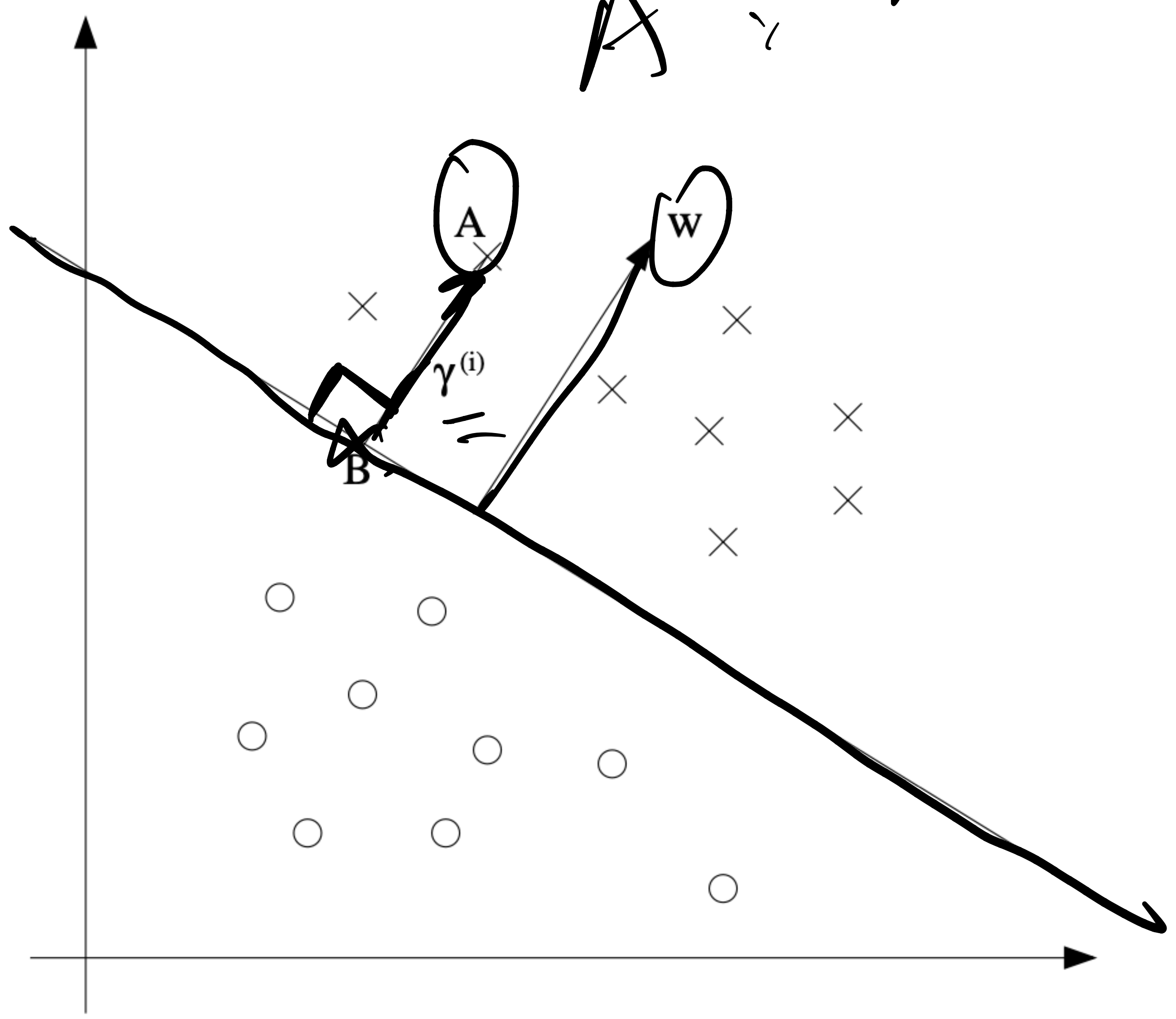
Geometric Margin

Geometric Margin



Geometric Margin

$\vec{A}$, $x^{(i)}$

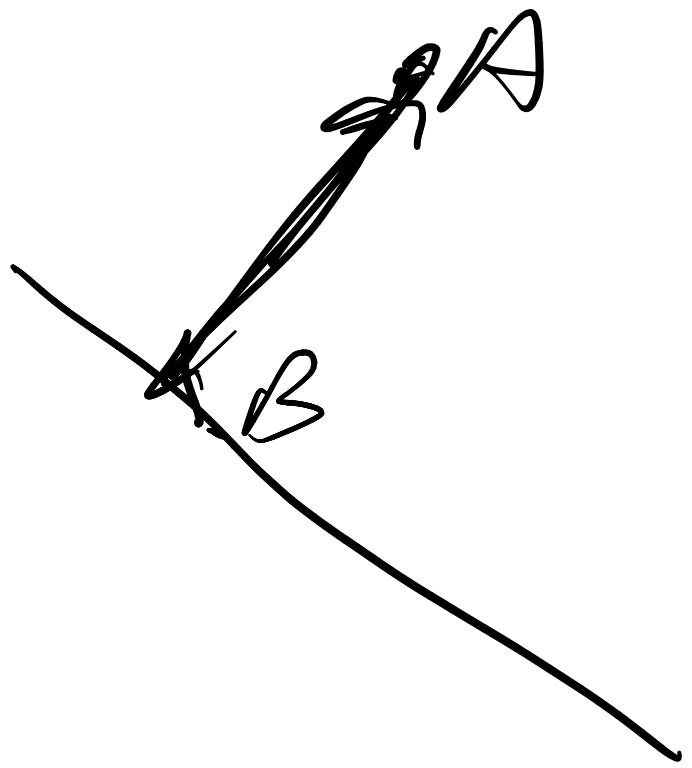


$||\vec{A} - \vec{B}||$

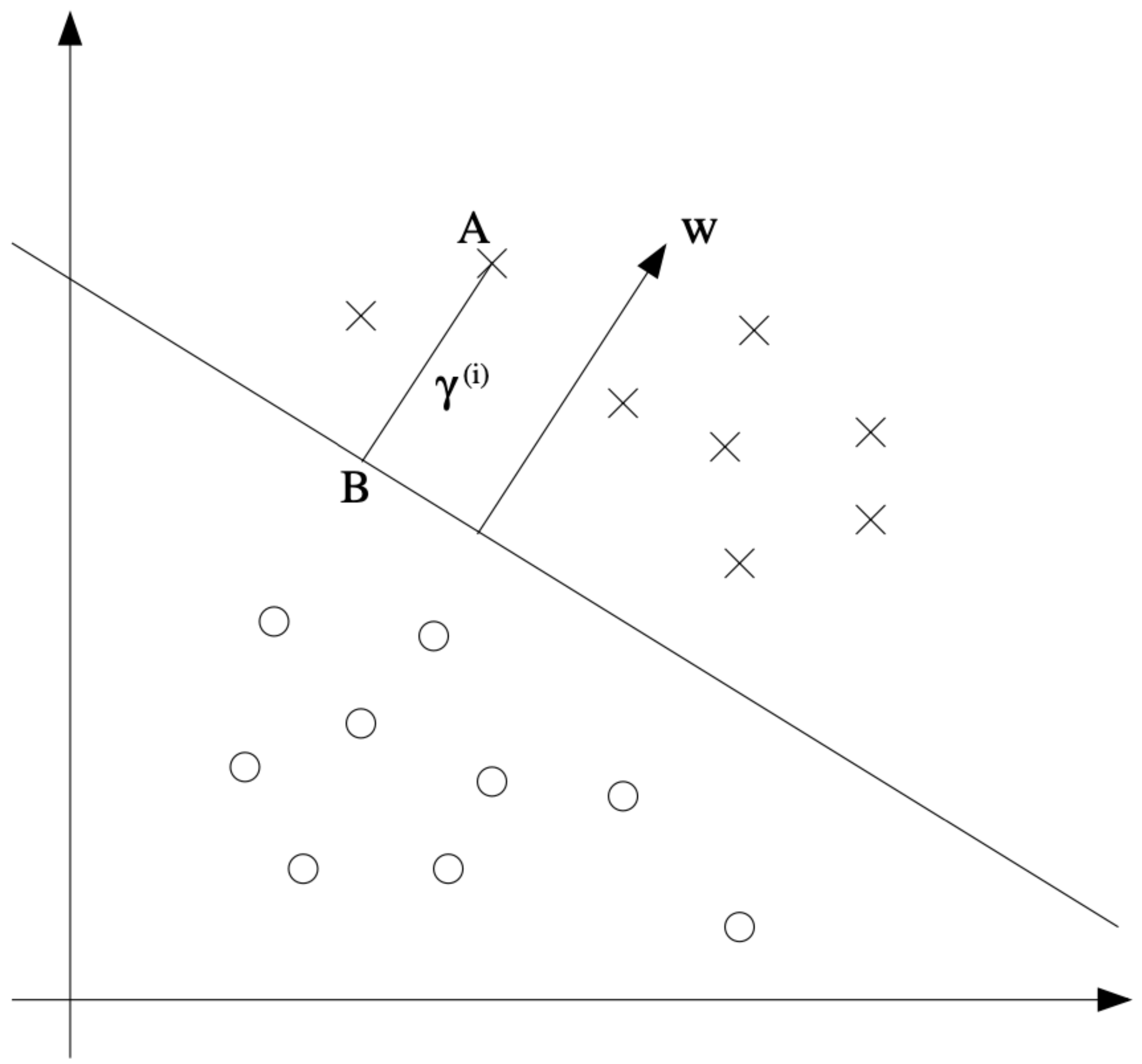
$$w^T \left(x^{(i)} - \gamma^{(i)} \frac{w}{||w||} \right) + b = 0.$$

$$\vec{B} = \vec{A} - \gamma \frac{\vec{w}}{||w||}$$

$$w^T \vec{B} + b = 0$$



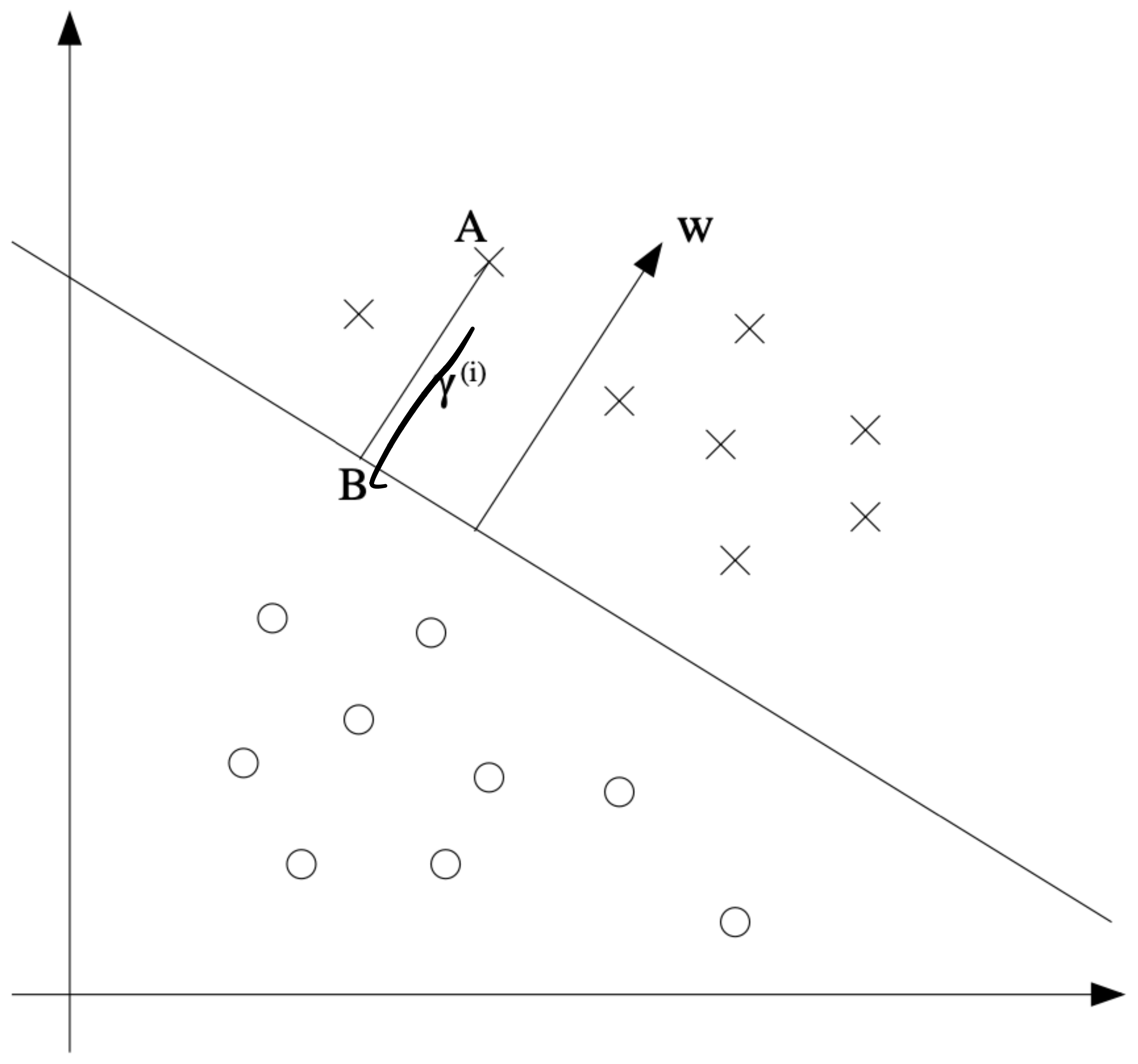
Geometric Margin



$$w^T \left(x^{(i)} - \gamma^{(i)} \frac{w}{||w||} \right) + b = 0.$$

$$\gamma^{(i)} = \frac{w^T x^{(i)} + b}{||w||} = \left(\frac{w}{||w||} \right)^T x^{(i)} + \frac{b}{||w||}$$

Geometric Margin

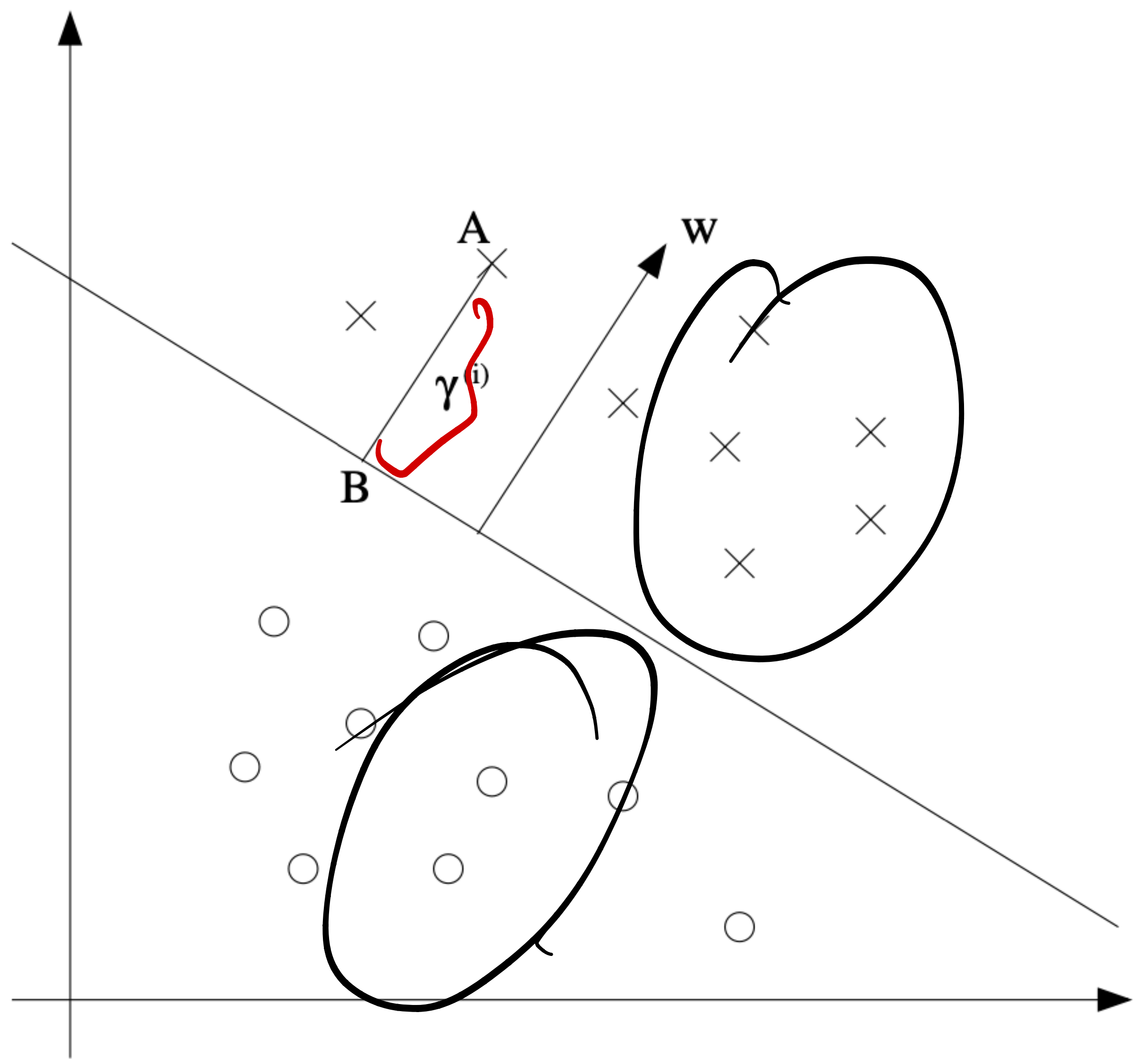


$$w^T \left(x^{(i)} - \gamma^{(i)} \frac{w}{||w||} \right) + b = 0.$$

$$\gamma^{(i)} = \frac{w^T x^{(i)} + b}{||w||} = \left(\frac{w}{||w||} \right)^T x^{(i)} + \frac{b}{||w||}$$

Generally

Geometric Margin



$$w^T\left(x^{(i)} - \gamma^{(i)}\frac{w}{||w||}\right) + b = 0.$$

$$\gamma^{(i)} = \frac{w^T x^{(i)} + b}{||w||} = \left(\frac{w}{||w||}\right)^T x^{(i)} + \frac{b}{||w||}$$

Generally

$$\gamma^{(i)} = y^{(i)}\left(\left(\frac{w}{||w||}\right)^T x^{(i)} + \frac{b}{||w||}\right)$$

Geometric Margin

Given a training set $S = \{(x^{(i)}, y^{(i)}); i = 1,\ldots,n\}$

$$\gamma = \min_{i=1,\ldots,n} \gamma^{(i)}$$

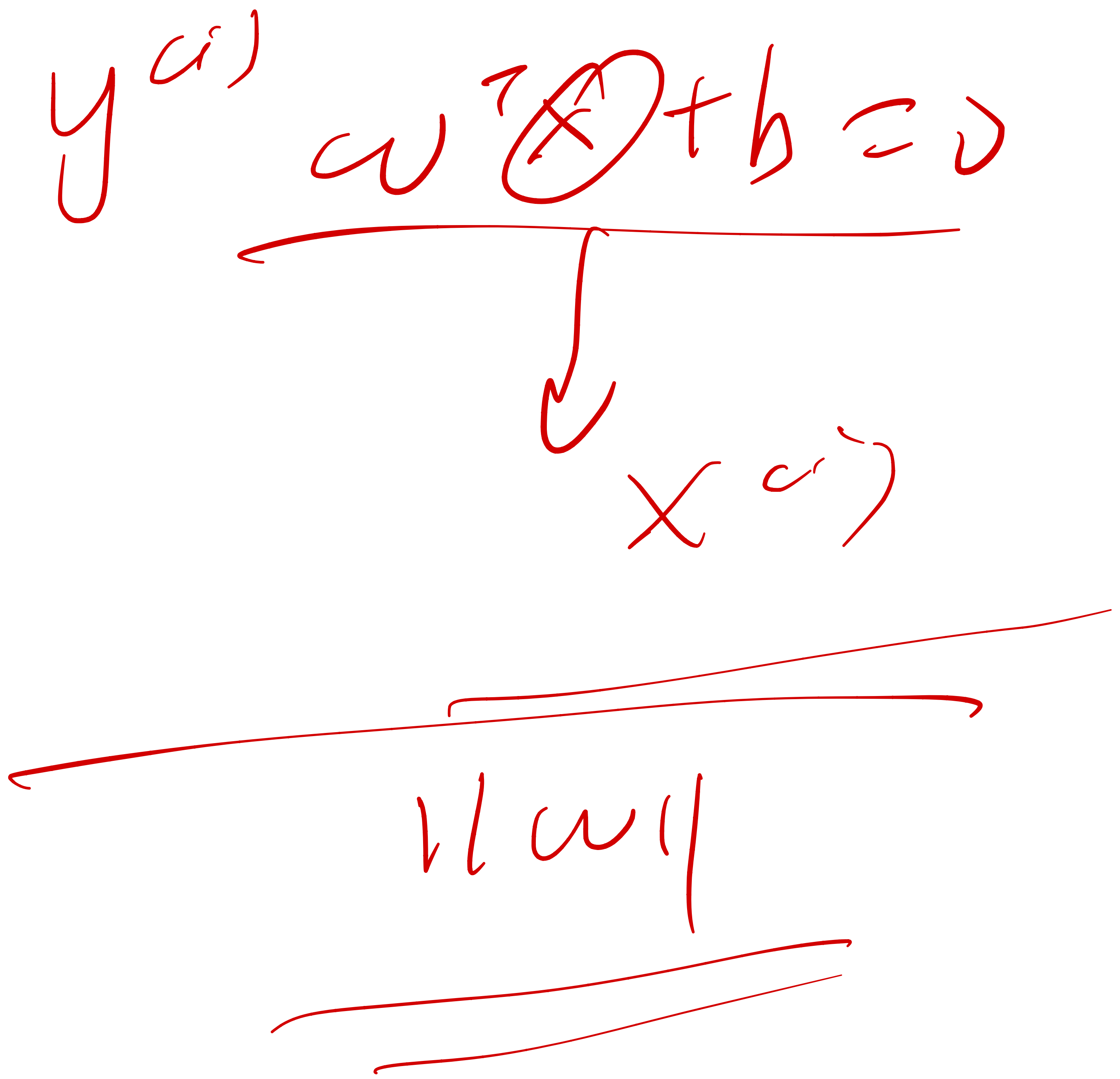
max γ

$$\max \min_{i=1,\ldots,n} \gamma^{(i)}$$

Functional Margin

Given a training example $(x^{(i)}, y^{(i)})$

$$\hat{\gamma}^{(i)} = y^{(i)}(w^T x^{(i)} + b).$$



Functional Margin

Given a training example $(x^{(i)}, y^{(i)})$

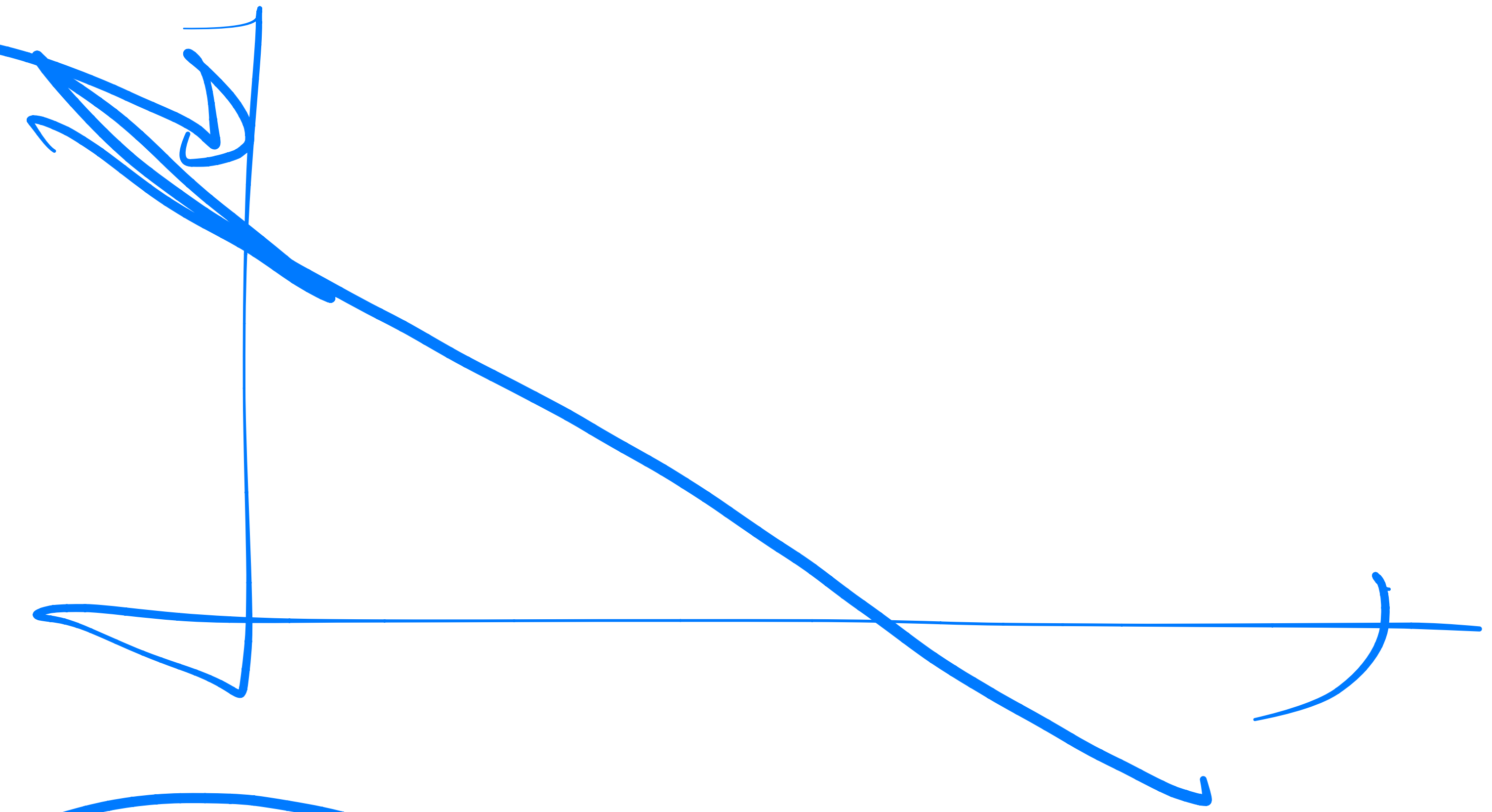
$$\hat{\gamma}^{(i)} = y^{(i)}(w^T x^{(i)} + b).$$

$$\gamma = \frac{\hat{\gamma}}{\|w\|}$$

Given a training set $S = \{(x^{(i)}, y^{(i)}); i = 1,\ldots,n\}$

$$\hat{\gamma} = \min_{i=1,\ldots,n} \hat{\gamma}^{(i)}$$

Functional Margin



Given a training example $(x^{(i)}, y^{(i)})$

$$\hat{\gamma}^{(i)} = y^{(i)}(w^T x^{(i)} + b).$$

Given a training set $S = \{(x^{(i)}, y^{(i)}); i = 1,\ldots,n\}$

$$\hat{\gamma} = \min_{i=1,\ldots,n} \hat{\gamma}^{(i)}$$

Functional margin changes when rescaling parameters, making it a bad objective, e.g. when w->2w, b->2b, the functional margin changes while the separating plane does not really change

The Optimization Problem

geometric

$$\max_{w,b} \min_{i=1,\dots,n} \gamma^{(i)}$$

The Optimization Problem

$$\max_{w,b} \min_{i=1,\ldots,n} \gamma^{(i)}$$

$$\max_{\hat{\gamma},w,b} \frac{\hat{\gamma}}{||w||}$$

$$\text{s.t.} \quad y^{(i)}(w^T x^{(i)} + b) \geq \hat{\gamma}, \quad i = 1, \ldots, n$$

$$\frac{y^{(i)}(w^T x^{(i)} + b)}{||w||} \geq \frac{\hat{\gamma}}{||w||}$$

The Optimization Problem

$$\max_{w,b} \min_{i=1,\ldots,n} \gamma^{(i)}$$

$w \to 2w$

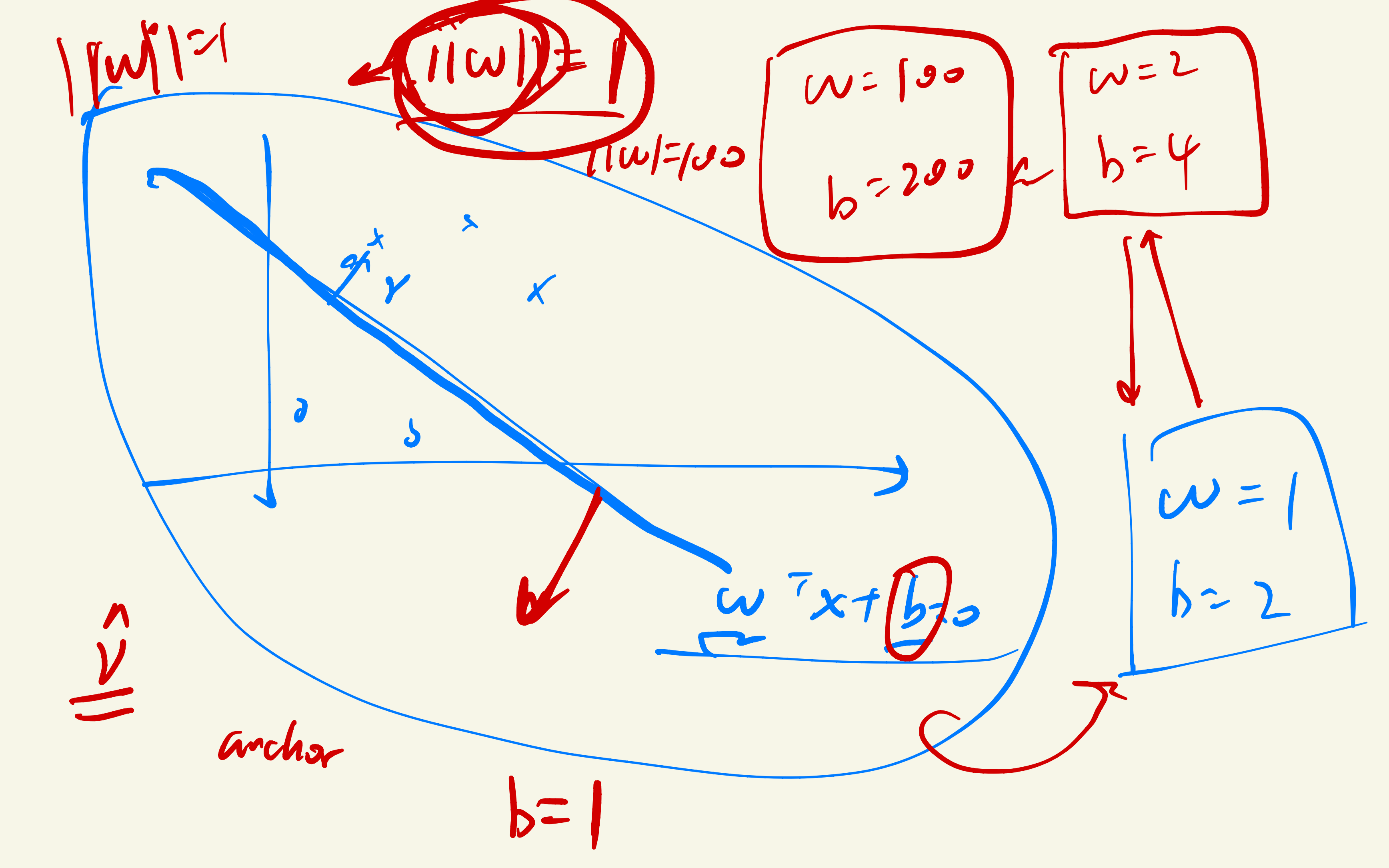
$b \to 2b$

$\hat{\gamma} \to 2\hat{\gamma}$

$$\max_{\hat{\gamma},w,b} \frac{\hat{\gamma}}{||w||}$$

$$\text{s.t.} \quad y^{(i)}(w^T x^{(i)} + b) \geq \hat{\gamma}, \quad i = 1, \ldots, n$$

Infinite solutions, as $\hat{\gamma}$ can be at any scale without changing the classifier


||w||=1
||w|| = 1
||w||=100
w=100
b=200
w=2
b=4
w=1
b=2
wᵀx+b=0
anchor
b=1

The Optimization Problem

$$\max_{w,b} \quad \min_{i=1,\ldots,n} \gamma^{(i)}$$

⇩

$$\max_{\hat{\gamma},w,b} \quad \frac{\hat{\gamma}}{||w||}$$

$$\text{s.t.} \quad y^{(i)}(w^T x^{(i)} + b) \geq \hat{\gamma}, \quad i = 1, \ldots, n$$

Infinite solutions, as $\hat{\gamma}$ can be at any scale without changing the classifier

||w|| is not easy to deal with, non-convex objective

The Optimization Problem

$$\max_{\hat{\gamma},w,b} \quad \frac{\hat{\gamma}}{||w||}$$
$$\text{s.t.} \quad y^{(i)}(w^T x^{(i)} + b) \geq \hat{\gamma}, \quad i = 1, \ldots, n$$

The Optimization Problem

$$\max_{\hat{\gamma},w,b} \quad \frac{\hat{\gamma}}{||w||}$$

$$\text{s.t.} \quad y^{(i)}(w^T x^{(i)} + b) \geq \hat{\gamma}, \quad i = 1, \ldots, n$$

Add constraint $\hat{\gamma} = 1$

The Optimization Problem

$$\max_{\hat{\gamma},w,b} \quad \frac{\hat{\gamma}}{||w||}$$

$$\text{s.t.} \quad y^{(i)}(w^T x^{(i)} + b) \geq \hat{\gamma}, \quad i = 1, \ldots, n$$

Add constraint $\hat{\gamma} = 1$

$$\min_{w,b} \quad \frac{1}{2}||w||^2$$

$$\text{s.t.} \quad y^{(i)}(w^T x^{(i)} + b) \geq 1, \quad i = 1, \ldots, n$$

The Optimization Problem

$$\max_{\hat{\gamma},w,b} \quad \frac{\hat{\gamma}}{||w||}$$

$$\text{s.t.} \quad y^{(i)}(w^T x^{(i)} + b) \geq \hat{\gamma}, \quad i = 1, \ldots, n$$

Add constraint $\hat{\gamma} = 1$

$$\min_{w,b} \quad \frac{1}{2}||w||^2$$

$$\text{s.t.} \quad y^{(i)}(w^T x^{(i)} + b) \geq 1, \quad i = 1, \ldots, n$$

Assumption: the training dataset is linearly separable

Lagrange Duality — Lagrange Multiplier

Lagrange Duality — Lagrange Multiplier

$$\min_w \quad f(w)$$
$$\text{s.t.} \quad h_i(w) = 0, \quad i = 1, \ldots, l.$$

Lagrange Duality — Lagrange Multiplier

$$\min_w \quad f(w)$$
$$\text{s.t.} \quad h_i(w) = 0, \quad i = 1, \ldots, l.$$

$$\mathcal{L}(w, \beta) = f(w) + \sum_{i=1}^{l} \beta_i h_i(w)$$

Lagrange Duality — Lagrange Multiplier

$$\begin{aligned} \min_w \quad & f(w) \\ \text{s.t.} \quad & h_i(w) = 0, \quad i = 1, \ldots, l. \end{aligned}$$

min $$\mathcal{L}(w, \beta) = f(w) + \sum_{i=1}^{l} \beta_i h_i(w)$$

Solve w, β

$$\frac{\partial \mathcal{L}}{\partial w_i} = 0; \quad \frac{\partial \mathcal{L}}{\partial \beta_i} = 0,$$

Lagrange Multiplier: Example

$$\min_{x,y} 5x - 3y$$

$$\text{s.t.} \quad x^2 + y^2 = 136$$

$\min_{x,y,\beta} L(x, y, \beta)$

$L(x, y, \beta) = 5x - 3y + \beta(x^2 + y^2 - 136)$

$\frac{\partial L}{\partial x} = 5 + 2\beta x = 0$

$\frac{\partial L}{\partial y} = -3 + 2\beta y = 0$

$\frac{\partial L}{\partial \beta} = x^2 + y^2 - 136 = 0$

Generalized Lagrangian

Generalized Lagrangian

Primal optimization problem

$$\begin{aligned} \min_w \quad & f(w) \\ \text{s.t.} \quad & g_i(w) \leq 0, \quad i = 1, \ldots, k \\ & h_i(w) = 0, \quad i = 1, \ldots, l. \end{aligned}$$

$y(w^T x + b) \geq 1$

$1 - y(w^T x + b) \leq 0$

Generalized Lagrangian

Primal optimization problem

$$\begin{aligned} \min_w \quad & f(w) \\ \text{s.t.} \quad & g_i(w) \le 0, \quad i = 1, \ldots, k \\ & h_i(w) = 0, \quad i = 1, \ldots, l. \end{aligned}$$

Generalized Lagrangian

$$\mathcal{L}(w, \alpha, \beta) = f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{l} \beta_i h_i(w)$$

Generalized Lagrangian

$$\mathcal{L}(w, \alpha, \beta) = f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{l} \beta_i h_i(w)$$

$$\min_{w,\beta} L(w, \beta)$$

Generalized Lagrangian

$$\mathcal{L}(w,\alpha,\beta) = f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{l} \beta_i h_i(w)$$

$$\theta_{\mathcal{P}}(w) = \max_{\alpha,\beta\,:\,\alpha_i \geq 0} f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{l} \beta_i h_i(w)$$

$\min \theta_p(w)$

$h_i(w) = 0$

$\max_{\beta} f(w) + \sum_{i=1}^{l} \beta_i h_i(w)$

Generalized Lagrangian

$$\mathcal{L}(w, \alpha, \beta) = f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{l} \beta_i h_i(w)$$

$$\theta_{\mathcal{P}}(w) \quad = \quad \max_{\alpha,\beta\,:\,\alpha_i \geq 0} f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{l} \beta_i h_i(w)$$

$$\theta_{\mathcal{P}}(w) = \begin{cases} f(w) & \text{if } w \text{ satisfies primal constraints} \\ \infty & \text{otherwise.} \end{cases}$$

Generalized Lagrangian

Consider this optimization problem

$$\min_w \theta_{\mathcal{P}}(w) = \min_w \max_{\alpha,\beta\,:\,\alpha_i \geq 0} \mathcal{L}(w, \alpha, \beta)$$

Generalized Lagrangian

Consider this optimization problem

$$\min_{w} \theta_{\mathcal{P}}(w) = \min_{w} \max_{\alpha,\beta\,:\,\alpha_i \geq 0} \mathcal{L}(w, \alpha, \beta)$$

It has exactly the same solution as our original problem

Generalized Lagrangian

Consider this optimization problem

$$\min_w \theta_{\mathcal{P}}(w) = \min_w \max_{\alpha,\beta\,:\,\alpha_i \geq 0} \mathcal{L}(w, \alpha, \beta)$$

It has exactly the same solution as our original problem

$$p^* = \min_w \theta_{\mathcal{P}}(w)$$

$$\max_{\alpha,\beta:\alpha_i \geq 0} \min_w L(w,\alpha,\beta)$$

The Dual Problem in Optimization

In optimization, sometimes the primal optimization is hard to solve, then we may find a related alternative optimization problem that can be solved more easily, to solve the orignal problem in an indirect way

The Dual Problem

$$\theta_{\mathcal{D}}(\alpha, \beta) = \min_{w} \mathcal{L}(w, \alpha, \beta)$$

The Dual Problem

$$\theta_{\mathcal{D}}(\alpha, \beta) = \min_{w} \mathcal{L}(w, \alpha, \beta)$$

The dual optimization problem

$$\max_{\alpha,\beta\,:\,\alpha_i \geq 0} \theta_{\mathcal{D}}(\alpha, \beta) = \max_{\alpha,\beta\,:\,\alpha_i \geq 0} \min_{w} \mathcal{L}(w, \alpha, \beta)$$

The Dual Problem

$$\theta_{\mathcal{D}}(\alpha, \beta) = \min_{w} \mathcal{L}(w, \alpha, \beta)$$

The dual optimization problem

$$\max_{\alpha,\beta : \alpha_i \geq 0} \theta_{\mathcal{D}}(\alpha, \beta) = \max_{\alpha,\beta : \alpha_i \geq 0} \min_{w} \mathcal{L}(w, \alpha, \beta)$$

The primal optimization problem

$$\min_{w} \theta_{\mathcal{P}}(w) = \min_{w} \max_{\alpha,\beta : \alpha_i \geq 0} \mathcal{L}(w, \alpha, \beta)$$

KKT Conditions

KKT Conditions

Denote the solution to the primal problem as $w*$, the solution to the dual problem as $\alpha*, \beta*$, then zero duality gap is sufficient and necessary (i.e. equivalent) to satisfy KKT Conditions:

KKT Conditions

Denote the solution to the primal problem as $w*$, the solution to the dual problem as $\alpha*, \beta*$, then zero duality gap is sufficient and necessary (i.e. equivalent) to satisfy KKT Conditions:

$$\mathcal{L}(w, \alpha, \beta) = f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{l} \beta_i h_i(w)$$

KKT Conditions

Denote the solution to the primal problem as w^*, the solution to the dual problem as α^*, β^*, then zero duality gap is sufficient and necessary (i.e. equivalent) to satisfy KKT Conditions:

$$\mathcal{L}(w, \alpha, \beta) = f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{l} \beta_i h_i(w)$$

$$\begin{aligned}
\frac{\partial}{\partial w_i}\mathcal{L}(w^*, \alpha^*, \beta^*) &= 0, \quad i = 1, \ldots, d \\
\frac{\partial}{\partial \beta_i}\mathcal{L}(w^*, \alpha^*, \beta^*) &= 0, \quad i = 1, \ldots, l \\
\alpha_i^* g_i(w^*) &= 0, \quad i = 1, \ldots, k \\
g_i(w^*) &\leq 0, \quad i = 1, \ldots, k \\
\alpha^* &\geq 0, \quad i = 1, \ldots, k
\end{aligned}$$

KKT Conditions

Denote the solution to the primal problem as w^*, the solution to the dual problem as α^*, β^*, then zero duality gap is sufficient and necessary (i.e. equivalent) to satisfy KKT Conditions:

$$\mathcal{L}(w, \alpha, \beta) = f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{l} \beta_i h_i(w)$$

$$\begin{aligned}
\frac{\partial}{\partial w_i}\mathcal{L}(w^*, \alpha^*, \beta^*) &= 0, \quad i = 1, \dots, d \\
\frac{\partial}{\partial \beta_i}\mathcal{L}(w^*, \alpha^*, \beta^*) &= 0, \quad i = 1, \dots, l \\
\alpha_i^* g_i(w^*) &= 0, \quad i = 1, \dots, k \\
g_i(w^*) &\leq 0, \quad i = 1, \dots, k \\
\alpha^* &\geq 0, \quad i = 1, \dots, k
\end{aligned}$$

Normal Lagrange multiplier equations

KKT Conditions

Denote the solution to the primal problem as w^*, the solution to the dual problem as α^*, β^*, then zero duality gap is sufficient and necessary (i.e. equivalent) to satisfy KKT Conditions:

$$\mathcal{L}(w, \alpha, \beta) = f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{l} \beta_i h_i(w)$$

$$\begin{aligned}
\frac{\partial}{\partial w_i}\mathcal{L}(w^*, \alpha^*, \beta^*) &= 0, \quad i = 1, \dots, d \\
\frac{\partial}{\partial \beta_i}\mathcal{L}(w^*, \alpha^*, \beta^*) &= 0, \quad i = 1, \dots, l \\
\alpha_i^* g_i(w^*) &= 0, \quad i = 1, \dots, k \\
g_i(w^*) &\leq 0, \quad i = 1, \dots, k \\
\alpha^* &\geq 0, \quad i = 1, \dots, k
\end{aligned}$$

Normal Lagrange multiplier equations (first two conditions)

The original constraints (last two conditions)

KKT Conditions

Denote the solution to the primal problem as w^*, the solution to the dual problem as α^*, β^*, then zero duality gap is sufficient and necessary (i.e. equivalent) to satisfy KKT Conditions:

$$\mathcal{L}(w, \alpha, \beta) = f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{l} \beta_i h_i(w)$$

$$\begin{aligned}
\frac{\partial}{\partial w_i} \mathcal{L}(w^*, \alpha^*, \beta^*) &= 0, \quad i = 1, \ldots, d \\
\frac{\partial}{\partial \beta_i} \mathcal{L}(w^*, \alpha^*, \beta^*) &= 0, \quad i = 1, \ldots, l \\
\alpha_i^* g_i(w^*) &= 0, \quad i = 1, \ldots, k \\
g_i(w^*) &\leq 0, \quad i = 1, \ldots, k \\
\alpha^* &\geq 0, \quad i = 1, \ldots, k
\end{aligned}$$

KKT Conditions

Denote the solution to the primal problem as w^*, the solution to the dual problem as α^*, β^*, then zero duality gap is sufficient and necessary (i.e. equivalent) to satisfy KKT Conditions:

$$\mathcal{L}(w, \alpha, \beta) = f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{l} \beta_i h_i(w)$$

$$\begin{aligned}
\frac{\partial}{\partial w_i}\mathcal{L}(w^*, \alpha^*, \beta^*) &= 0, \quad i = 1, \ldots, d \\
\frac{\partial}{\partial \beta_i}\mathcal{L}(w^*, \alpha^*, \beta^*) &= 0, \quad i = 1, \ldots, l \\
\alpha_i^* g_i(w^*) &= 0, \quad i = 1, \ldots, k \\
g_i(w^*) &\leq 0, \quad i = 1, \ldots, k \\
\alpha^* &\geq 0, \quad i = 1, \ldots, k
\end{aligned}$$

If $\alpha_i^* > 0$, then $g_i(w^*) = 0$, the inequality is actually equality

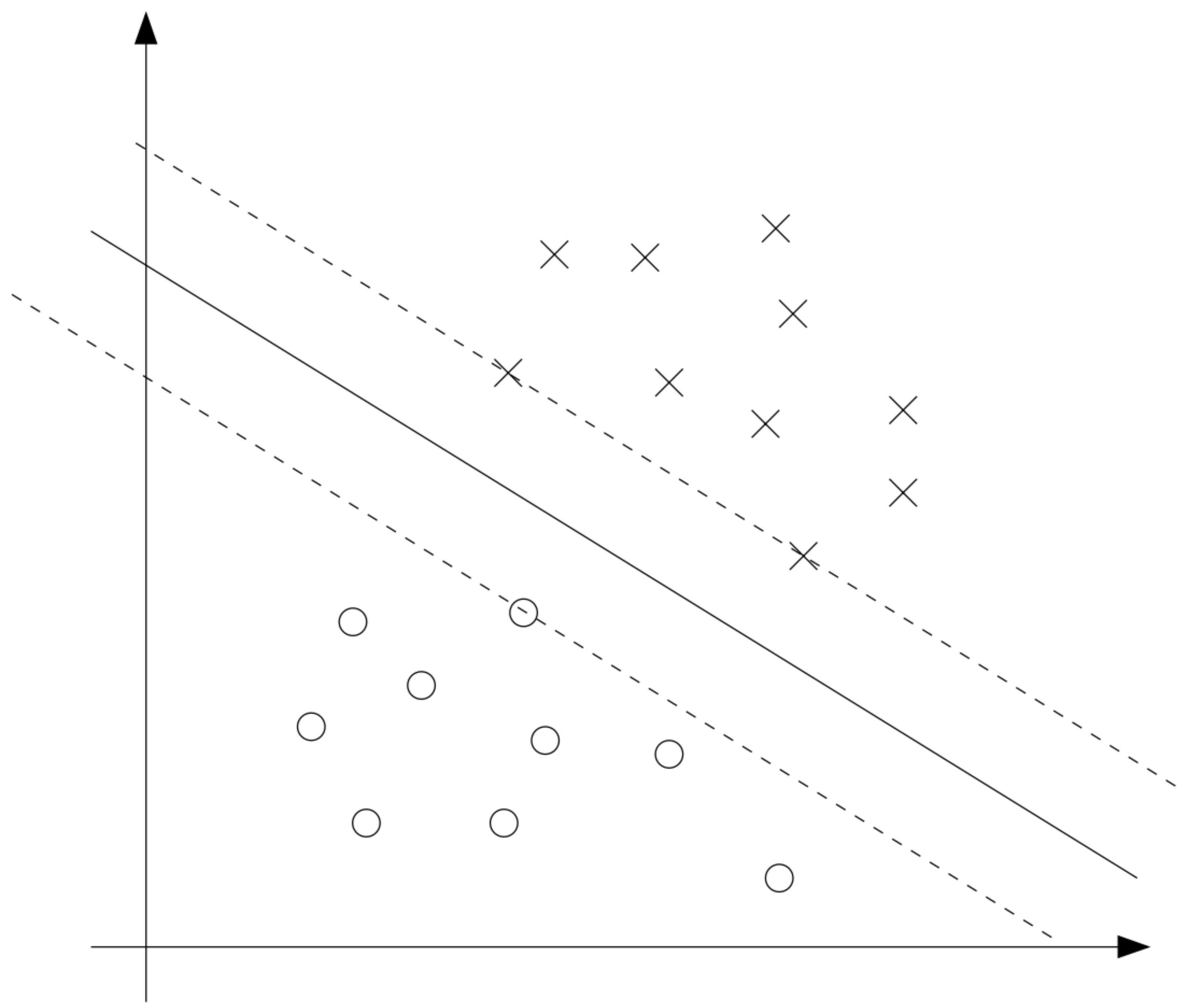
Supporting Vectors

Supporting Vectors

$$\alpha_i^* g_i(w^*) \quad = \quad 0, \quad i = 1, \ldots, k$$

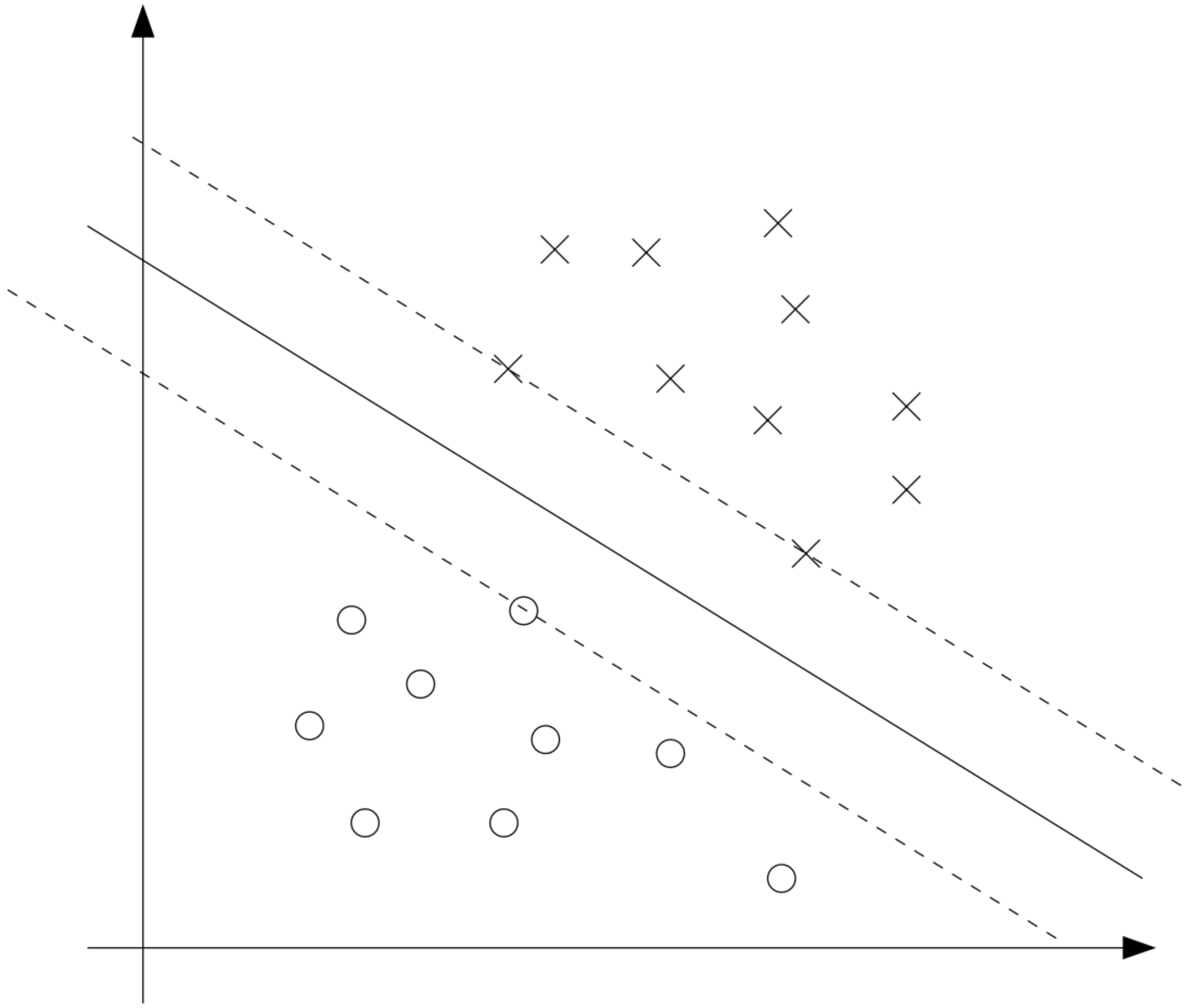
Supporting Vectors

$$\alpha_i^* g_i(w^*) \quad = \quad 0, \quad i = 1, \ldots, k$$



Supporting Vectors

$$\alpha_i^* g_i(w^*) \;=\; 0, \;\; i = 1, \ldots, k$$



Only the 3 points have non-zero α_i, and they are called supporting vectors

Lagrangian for SVM

Lagrangian for SVM

$$\mathcal{L}(w, b, \alpha) = \frac{1}{2}||w||^2 - \sum_{i=1}^{n} \alpha_i \left[y^{(i)}(w^T x^{(i)} + b) - 1 \right]$$

Lagrangian for SVM

$$\mathcal{L}(w, b, \alpha) = \frac{1}{2}||w||^2 - \sum_{i=1}^{n} \alpha_i \left[y^{(i)} (w^T x^{(i)} + b) - 1 \right]$$

The dual optimization problem

$$\max_{\alpha,\beta\,:\,\alpha_i \geq 0} \theta_{\mathcal{D}}(\alpha, \beta) = \max_{\alpha,\beta\,:\,\alpha_i \geq 0} \min_{w} \mathcal{L}(w, \alpha, \beta)$$

Lagrangian for SVM

$$\mathcal{L}(w, b, \alpha) = \frac{1}{2}||w||^2 - \sum_{i=1}^{n} \alpha_i \left[y^{(i)}(w^T x^{(i)} + b) - 1 \right]$$

The dual optimization problem

$$\max_{\alpha,\beta\,:\,\alpha_i \geq 0} \theta_{\mathcal{D}}(\alpha, \beta) = \max_{\alpha,\beta\,:\,\alpha_i \geq 0} \min_{w} \mathcal{L}(w, \alpha, \beta)$$

$$\nabla_w \mathcal{L}(w, b, \alpha) = w - \sum_{i=1}^{n} \alpha_i y^{(i)} x^{(i)} = 0$$

Lagrangian for SVM

$$\mathcal{L}(w,b,\alpha) = \frac{1}{2}||w||^2 - \sum_{i=1}^{n} \alpha_i \left[y^{(i)}(w^T x^{(i)} + b) - 1 \right]$$

The dual optimization problem

$$\max_{\alpha,\beta\,:\,\alpha_i \geq 0} \theta_{\mathcal{D}}(\alpha,\beta) = \max_{\alpha,\beta\,:\,\alpha_i \geq 0} \min_{w} \mathcal{L}(w,\alpha,\beta)$$

$$\nabla_w \mathcal{L}(w,b,\alpha) = w - \sum_{i=1}^{n} \alpha_i y^{(i)} x^{(i)} = 0 \qquad w = \sum_{i=1}^{n} \alpha_i y^{(i)} x^{(i)}$$

Lagrangian for SVM

$$\mathcal{L}(w,b,\alpha) = \frac{1}{2}||w||^2 - \sum_{i=1}^{n} \alpha_i \left[y^{(i)}(w^T x^{(i)} + b) - 1 \right]$$

The dual optimization problem

$$\max_{\alpha,\beta\,:\,\alpha_i \geq 0} \theta_{\mathcal{D}}(\alpha,\beta) = \max_{\alpha,\beta\,:\,\alpha_i \geq 0} \min_{w} \mathcal{L}(w,\alpha,\beta)$$

$$\nabla_w \mathcal{L}(w,b,\alpha) = w - \sum_{i=1}^{n} \alpha_i y^{(i)} x^{(i)} = 0 \qquad w = \sum_{i=1}^{n} \alpha_i y^{(i)} x^{(i)} \qquad \frac{\partial}{\partial b}\mathcal{L}(w,b,\alpha) = \sum_{i=1}^{n} \alpha_i y^{(i)} = 0$$

Lagrangian for SVM

$$\mathcal{L}(w,b,\alpha) = \frac{1}{2}||w||^2 - \sum_{i=1}^{n} \alpha_i \left[y^{(i)}(w^T x^{(i)} + b) - 1 \right]$$

The dual optimization problem

$$\max_{\alpha,\beta\,:\,\alpha_i \geq 0} \theta_{\mathcal{D}}(\alpha,\beta) = \max_{\alpha,\beta\,:\,\alpha_i \geq 0} \min_{w} \mathcal{L}(w,\alpha,\beta)$$

$$\nabla_w \mathcal{L}(w,b,\alpha) = w - \sum_{i=1}^{n} \alpha_i y^{(i)} x^{(i)} = 0 \qquad w = \sum_{i=1}^{n} \alpha_i y^{(i)} x^{(i)} \qquad \frac{\partial}{\partial b}\mathcal{L}(w,b,\alpha) = \sum_{i=1}^{n} \alpha_i y^{(i)} = 0$$

$$\theta(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y^{(i)} y^{(j)} \alpha_i \alpha_j (x^{(i)})^T x^{(j)}$$

The Dual Problem of SVM

The Dual Problem of SVM

$$
\begin{aligned}
\max_{\alpha} \quad & W(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y^{(i)} y^{(j)} \alpha_i \alpha_j \langle x^{(i)}, x^{(j)} \rangle \\
\text{s.t.} \quad & \alpha_i \geq 0, \quad i = 1, \ldots, n \\
& \sum_{i=1}^{n} \alpha_i y^{(i)} = 0,
\end{aligned}
$$

The Dual Problem of SVM

$$\max_\alpha \quad W(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y^{(i)} y^{(j)} \alpha_i \alpha_j \langle x^{(i)}, x^{(j)} \rangle$$

$$\text{s.t.} \quad \alpha_i \geq 0, \quad i = 1, \dots, n$$

$$\sum_{i=1}^{n} \alpha_i y^{(i)} = 0,$$

Kernel is all we need!

The Dual Problem of SVM

$$\max_\alpha \quad W(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y^{(i)} y^{(j)} \alpha_i \alpha_j \langle x^{(i)}, x^{(j)} \rangle$$

$$\text{s.t.} \quad \alpha_i \geq 0, \quad i = 1, \ldots, n$$

$$\sum_{i=1}^{n} \alpha_i y^{(i)} = 0,$$

Kernel is all we need!

After solving α (with standard quadratic optimization algorithms)

The Dual Problem of SVM

$$\max_\alpha \quad W(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y^{(i)} y^{(j)} \alpha_i \alpha_j \langle x^{(i)}, x^{(j)} \rangle$$

$$\text{s.t.} \quad \alpha_i \geq 0, \quad i = 1, \ldots, n$$

$$\sum_{i=1}^{n} \alpha_i y^{(i)} = 0,$$

Kernel is all we need!

After solving α (with standard quadratic optimization algorithms)

$$w = \sum_{i=1}^{n} \alpha_i y^{(i)} x^{(i)}$$

The Dual Problem of SVM

$$\max_\alpha \quad W(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y^{(i)} y^{(j)} \alpha_i \alpha_j \langle x^{(i)}, x^{(j)} \rangle$$

$$\text{s.t.} \quad \alpha_i \geq 0, \quad i = 1, \ldots, n$$

$$\sum_{i=1}^{n} \alpha_i y^{(i)} = 0,$$

Kernel is all we need!

After solving α (with standard quadratic optimization algorithms)

$$w = \sum_{i=1}^{n} \alpha_i y^{(i)} x^{(i)}$$

From KKT Conditions

The Dual Problem of SVM

$$\max_\alpha \quad W(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y^{(i)} y^{(j)} \alpha_i \alpha_j \langle x^{(i)}, x^{(j)} \rangle$$

$$\text{s.t.} \quad \alpha_i \geq 0, \quad i = 1, \ldots, n$$

$$\sum_{i=1}^{n} \alpha_i y^{(i)} = 0,$$

Kernel is all we need!

After solving α (with standard quadratic optimization algorithms)

$$w = \sum_{i=1}^{n} \alpha_i y^{(i)} x^{(i)} \qquad b^* = -\frac{\max_{i:y^{(i)}=-1} {w^*}^T x^{(i)} + \min_{i:y^{(i)}=1} {w^*}^T x^{(i)}}{2}$$

From KKT Conditions

The Dual Problem of SVM

$$\max_\alpha \quad W(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y^{(i)} y^{(j)} \alpha_i \alpha_j \langle x^{(i)}, x^{(j)} \rangle$$

$$\text{s.t.} \quad \alpha_i \geq 0, \quad i = 1, \ldots, n$$

$$\sum_{i=1}^{n} \alpha_i y^{(i)} = 0,$$

Kernel is all we need!

After solving α (with standard quadratic optimization algorithms)

$$w = \sum_{i=1}^{n} \alpha_i y^{(i)} x^{(i)}$$

From KKT Conditions

$$b^* = -\frac{\max_{i:y^{(i)}=-1} w^{*T} x^{(i)} + \min_{i:y^{(i)}=1} w^{*T} x^{(i)}}{2}$$

From the original constraints

Inference

Inference

$$
\begin{aligned}
w^T x + b &= \left(\sum_{i=1}^{n} \alpha_i y^{(i)} x^{(i)} \right)^T x + b \\
&= \sum_{i=1}^{n} \alpha_i y^{(i)} \langle x^{(i)}, x \rangle + b.
\end{aligned}
$$

Inference

$$
\begin{aligned}
w^T x + b &= \left(\sum_{i=1}^{n} \alpha_i y^{(i)} x^{(i)}\right)^T x + b \\
&= \sum_{i=1}^{n} \alpha_i y^{(i)} \langle x^{(i)}, x \rangle + b.
\end{aligned}
$$

We never need to really compute w

Inference

$$
\begin{aligned}
w^T x + b &= \left(\sum_{i=1}^{n} \alpha_i y^{(i)} x^{(i)} \right)^T x + b \\
&= \sum_{i=1}^{n} \alpha_i y^{(i)} \langle x^{(i)}, x \rangle + b.
\end{aligned}
$$

We never need to really compute w

$$
\alpha_i^* g_i(w^*) \quad = \quad 0, \quad i = 1, \ldots, k
$$

Inference

$$w^T x + b = \left(\sum_{i=1}^{n} \alpha_i y^{(i)} x^{(i)}\right)^T x + b$$
$$= \sum_{i=1}^{n} \alpha_i y^{(i)} \langle x^{(i)}, x \rangle + b.$$

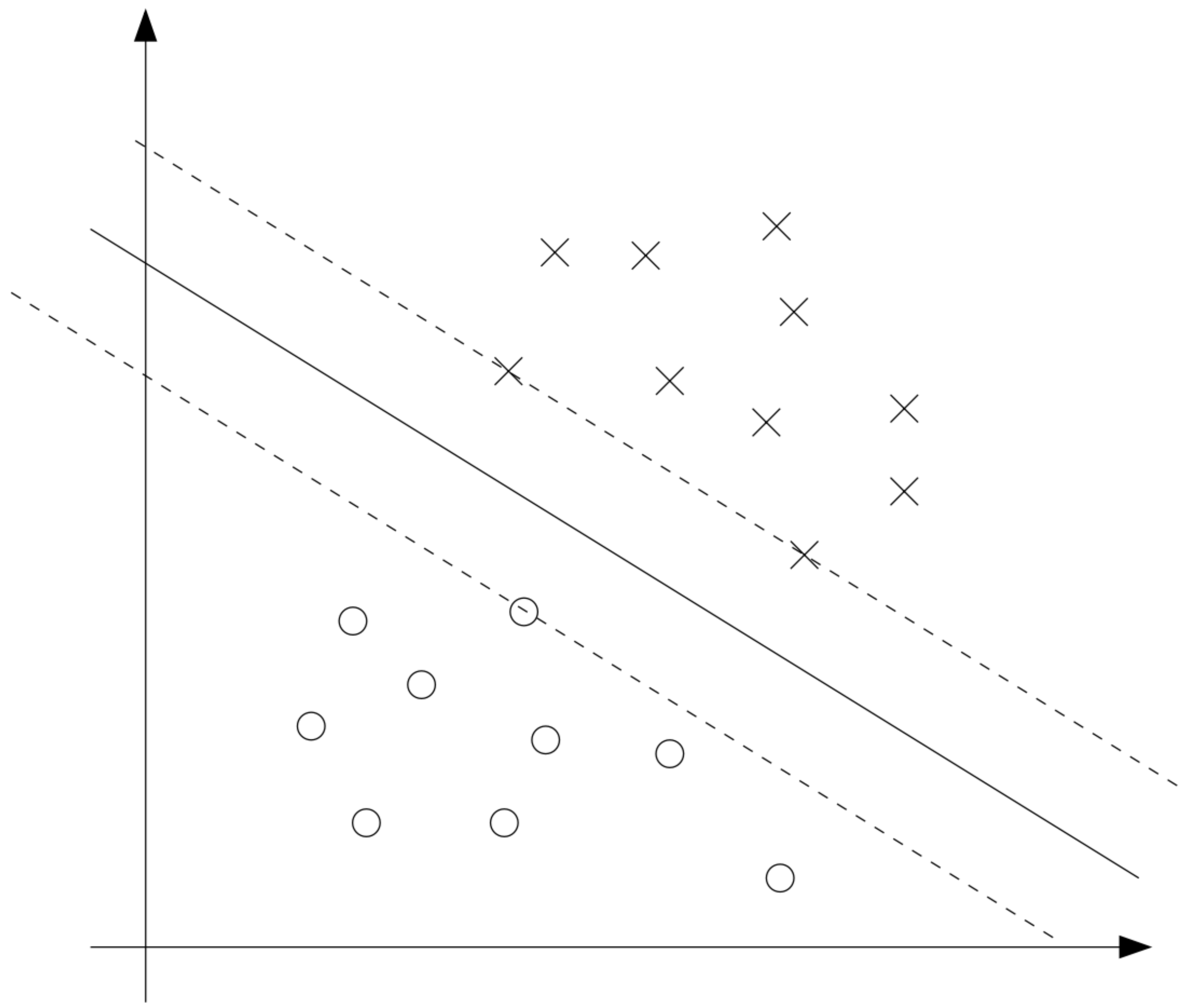
We never need to really compute w

$$\alpha_i^* g_i(w^*) = 0, \quad i = 1, \ldots, k$$

Most α_i are 0, only the supporting examples will influence the final prediction

Inference

$$w^T x + b = \left(\sum_{i=1}^{n} \alpha_i y^{(i)} x^{(i)}\right)^T x + b$$
$$= \sum_{i=1}^{n} \alpha_i y^{(i)} \langle x^{(i)}, x \rangle + b.$$

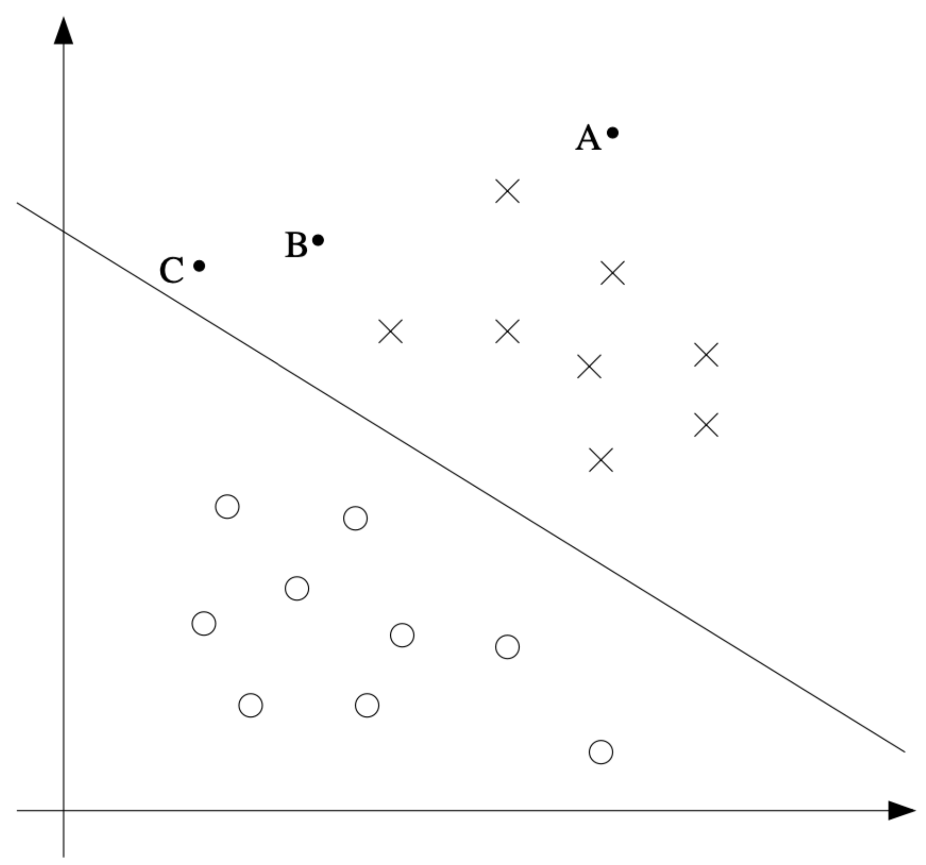


We never need to really compute w

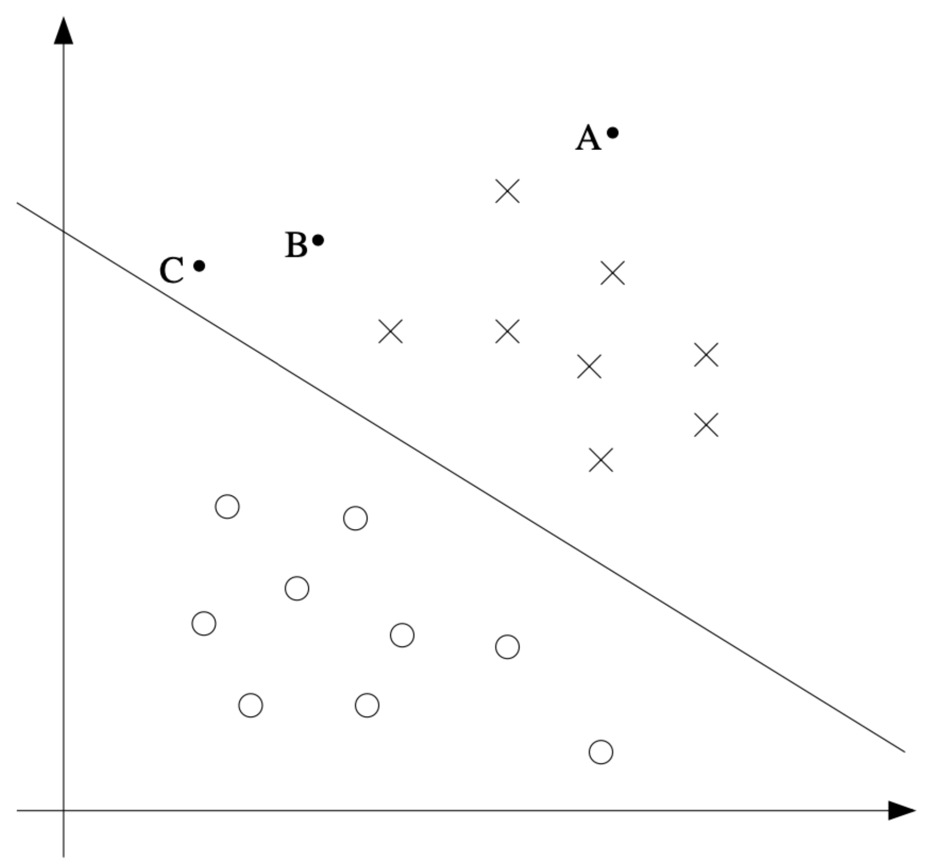
$$\alpha_i^* g_i(w^*) = 0, \quad i = 1, \ldots, k$$

Most α_i are 0, only the supporting examples will influence the final prediction

Review of the High-Level Logic

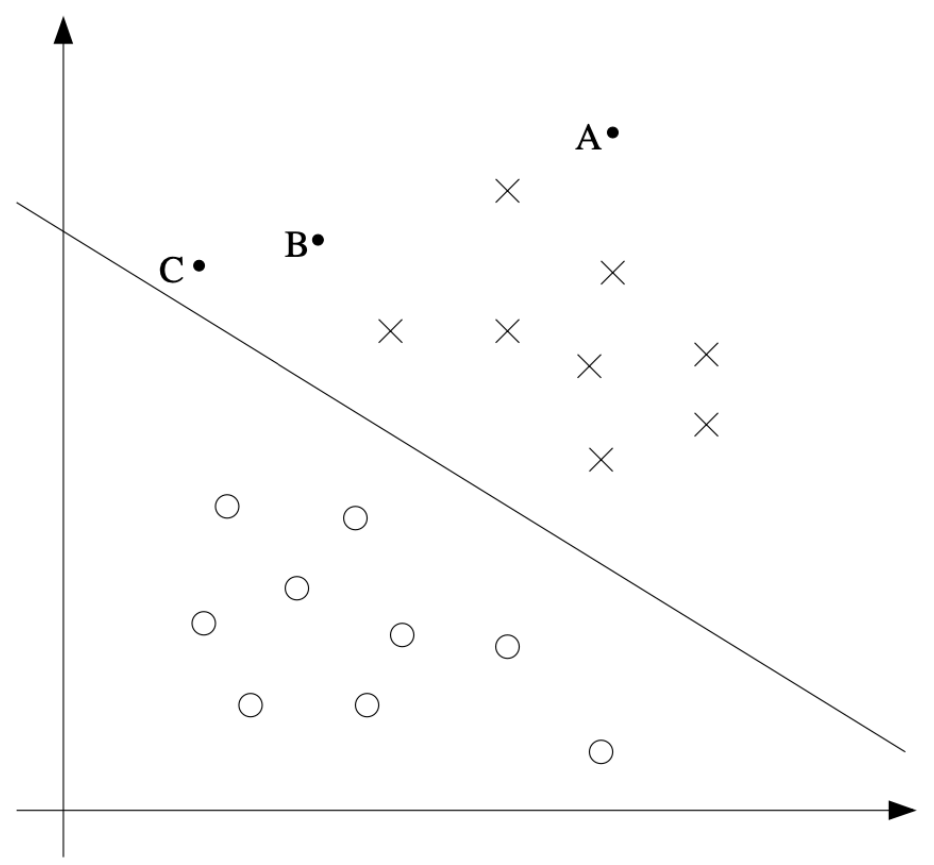


Review of the High-Level Logic



Review of the High-Level Logic

$$h_{w,b}(x) = g(w^T x + b).$$

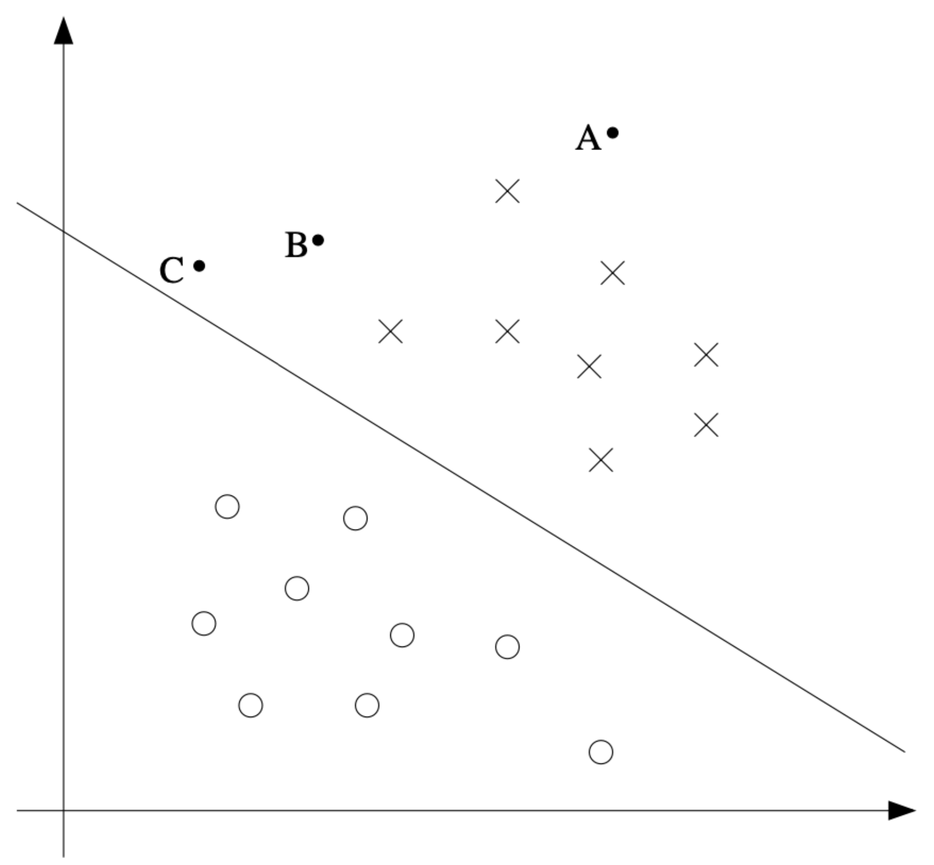


Review of the High-Level Logic

$$h_{w,b}(x) = g(w^T x + b).$$

Maximize geometric margin

$$\gamma^{(i)} = y^{(i)} \left(\left(\frac{w}{||w||} \right)^T x^{(i)} + \frac{b}{||w||} \right)$$



Review of the High-Level Logic

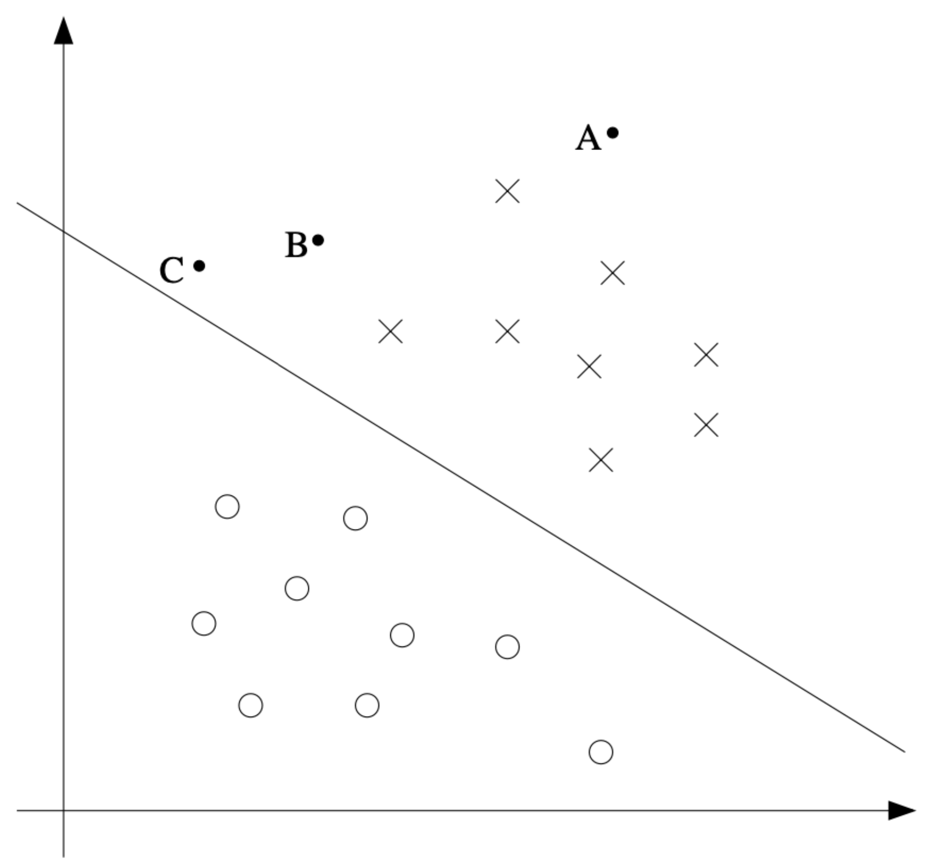
$$h_{w,b}(x) = g(w^T x + b).$$

Maximize geometric margin

Problem rewriting →

$$\gamma^{(i)} = y^{(i)} \left(\left(\frac{w}{||w||} \right)^T x^{(i)} + \frac{b}{||w||} \right)$$

Review of the High-Level Logic

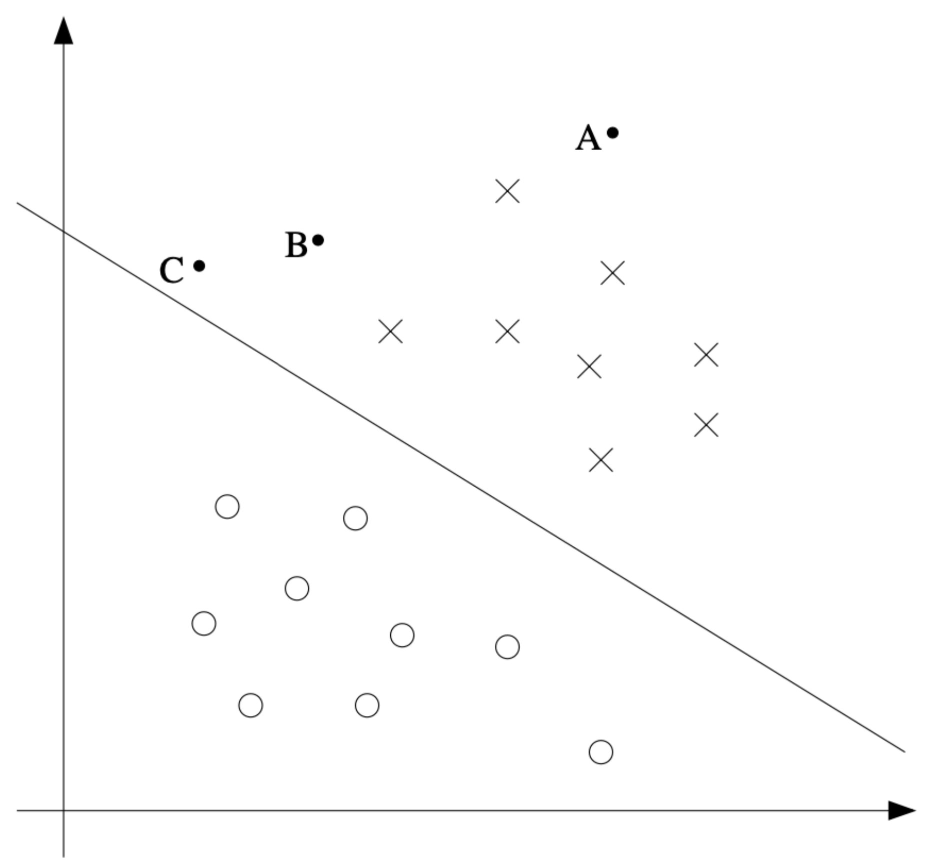


$$h_{w,b}(x) = g(w^T x + b).$$

Maximize geometric margin → (Problem rewriting) → **Quadratic Optimization Problem**

$$\gamma^{(i)} = y^{(i)} \left(\left(\frac{w}{||w||} \right)^T x^{(i)} + \frac{b}{||w||} \right)$$

$$\begin{aligned} \min_{w,b} \quad & \frac{1}{2}||w||^2 \\ \text{s.t.} \quad & y^{(i)}(w^T x^{(i)} + b) \geq 1, \quad i = 1, \ldots, n \end{aligned}$$



Review of the High-Level Logic

$$h_{w,b}(x) = g(w^T x + b).$$

Maximize geometric margin

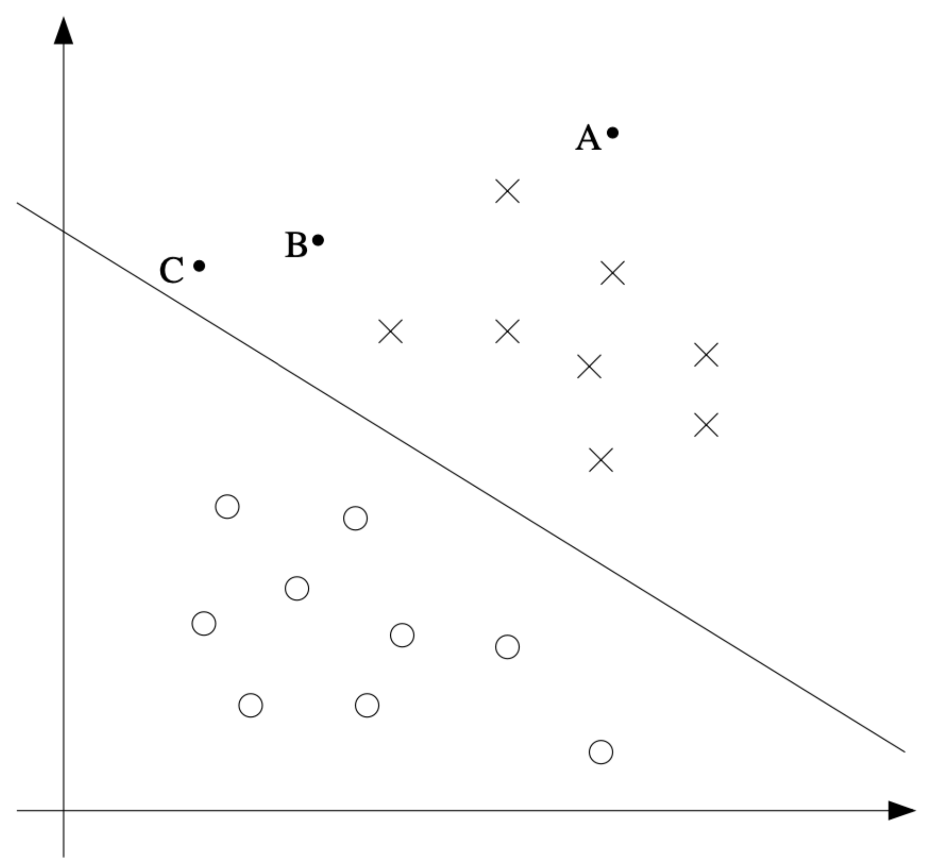
Problem rewriting →

Quadratic Optimization Problem

$$\gamma^{(i)} = y^{(i)} \left(\left(\frac{w}{||w||} \right)^T x^{(i)} + \frac{b}{||w||} \right)$$

$$\begin{aligned} \min_{w,b} \quad & \frac{1}{2}||w||^2 \\ \text{s.t.} \quad & y^{(i)}(w^T x^{(i)} + b) \geq 1, \quad i = 1, \ldots, n \end{aligned}$$

Not suitable for non-linear cases (high-dim feature map)



Review of the High-Level Logic

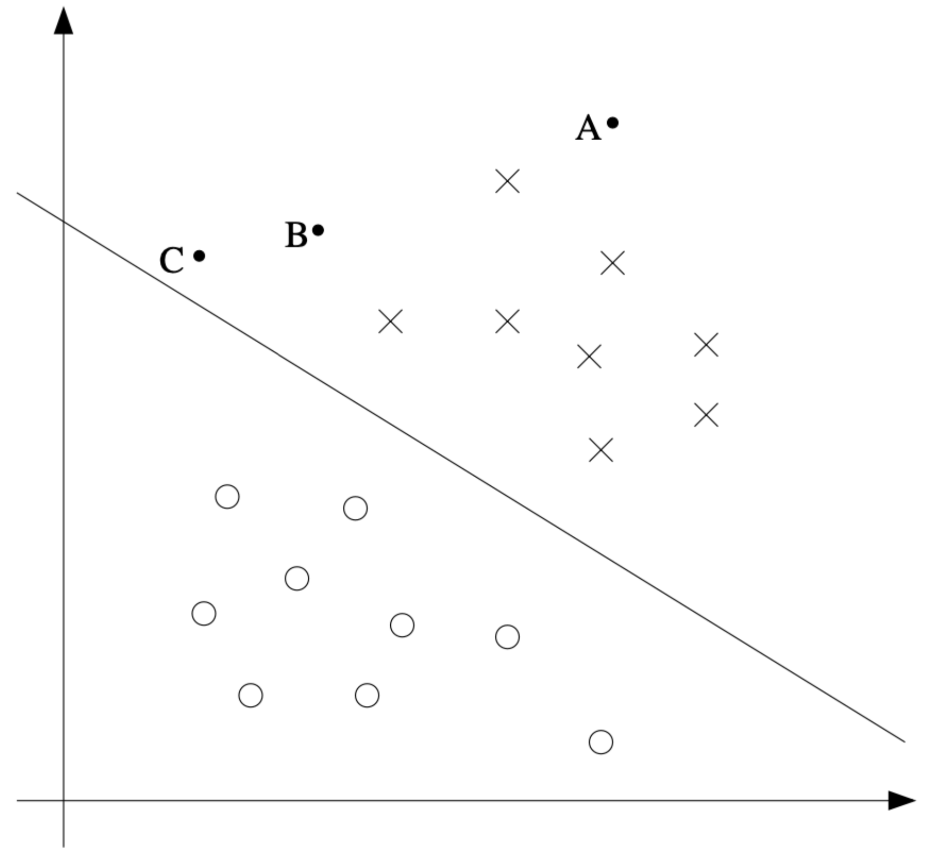
$$h_{w,b}(x) = g(w^T x + b).$$

Maximize geometric margin → (Problem rewriting) → Quadratic Optimization Problem → (Finding a related optimization problem that is easier)

$$\gamma^{(i)} = y^{(i)} \left(\left(\frac{w}{||w||} \right)^T x^{(i)} + \frac{b}{||w||} \right)$$

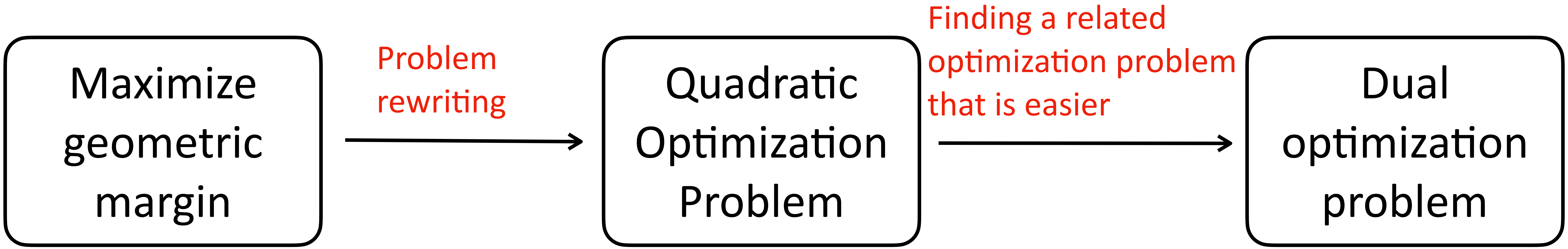
$$\begin{aligned} \min_{w,b} \quad & \frac{1}{2}||w||^2 \\ \text{s.t.} \quad & y^{(i)}(w^T x^{(i)} + b) \geq 1, \quad i = 1, \ldots, n \end{aligned}$$

Not suitable for non-linear cases (high-dim feature map)



Review of the High-Level Logic

$$h_{w,b}(x) = g(w^T x + b).$$

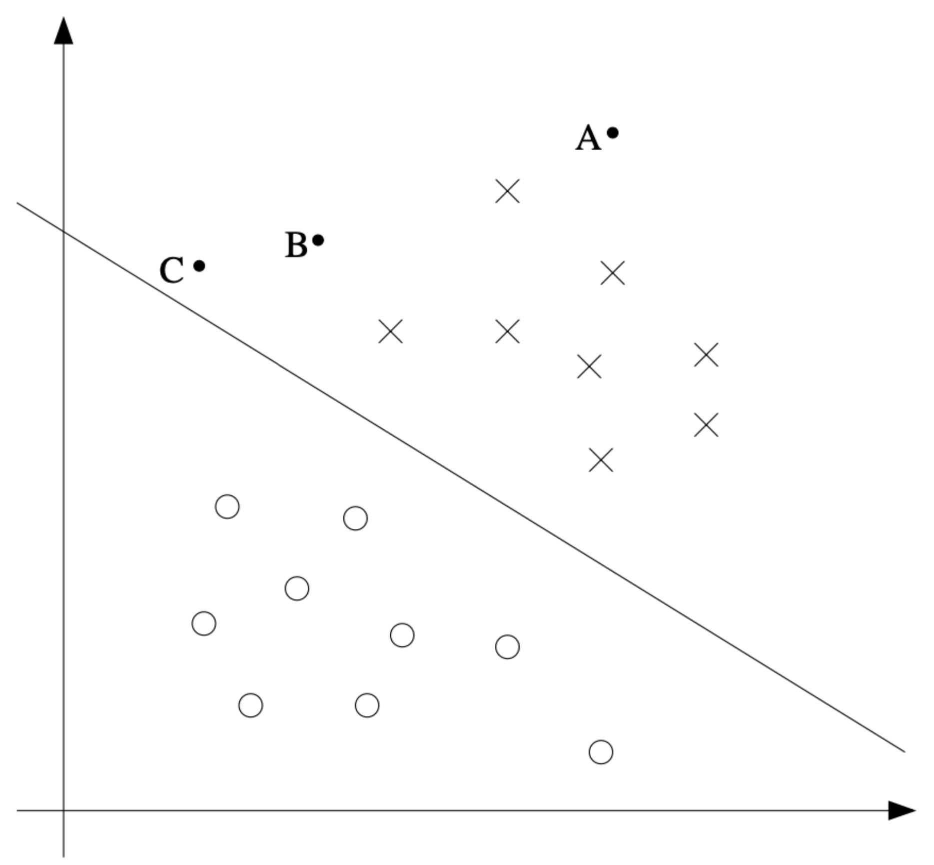


$$\gamma^{(i)} = y^{(i)} \left(\left(\frac{w}{||w||} \right)^T x^{(i)} + \frac{b}{||w||} \right)$$

$$\begin{aligned} \min_{w,b} \quad & \frac{1}{2}||w||^2 \\ \text{s.t.} \quad & y^{(i)}(w^T x^{(i)} + b) \geq 1, \quad i = 1, \ldots, n \end{aligned}$$

$$\begin{aligned} \max_\alpha \quad & W(\alpha) = \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j=1}^n y^{(i)} y^{(j)} \alpha_i \alpha_j \langle x^{(i)}, x^{(j)} \rangle \\ \text{s.t.} \quad & \alpha_i \geq 0, \quad i = 1, \ldots, n \\ & \sum_{i=1}^n \alpha_i y^{(i)} = 0, \end{aligned}$$

Not suitable for non-linear cases (high-dim feature map)



Review of the High-Level Logic

$$h_{w,b}(x) = g(w^T x + b)$$

Maximize geometric margin → (Problem rewriting) → **Quadratic Optimization Problem** → (Finding a related optimization problem that is easier) → **Dual optimization problem**

$$\gamma^{(i)} = y^{(i)} \left(\left(\frac{w}{||w||} \right)^T x^{(i)} + \frac{b}{||w||} \right)$$

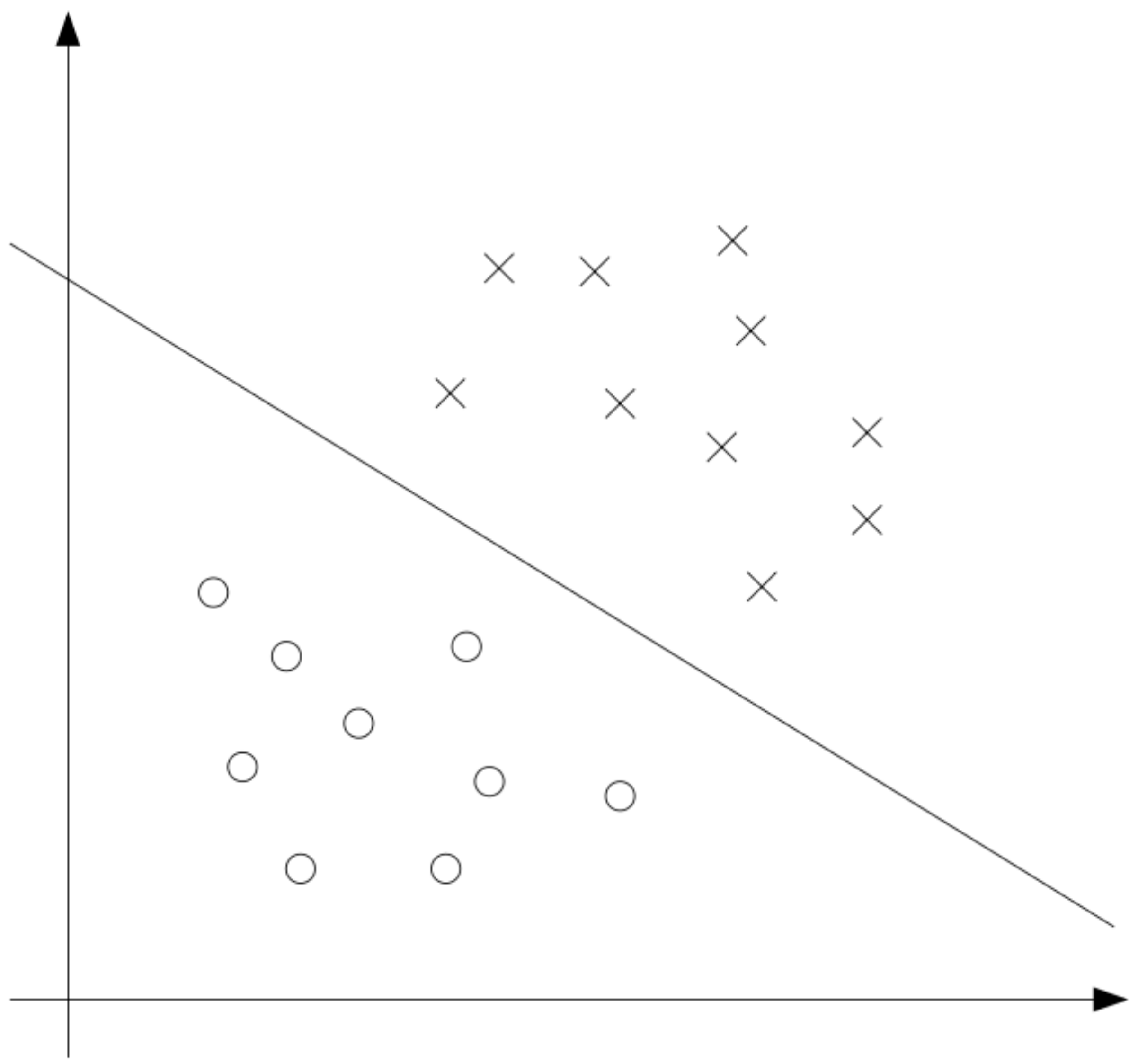
$$\begin{aligned} \min_{w,b} \quad & \frac{1}{2}||w||^2 \\ \text{s.t.} \quad & y^{(i)}(w^T x^{(i)} + b) \geq 1, \quad i = 1, \ldots, n \end{aligned}$$

Not suitable for non-linear cases (high-dim feature map)

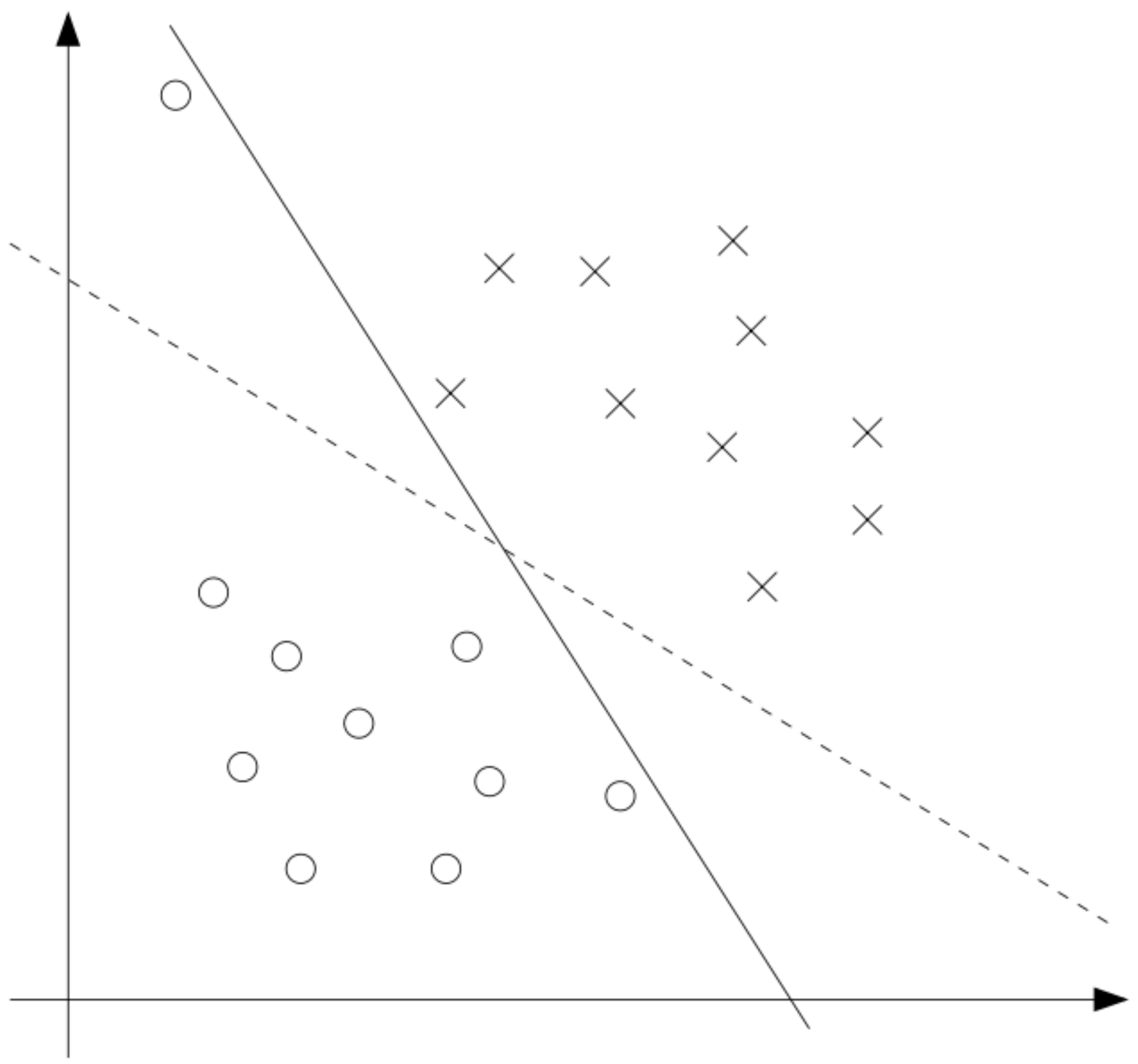
$$\begin{aligned} \max_\alpha \quad & W(\alpha) = \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j=1}^n y^{(i)} y^{(j)} \alpha_i \alpha_j \langle x^{(i)}, x^{(j)} \rangle \\ \text{s.t.} \quad & \alpha_i \geq 0, \quad i = 1, \ldots, n \\ & \sum_{i=1}^n \alpha_i y^{(i)} = 0, \end{aligned}$$

Kernel makes it very flexible in non-linear cases!

The Non-Separable Case



Linearly Separable



Linearly Non-Separable

The Non-Separable Case

Primal opt problem:

$$\begin{aligned} \min_{\gamma,w,b} \quad & \frac{1}{2}||w||^2 + C\sum_{i=1}^{n}\xi_i \\ \text{s.t.} \quad & y^{(i)}(w^T x^{(i)} + b) \geq 1 - \xi_i, \quad i = 1,\ldots,n \\ & \xi_i \geq 0, \quad i = 1,\ldots,n. \end{aligned}$$

Dual opt problem

$$\begin{aligned} \max_{\alpha} \quad & W(\alpha) = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} y^{(i)}y^{(j)}\alpha_i\alpha_j\langle x^{(i)}, x^{(j)}\rangle \\ \text{s.t.} \quad & 0 \leq \alpha_i \leq C, \quad i = 1,\ldots,n \\ & \sum_{i=1}^{n}\alpha_i y^{(i)} = 0, \end{aligned}$$

Thank You!

Q & A