



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

COMP 5212
Machine Learning
Lecture 7

Support Vector Machines

Junxian He
Feb 21, 2024

HW1 is Out on Canvas

- Start Early
- Due in 2 weeks on Friday
- No submissions will be accepted more than 3 days late

Update of the Website

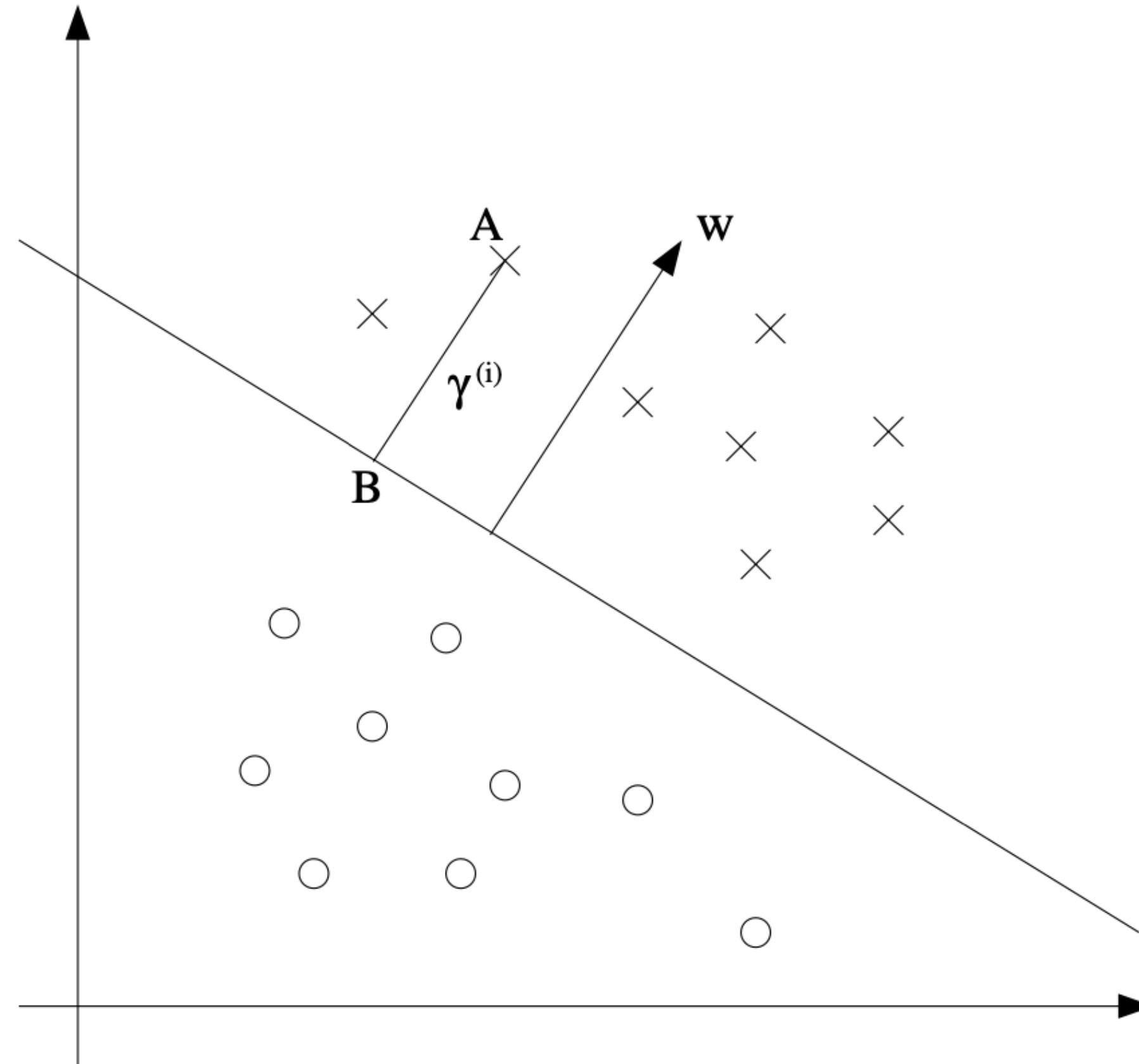
Lecture Schedule

The lecture schedule below is tentative and subject to change.

Slides	Date	Topic	Readings	Assignments
Lecture 0	31/01 Wed	Introduction		
Lecture 1	02/02 Fri	Math basics		
Lecture 2, draft2	07/02 Wed	Linear Regression		
Lecture 3, draft3	09/02 Fri	Logistic regression, Exponential Family		
Lecture 4, draft4	14/02 Wed	Generalized linear models, Kernel Methods	Section 3 of Notes	
Lecture 5, draft5	16/02 Fri	Kernel methods, SVM	Section 5 of Notes	
Lecture 6, draft6	21/02 Wed	SVM	Section 6 of Notes	

The Stanford CS229 Notes by Andrew Ng is the most important reading material
I will post the lecture handwritings after each lecture (very unstructured)

Recap: Support Vector Machines



Recap: The Optimization Problem

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

s.t. $y^{(i)} \left(\left(\frac{w}{\|w\|} \right)^T x^{(i)} + \frac{b}{\|w\|} \right) \geq \gamma, \quad i = 1, \dots, n$

Linear constraint

$$\xrightarrow{\text{Linear constraint}} \max_{\hat{\gamma},w,b} \frac{\hat{\gamma}}{\|w\|}$$

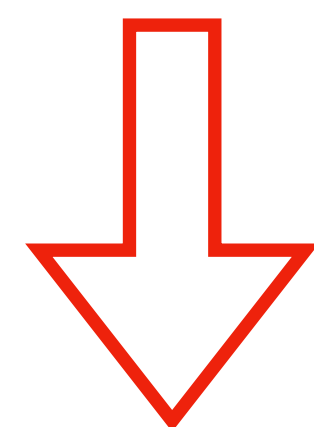
s.t. $y^{(i)} (w^T x^{(i)} + b) \geq \hat{\gamma}, \quad i = 1, \dots, n$

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

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

Recap: The Optimization Problem

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



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

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

This is a standard quadratic problem that can be directly solved with quadratic problem solvers

Assumption: the training dataset is linearly separable

Lagrange Duality — Lagrange Multiplier

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

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

Solve w, β

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

Lagrange Multiplier: Example

$$\begin{aligned} & \min_{x,y} 5x - 3y \\ \text{s.t. } & x^2 + y^2 = 136 \end{aligned}$$

Recap: Generalized Lagrangian

Primal optimization problem

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

Generalized Lagrangian

This is a general concept in optimization, beyond SVMs

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

Recap: Generalized Lagrangian

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

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

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

Recap: 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

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

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 original problem in an indirect way

The Dual Problem

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

The dual optimization problem

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

These are a general concepts in optimization, beyond SVMs

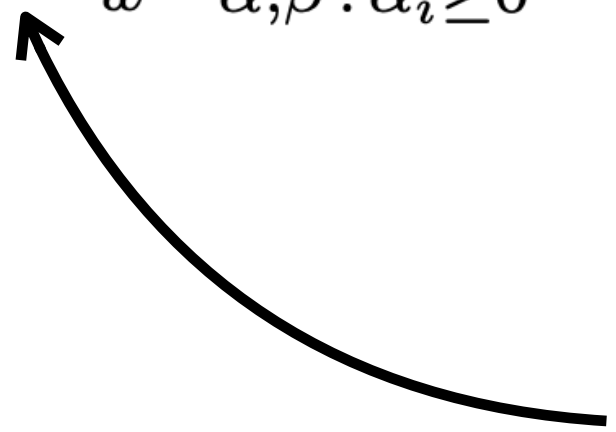
The primal optimization problem

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

What is the relation of the two problems?

The Dual Problem

$$d^* = \max_{\alpha, \beta: \alpha_i \geq 0} \min_w \mathcal{L}(w, \alpha, \beta) \leq \min_w \max_{\alpha, \beta: \alpha_i \geq 0} \mathcal{L}(w, \alpha, \beta) = p^*$$

$$\max_x \min_y f(x, y) \leq \min_y \max_x f(x, y)$$


Under certain conditions: $d^* = p^*$ Zero-duality Gap (Strong Duality)

What are the conditions?

Slater's Condition

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

- $f(w)$ and $g(w)$ are convex
- $h_i(w)$ is affine (i.e. linear)
- $g_i(w)$ are strictly feasible for all i , which means there exists some w so that $g_i(w) < 0$ for all i

If Slater's condition holds, then $d^* = p^*$

The primal optimization problem of SVM satisfies the Slater's condition

KKT Conditions

Zero duality gap is sufficient and necessary (i.e. equivalent) to satisfy KKT Conditions:

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

$$\frac{\partial}{\partial w_i} \mathcal{L}(w^*, \alpha^*, \beta^*) = 0, \quad i = 1, \dots, d$$

$$\frac{\partial}{\partial \beta_i} \mathcal{L}(w^*, \alpha^*, \beta^*) = 0, \quad i = 1, \dots, l$$

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

$$g_i(w^*) \leq 0, \quad i = 1, \dots, k$$

$$\alpha^* \geq 0, \quad i = 1, \dots, k$$

Normal Lagrange multiplier equations

The original constraints

KKT Conditions

Zero duality gap is sufficient and necessary (i.e. equivalent) to satisfy KKT Conditions:

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

$$\frac{\partial}{\partial w_i} \mathcal{L}(w^*, \alpha^*, \beta^*) = 0, \quad i = 1, \dots, d$$

$$\frac{\partial}{\partial \beta_i} \mathcal{L}(w^*, \alpha^*, \beta^*) = 0, \quad i = 1, \dots, l$$

If $\alpha_i^* > 0$, then

$g_i(w^*) = 0$, the inequality is actually equality

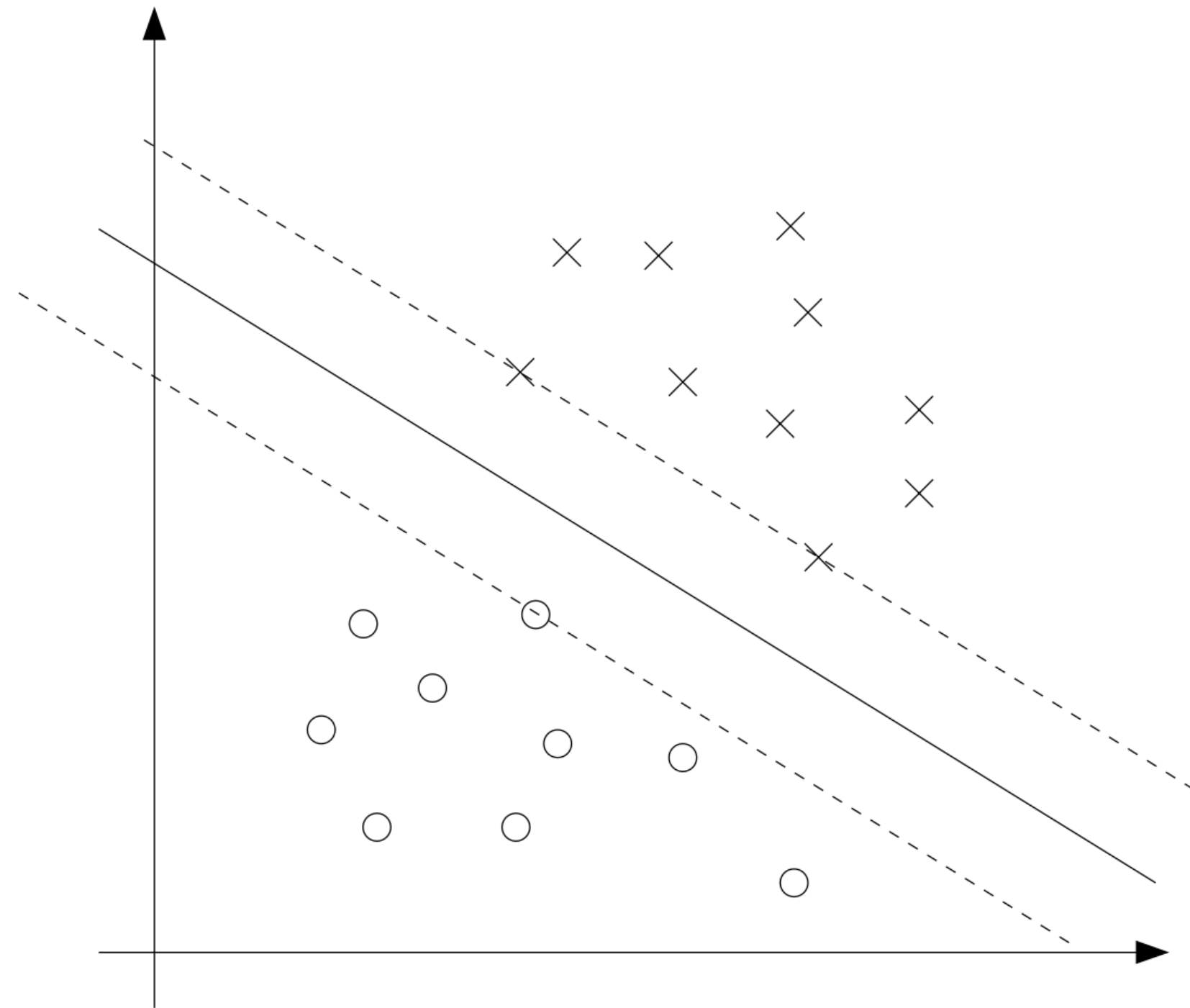
$$\alpha_i^* g_i(w^*) = 0, \quad i = 1, \dots, k$$

$$g_i(w^*) \leq 0, \quad i = 1, \dots, k$$

$$\alpha_i^* \geq 0, \quad i = 1, \dots, k$$

Supporting Vectors

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



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

Lagrangian for SVM

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

The dual optimization problem

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

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

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

The Dual Problem of SVM

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

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

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

Kernel is all we need!

After solving α (coordinate ascent with clipping, 6.8.2 of the CS229 Notes)

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

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

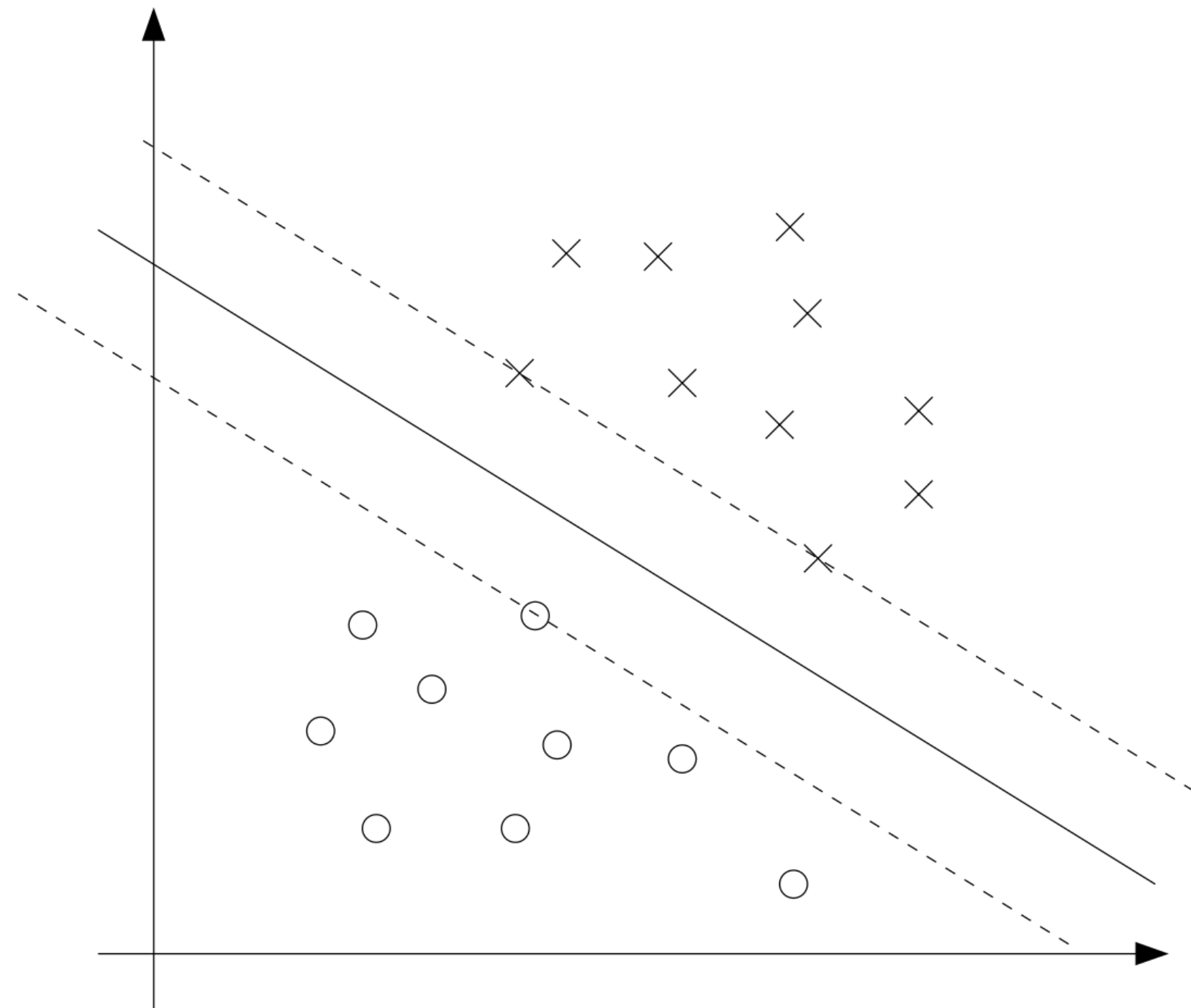
Inference

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

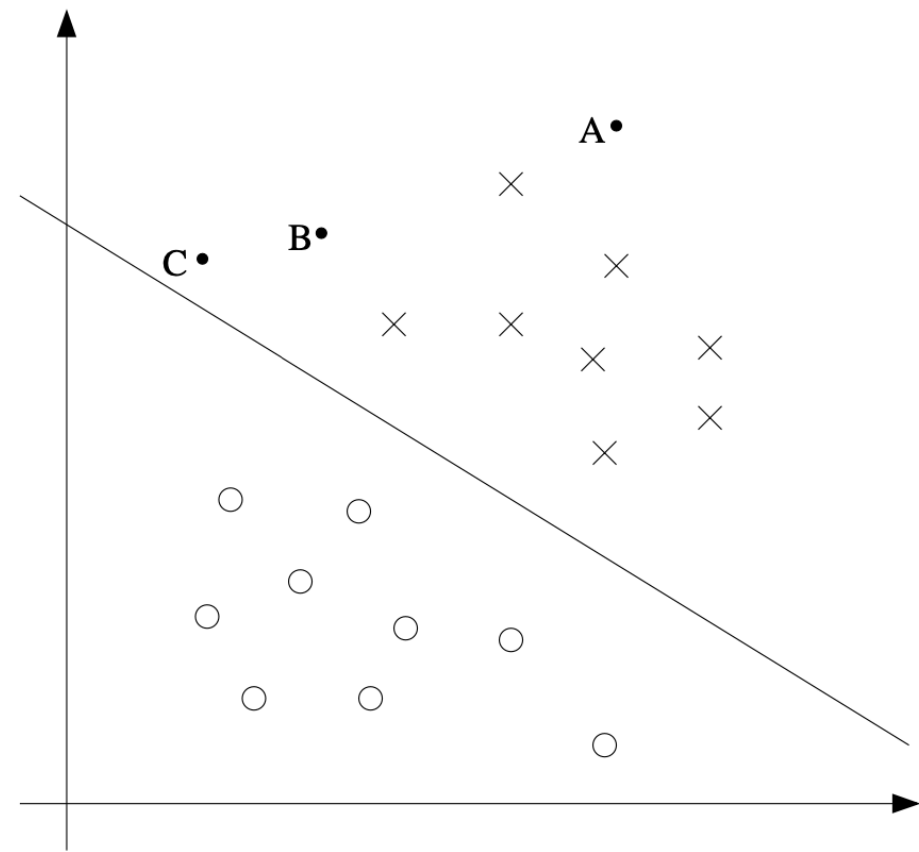
We never need to really compute w

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

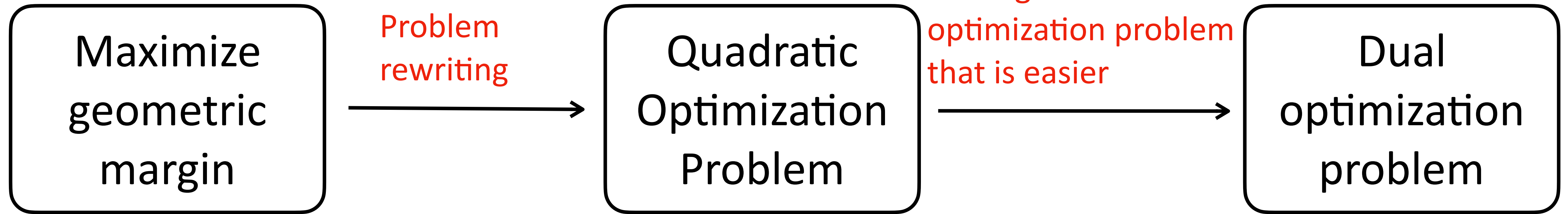
Most α_i are 0, only the supporting examples will influence the final prediction



Review of the High-Level Logic



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



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

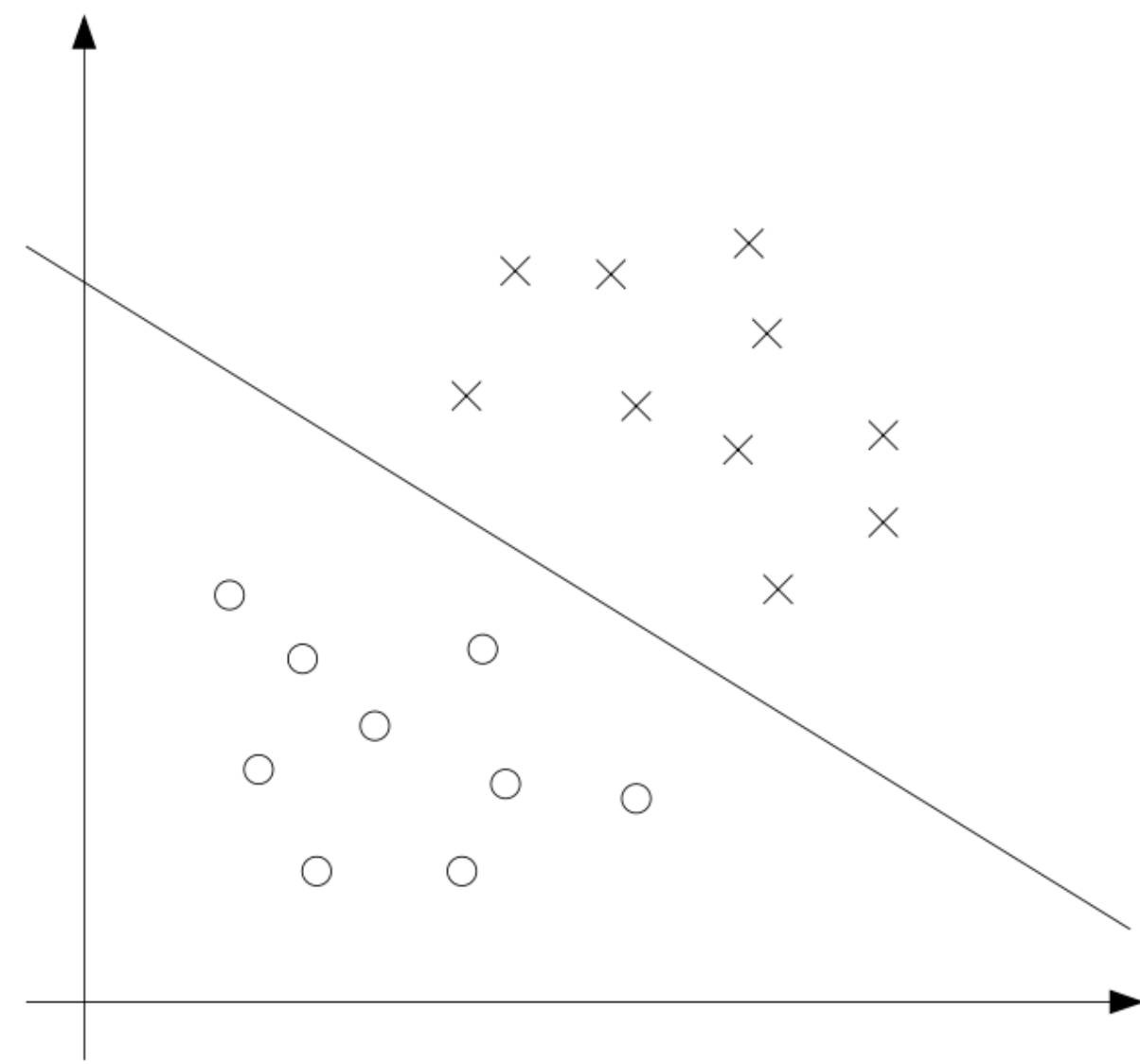
$$\begin{aligned} \min_{w,b} \quad & \frac{1}{2} \|w\|^2 \\ \text{s.t.} \quad & y^{(i)}(w^T x^{(i)} + b) \geq 1, \quad i = 1, \dots, n \end{aligned}$$

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

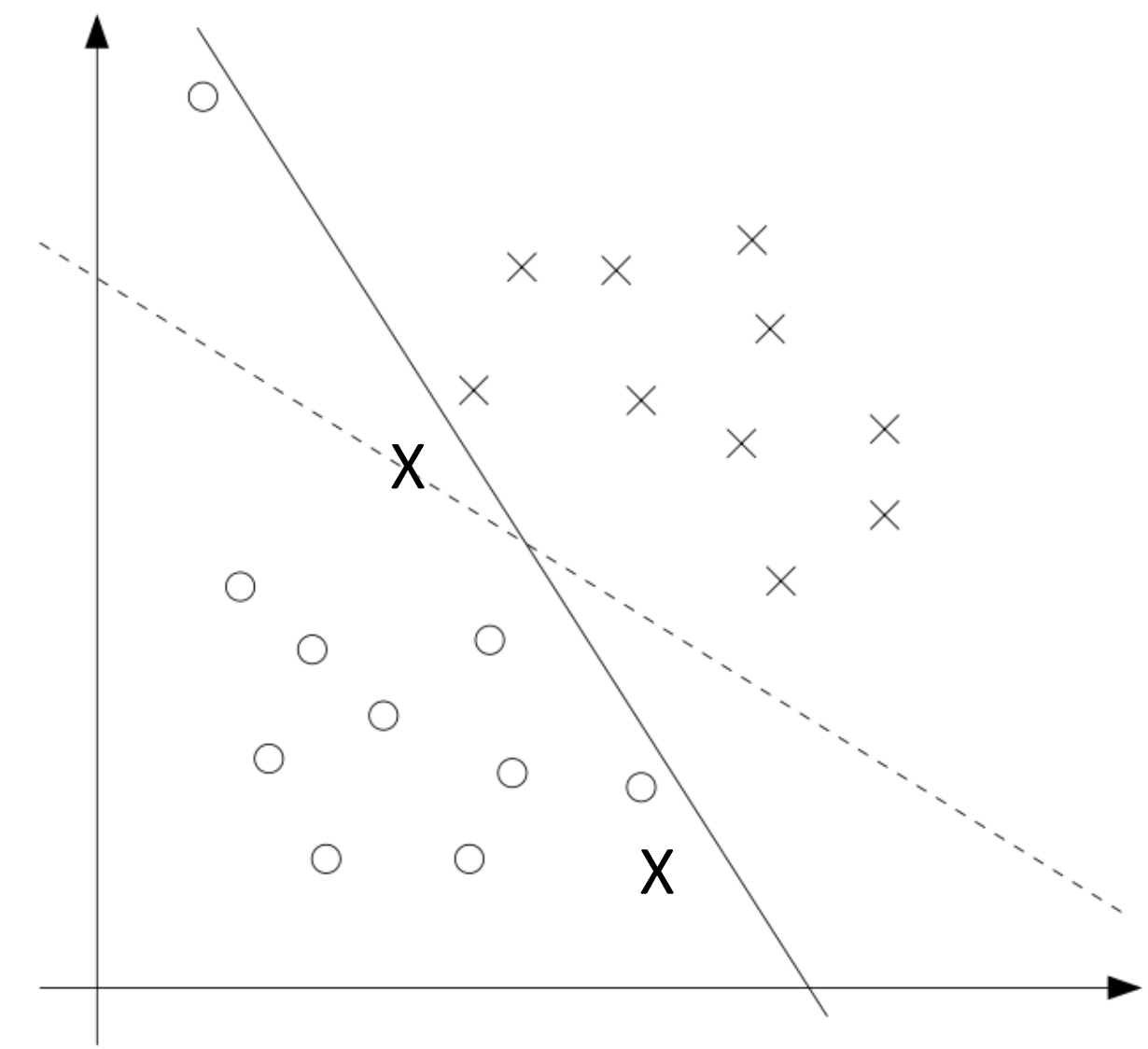
$$\begin{aligned} \max_{\alpha} \quad & W(\alpha) = \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j=1}^n y^{(i)} y^{(j)} \alpha_i \alpha_j \langle x^{(i)}, x^{(j)} \rangle \\ \text{s.t.} \quad & \alpha_i \geq 0, \quad i = 1, \dots, n \\ & \sum_{i=1}^n \alpha_i y^{(i)} = 0, \end{aligned}$$

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:

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

Dual opt problem

You will prove this in your hw

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

Thank You!
Q & A