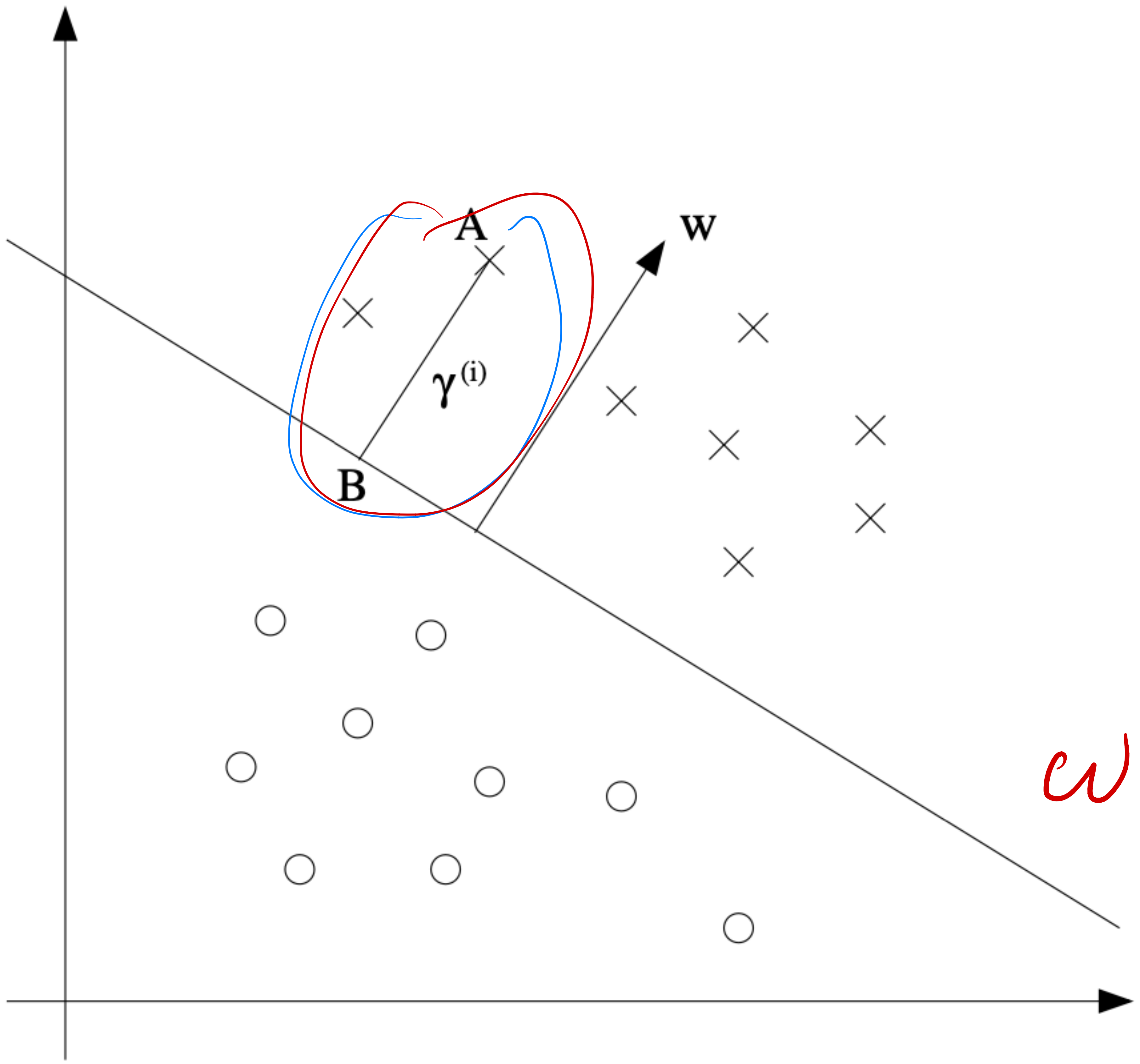
香港科技大學
THE HONG KONG
UNIVERSITY OF SCIENCE
AND TECHNOLOGY

COMP 5212
Machine Learning
Lecture 7

Support Vector Machines

Junxian He
Sep 26, 2024

Recap: Support Vector Machines



w, b

$w^T x + b = 0$

$w \to kw$

$b \to kw$

$k\, w^T x + kb = 0$

Recap: The Optimization Problem

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

Rewrite

$\max_{\gamma,w,b} \ \gamma$

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

$||w||$

Linear constraint

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

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

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

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

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

2, 3

$\alpha^* \rightarrow w^*$

?

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

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

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

not Kernel

Assumption: the training dataset is linearly separable

$y^{(i)}(w^T x^{(i)} + b) > 0$ only if prediction correct

Recap: The Dual Problem

$L(w, \alpha, \beta) = \frac{1}{2}\|w\|^2 + \alpha \cdot g(w) + \beta h(w)$

inequaly

equaliy

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

The dual optimization problem

α, β

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

$\boxed{\min_w L(w, \alpha, \beta)}$

w

The primal optimization problem

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

$= \begin{cases} f(w) & \text{if constraints are satisfied} \\ \infty & \text{otherwise} \end{cases}$

What is the relation of the two problems?

Recap: The Dual Problem

remove w, b.

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

The dual optimization problem

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

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

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

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

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

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

$$\sum_i \alpha_i y^{(i)} = 0$$

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

$$-\sum_i \alpha_i y^{(i)} b = 0$$

Recap: The Dual Problem

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

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

$\langle x^{(i)}, x^{(j)} \rangle$

$K(x^{(i)}, x^{(j)})$

$K(x^{(i)}, x^{(j)})$

$\Rightarrow \alpha^*$

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

What is the relation between solving this dual problem and solving the original problem

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

The Dual Problem

dual

primal

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

$$\boxed{\max_x \min_y f(x,y) \leq \min_y \max_x f(x,y)}$$

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

What are the conditions?

$$\boxed{\max_{x} \min_{y} f(x,y) \le \min_{y} \max_{x} f(x,y)}$$

$$\Downarrow \qquad\qquad\qquad\qquad \Downarrow$$

$$(x_0, y_0) \qquad\qquad\qquad\qquad (x_1, y_1)$$

$$f(x, y_0) \le f(x, y) \quad \text{for any } x, y$$

$$\rightleftharpoons$$

$$f(x_1, y) \ge f(x, y) \quad \text{for any } x, y$$

$$f(x_0, y_0) \le f(x_0, y_1) \le f(x_1, y_1)$$

Slater's Condition

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

$g_i(w) \le 0$

$\frac{1}{2}\|w\|^2$ $g(w)$ linear

- 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

no $h_i(w)$ for SVM

$1 - y^{(i)}(w^T x^{(i)} + b) \le 0$

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

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

Slaters condition ⟹ zero duality gap ⟺ KKT

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:

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

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

Normal Lagrange multiplier equations

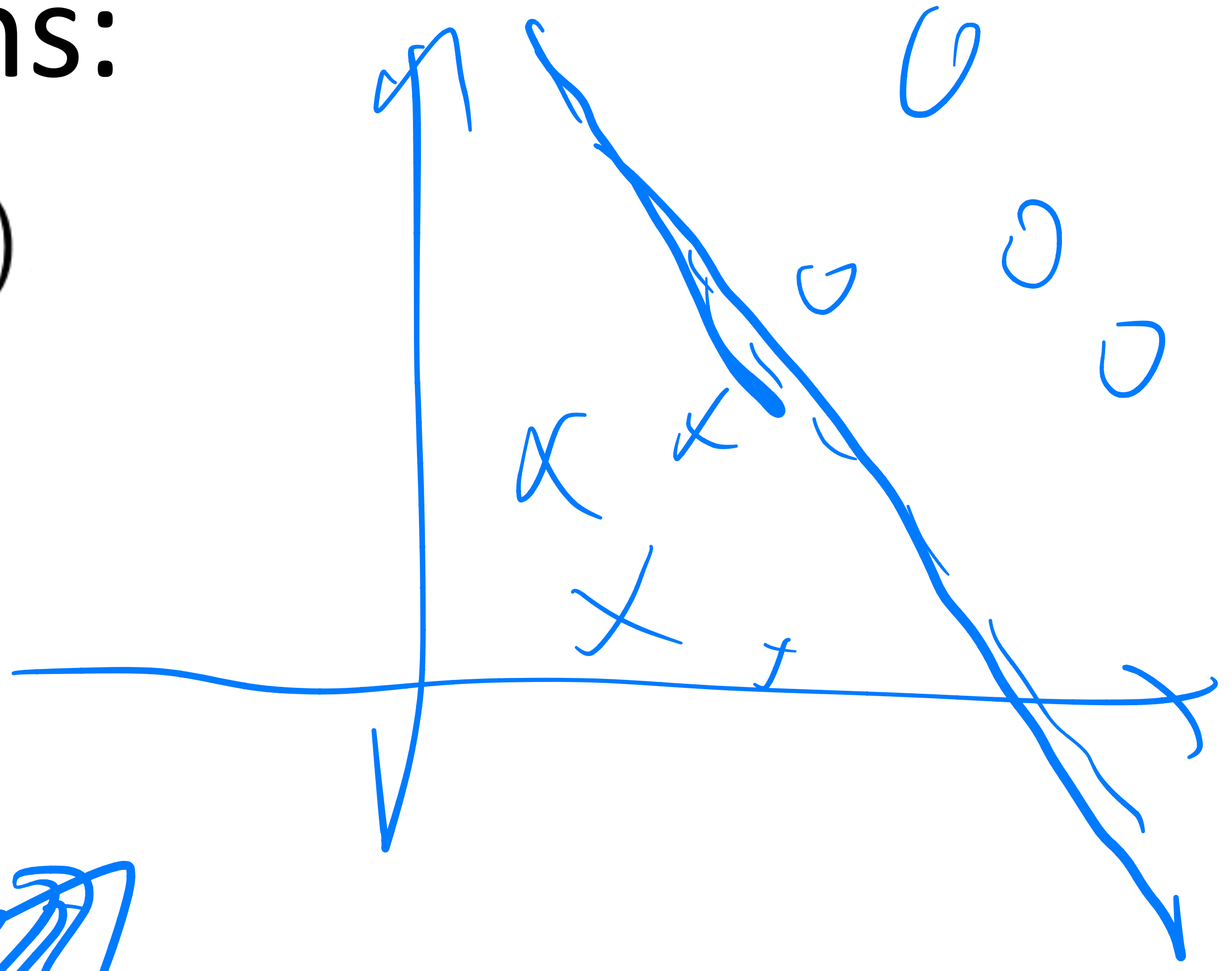
The original constraints

$g_i(w) \leq 0$

$\alpha \geq 0$

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:



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

$w = [w, b]$

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

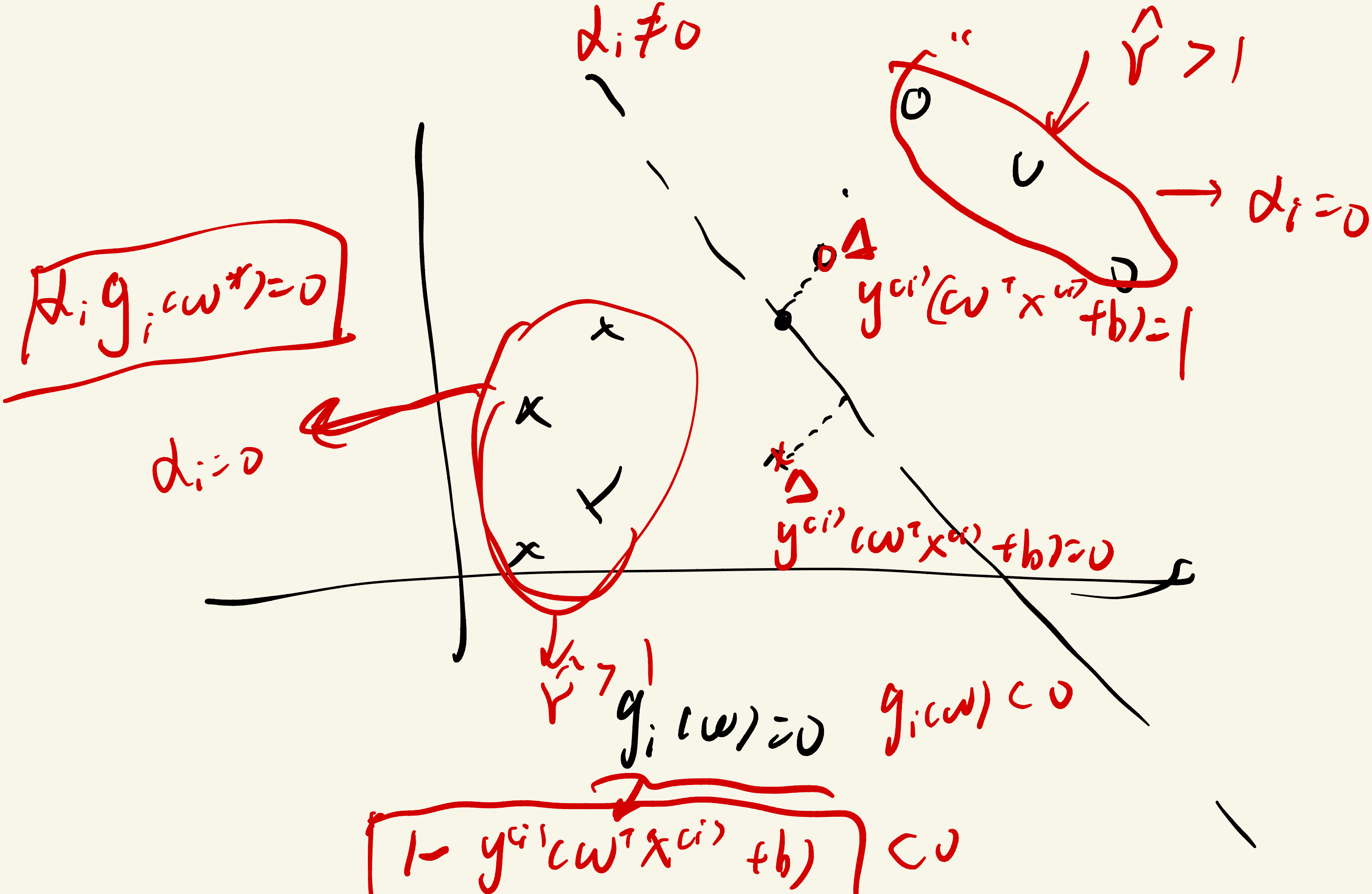
inequality

$g_i(w) = 0$

$= 1 - y^{(i)}(w^T x + b) \leq 0$

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

$g_i(w^*) < 0, \; \alpha_i^* = 0$



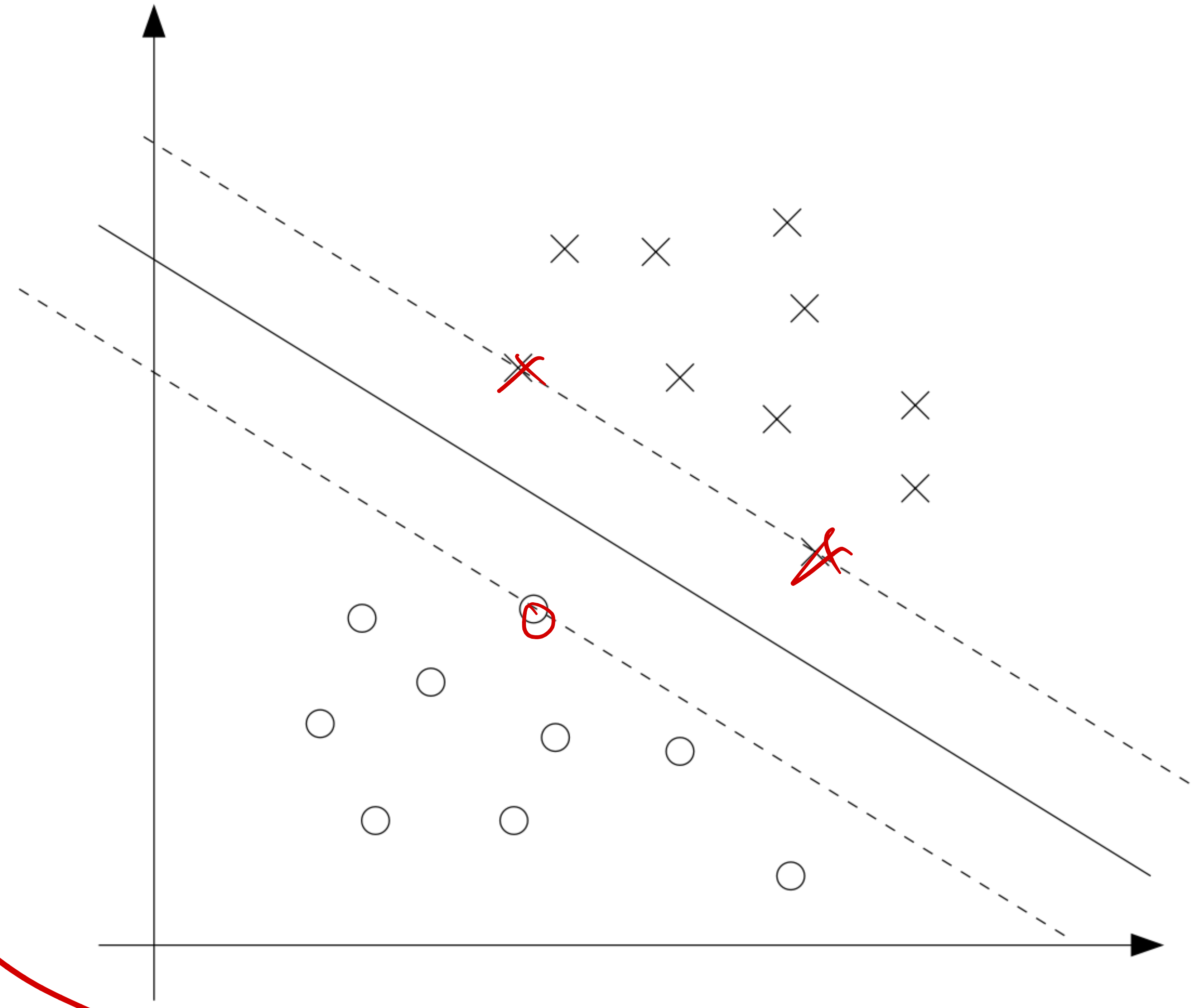
$\alpha_i \neq 0$
$\hat{\gamma} > 1$
$\alpha_i = 0$
$\alpha_i g_i(w^*) = 0$
$y^{(i)}(w^T x^{(i)} + b) = 1$
$\alpha_i = 0$
$y^{(i)}(w^T x^{(i)} + b) = 0$
$\hat{\gamma} > 1$
$g_i(w) = 0$
$g_i(w) < 0$
$1 - y^{(i)}(w^T x^{(i)} + b) < 0$

Supporting Vectors

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

$\alpha_i = 0$ $\alpha_i \neq 0$

imply ?



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

Lagrangian for SVM

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

The dual optimization problem

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

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

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

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

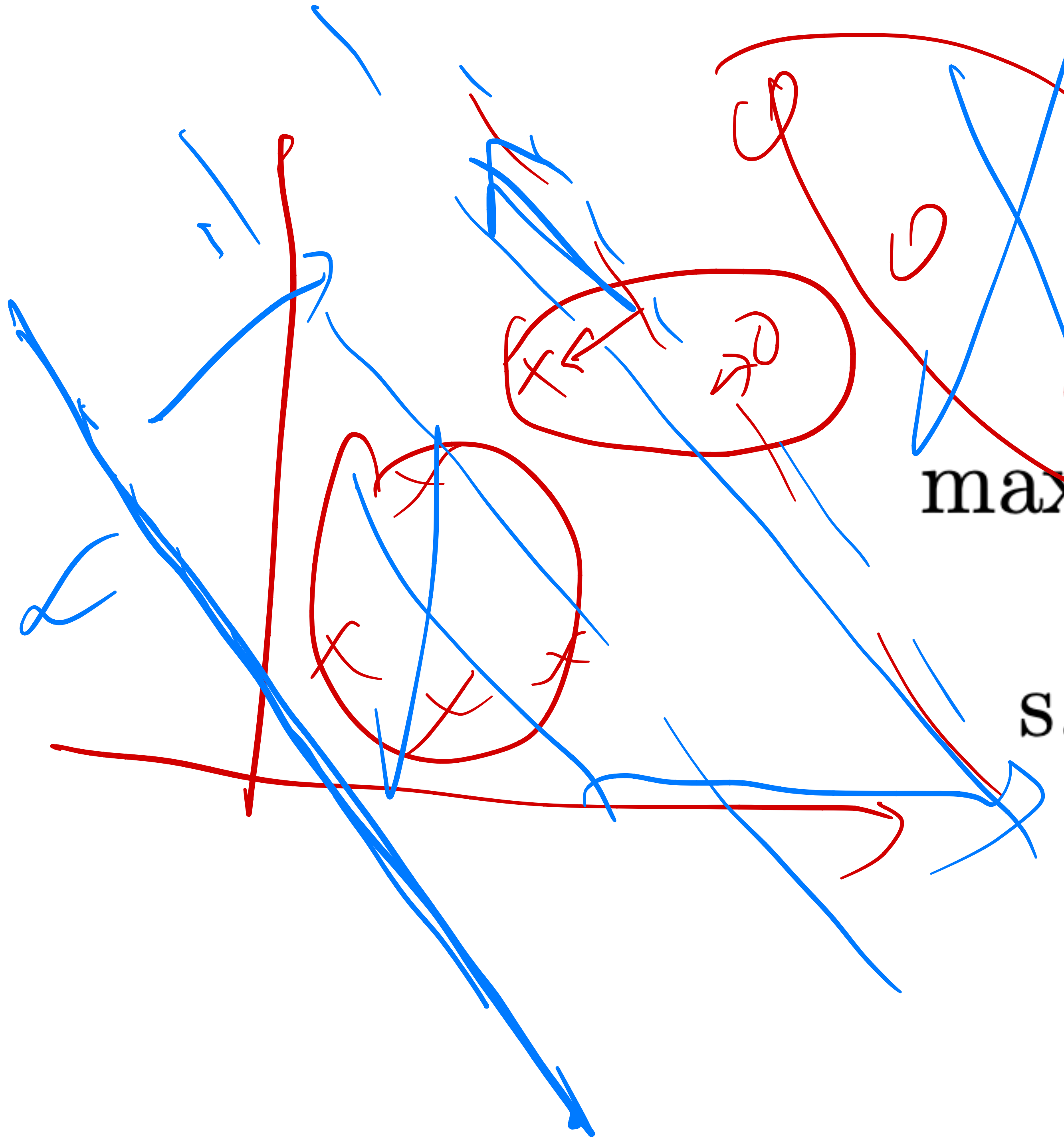
$$\frac{\partial}{\partial b} \mathcal{L}(w, b, \alpha) = \sum_{i=1}^{n} \alpha_i y^{(i)} = 0$$

$\alpha_i = 0$

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

$x^{(i)}$ is irrelevant

The Dual Problem of SVM



$K(x^{(i)}, x^{(j)})$

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

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

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

Kernel is all we need!

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

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

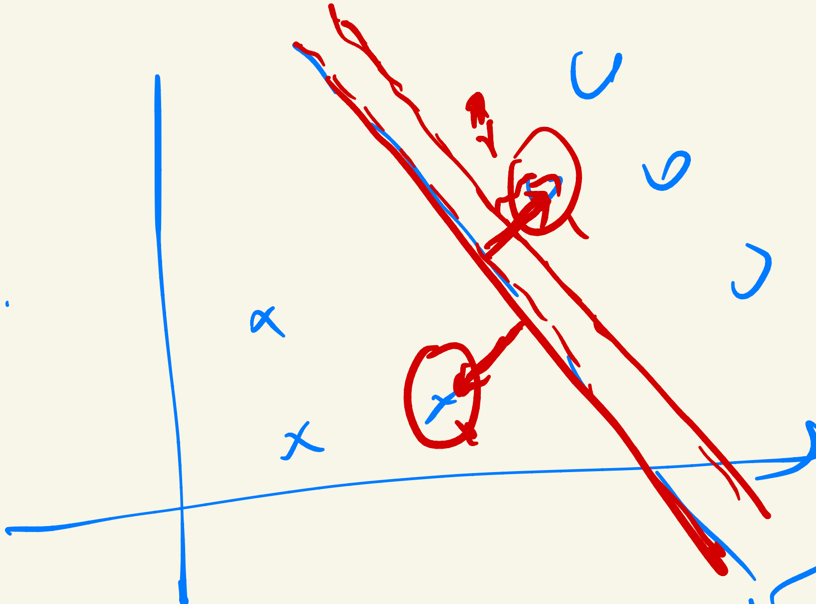
From KKT Conditions

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

From the original constraints

$y^{(i)}(w^T x^{(i)} + b) = 1$

$\max\min \hat{\gamma}$



$\boxed{w^*}^T + b^* = 0$

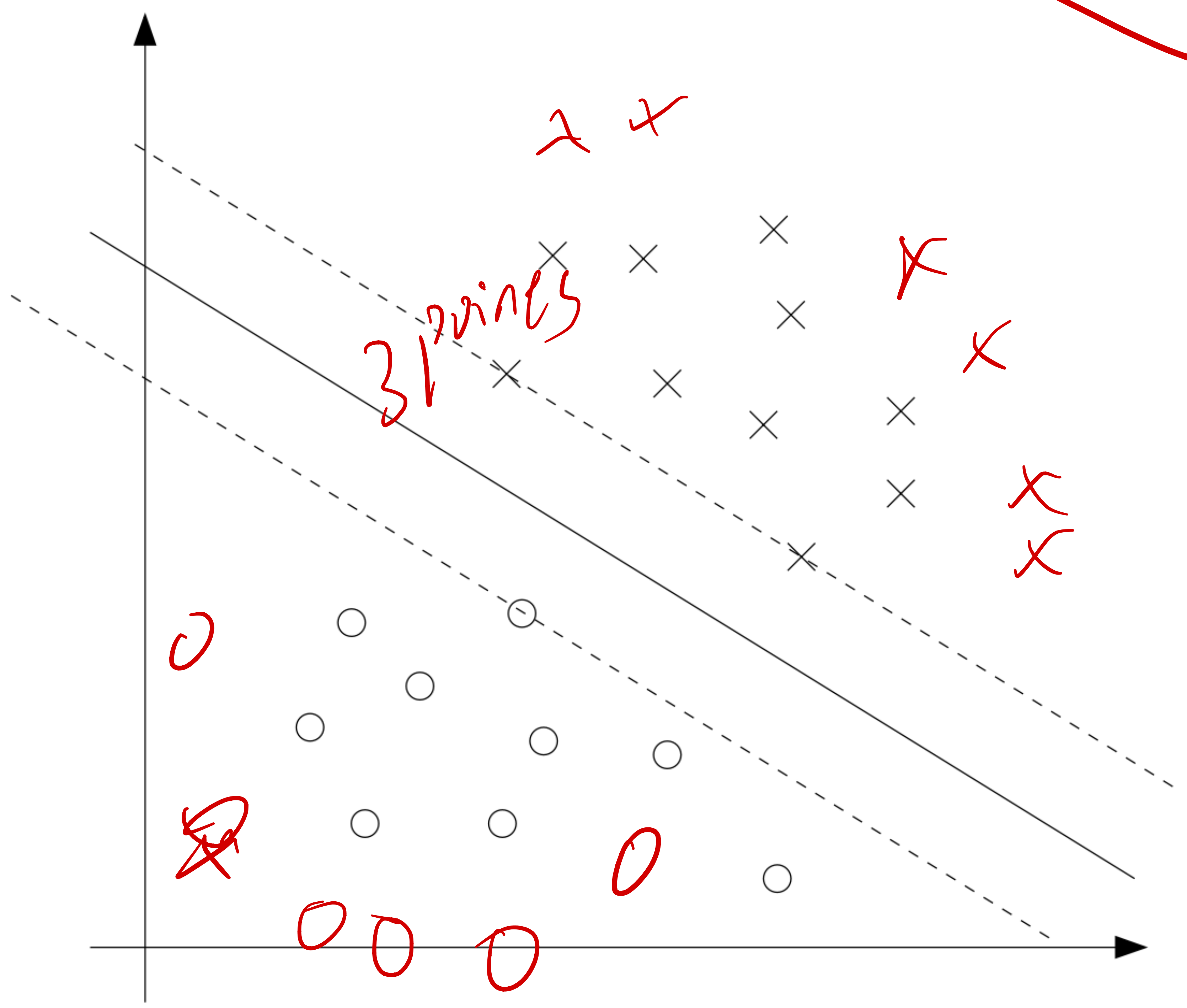
Inference

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

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

support

3 terms



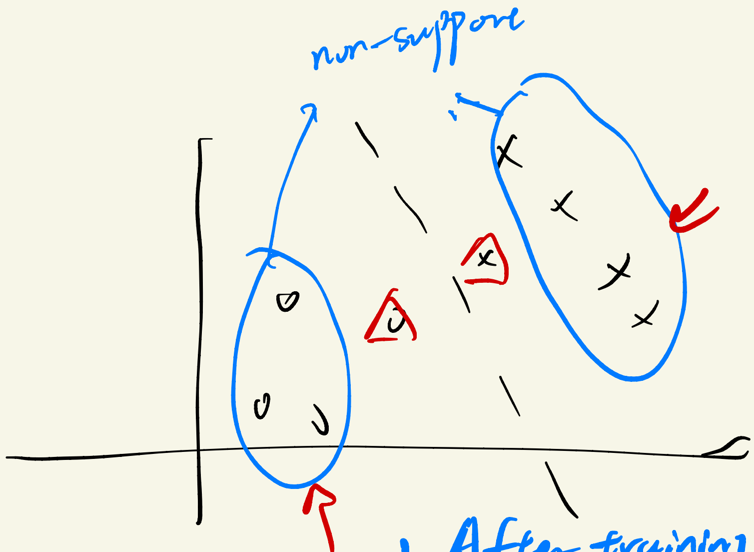
We never need to really compute w

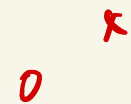
$$\alpha_i^* g_i(w^*) = 0, \quad i = 1, \ldots, k$$

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

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

$$\boxed{\hat{\gamma} = 1} \qquad \boxed{\gamma = \frac{1}{\|w\|}}$$

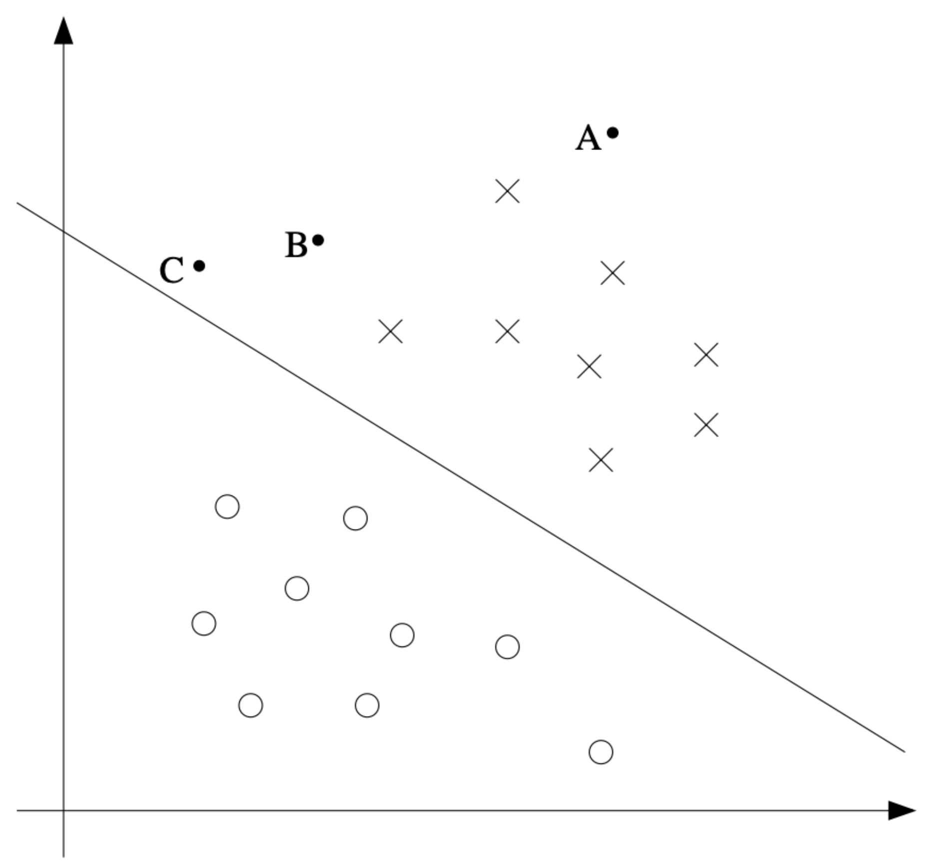




1. After training, if we delete all the non-support data samples, prediction change? No

2. If we know the non-support samples in advance, and we delete in the beginning, decision boundary change? No

3. Same to 2, whether training process change? dynamics? Yes



Review of the High-Level Logic

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

Maximize geometric margin → Quadratic Optimization Problem → Dual optimization problem

Problem rewriting

Finding a related optimization problem that is easier

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

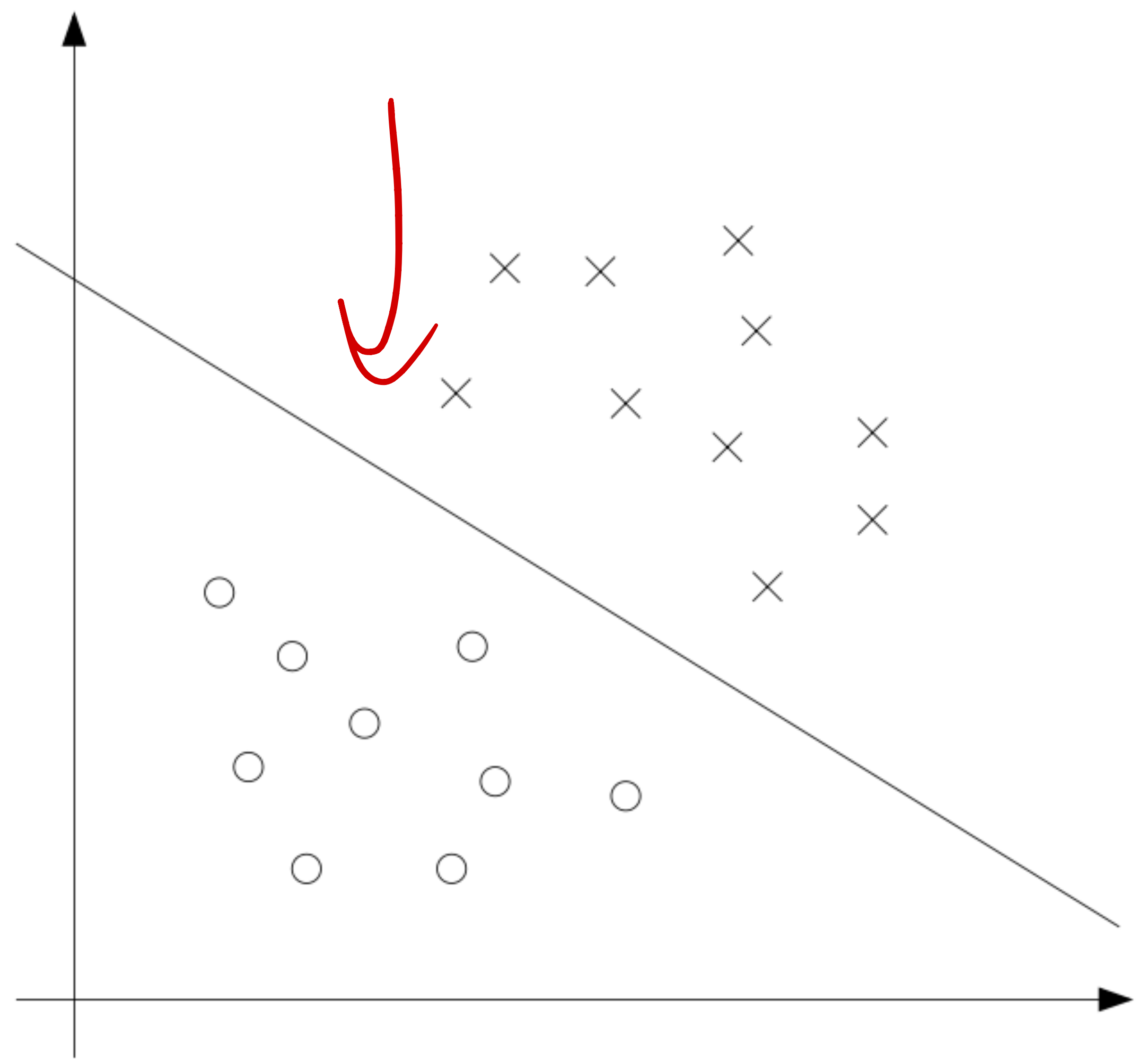
$$\begin{aligned} \min_{w,b} \quad & \frac{1}{2}||w||^2 \\ \text{s.t.} \quad & y^{(i)}(w^T x^{(i)} + b) \geq 1, \quad i = 1, \ldots, n \end{aligned}$$

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

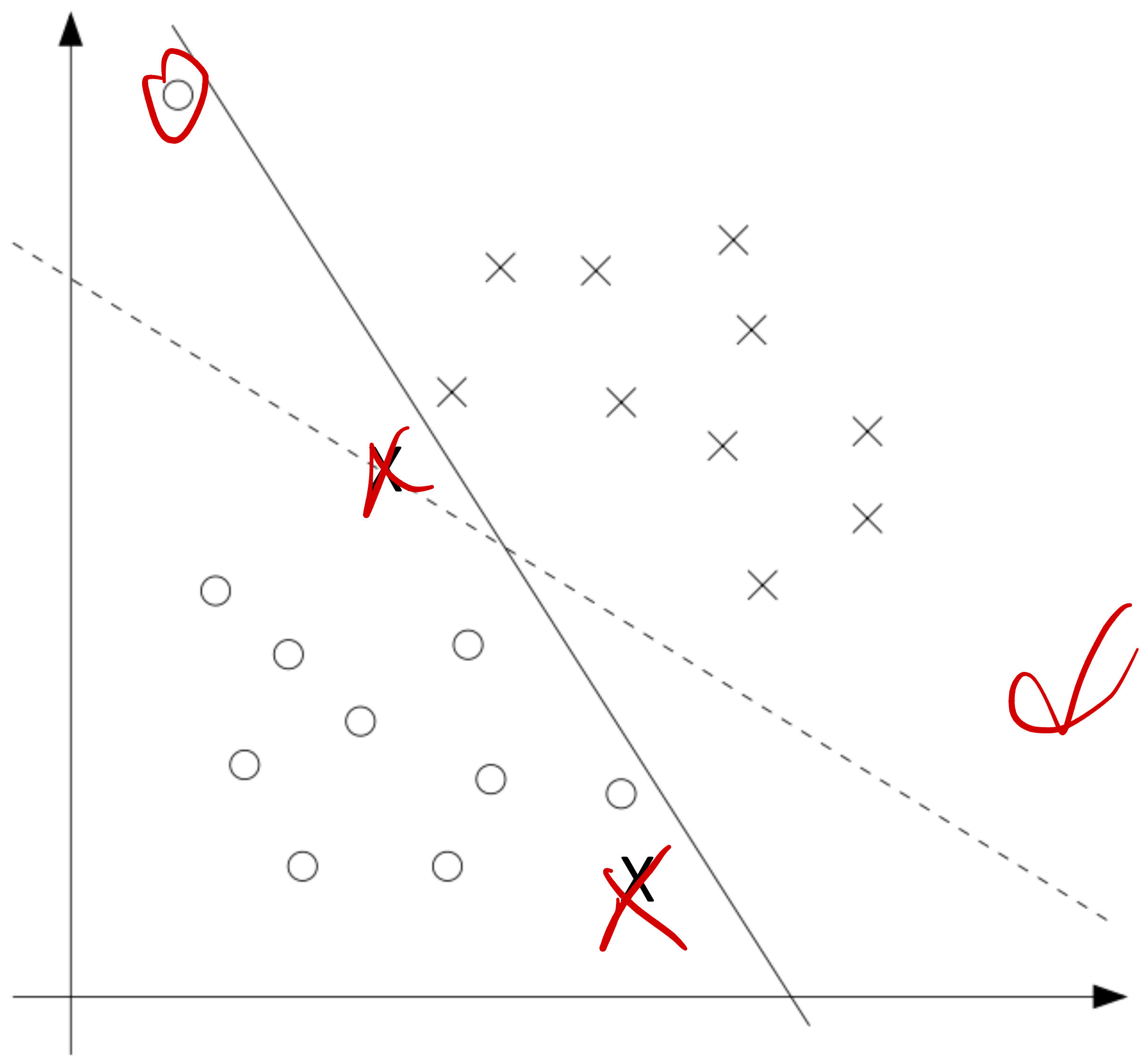
$$\begin{aligned} \max_\alpha \quad & W(\alpha) = \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j=1}^n y^{(i)} y^{(j)} \alpha_i \alpha_j \langle x^{(i)}, x^{(j)} \rangle \\ \text{s.t.} \quad & \alpha_i \geq 0, \quad i = 1, \ldots, n \\ & \sum_{i=1}^n \alpha_i y^{(i)} = 0, \end{aligned}$$

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:

$$y^{(i)}(w^T x^{(i)}+b)<0$$

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

$$1-\xi_i$$

Dual opt problem

You will prove this in your hw

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

$$0 \leq \alpha_i \leq C$$

Thank You!

Q & A