Last time

Perceptron
Support Vector Machines
  Primal Problem – Margin Maximization
  Dual Problem

Today – part 1

- - Kernel Version
-- What if there are outliers in the problem?

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

$$\text{s.t } \forall i \quad (w^T x_i)\, y_i \geq 1$$

DUAL PROBLEM

$$\max_{\boxed{\alpha \geq 0}} \quad \alpha^T 1 - \frac{1}{2}\alpha^T y^T X^T X y \alpha$$

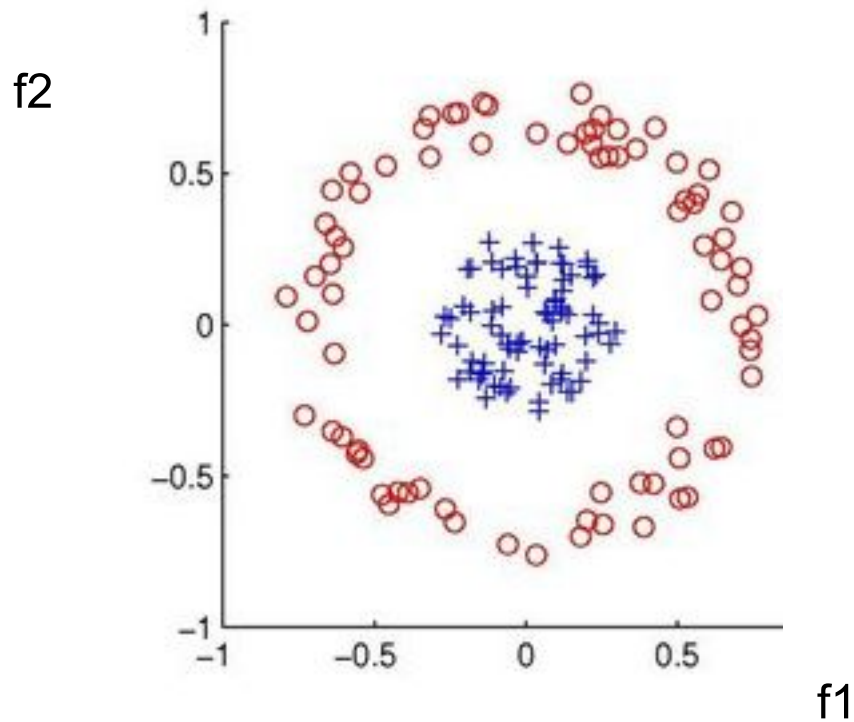
Solving in n instead of d.

easy constraints.

can be KERNELIZED!

nxn.

ISSUE → Features could be non-linearly related



In general, $(f_1 - a)^2 + (f_2 - b)^2 = r^2$

Every datapoint x roughly satisfies

$$f_1^2 + a^2 - 2f_1 a + f_2^2 + b^2 - 2f_2 \cdot b - r^2 = 0$$

$$\begin{bmatrix} f_1 & f_2 \end{bmatrix} \longrightarrow \begin{bmatrix} 1 & f_1 & f_2 & f_1 f_2 & f_1^2 & f_2^2 \end{bmatrix}$$

$$x \in \mathbb{R}^2 \longrightarrow \phi(x) \in \mathbb{R}^6$$

$u \in \mathbb{R}^6$

Datapoints $\phi(x)$ satisfies

$$\underbrace{\begin{bmatrix} 1 & f_1 & f_2 & f_1 f_2 & f_1^2 & f_2^2 \end{bmatrix}}_{\phi(x)} \begin{bmatrix} a^2+b^2-r^2 \\ -2a \\ -2b \\ 0 \\ 1 \\ 1 \end{bmatrix}$$

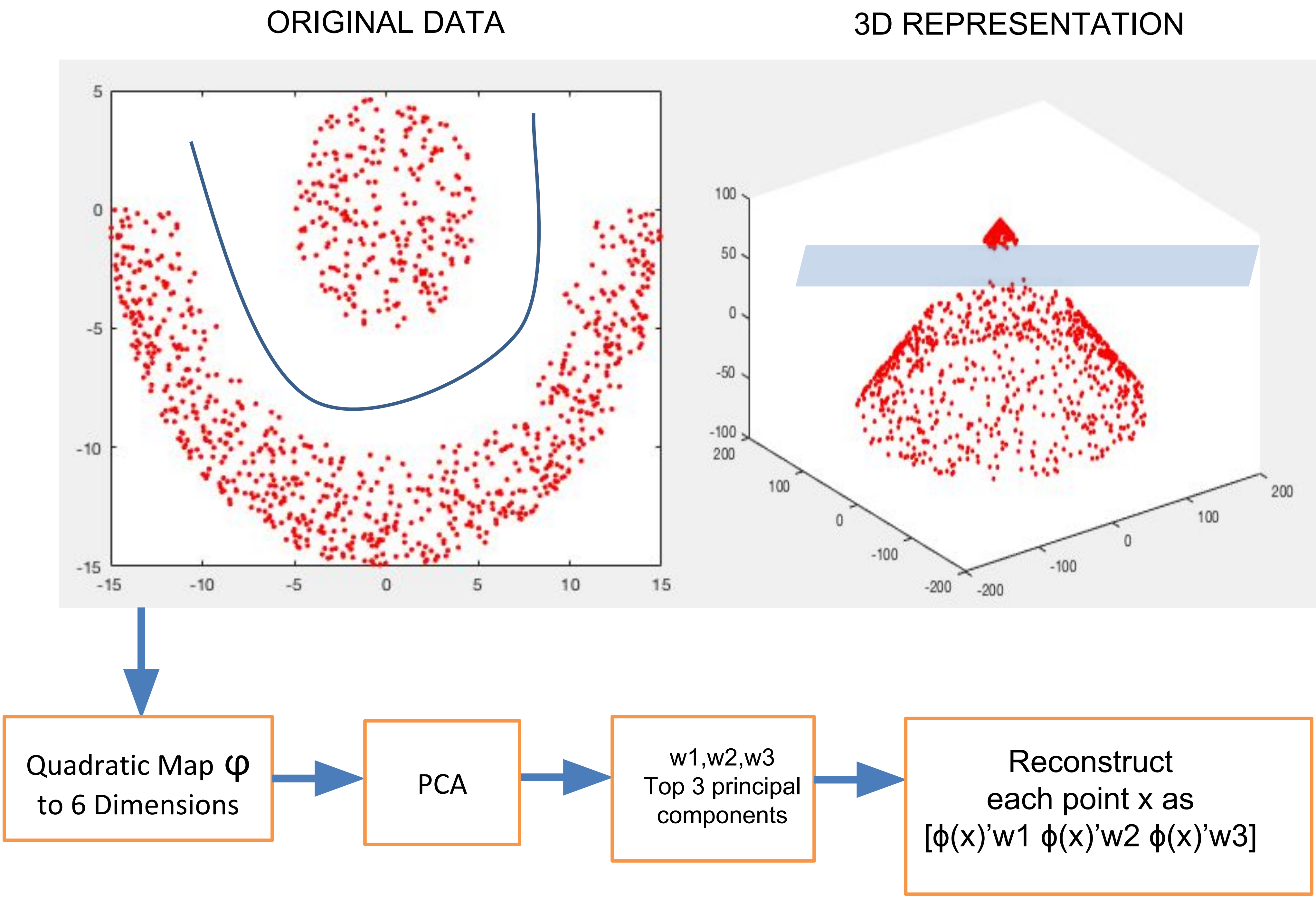
$$\boxed{\phi(x)^T u = 0}$$

Idea: Transform features from

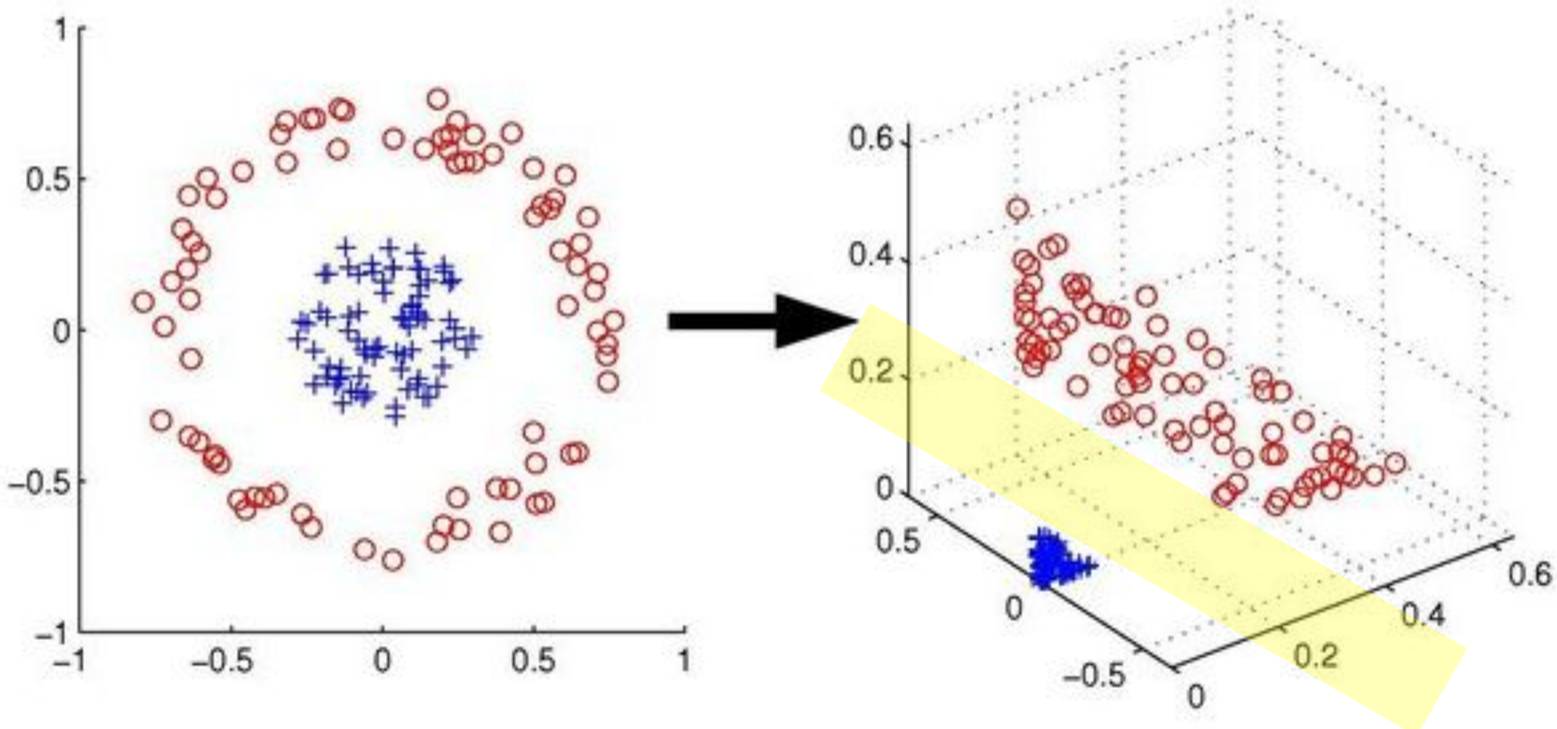
low dimension $\mathbb{R}^d$ to

high dimension $\mathbb{R}^D$

$$\underset{\mathbb{R}^d}{x} \longrightarrow \underset{\mathbb{R}^D}{\phi(x)}$$


ORIGINAL DATA
3D REPRESENTATION
5
0
-5
-10
-15
-15
-10
-5
0
5
10
15
100
50
0
-50
-100
200
100
0
-100
-200
-200
-100
0
100
200
Quadratic Map φ to 6 Dimensions
PCA
w1,w2,w3
Top 3 principal components
Reconstruct each point x as
[ϕ(x)'w1 ϕ(x)'w2 ϕ(x)'w3]

$\{x_1, x_2, \ldots, x_n\}$ $\{\Phi(x_1), \Phi(x_2), \ldots, \Phi(x_n)\}$



$\forall i,\ x_i \in R^2$ $\forall i,\ \phi(x_i) \in R^3$

$$x = [f_1 \quad f_2 \quad f_3 \quad f_4]$$

$\Downarrow$ Cubic relation

$$\phi(x) = [1 \quad \underbrace{f_1 \quad f_2 \quad f_3 \quad f_4}$$

$$1 + 4 \quad + \quad 4c^2 + 4c^3$$

$$x \in d$$

$$\phi(x) \in O(d^p)$$

ISSUE .

$\phi(x) \in \mathbb{R}^D \rightarrow$ might be too large

EXAMPLE

$$x = \begin{bmatrix} f_1 & f_2 \end{bmatrix} \qquad x' = \begin{bmatrix} g_1 & g_2 \end{bmatrix}$$

$$\boxed{\left(x^T x' + 1\right)^2} = \left(\begin{bmatrix} f_1 & f_2 \end{bmatrix} \begin{bmatrix} g_1 \\ g_2 \end{bmatrix} + 1 \right)^2$$

$$= \left(f_1 g_1 + f_2 g_2 + 1 \right)^2$$

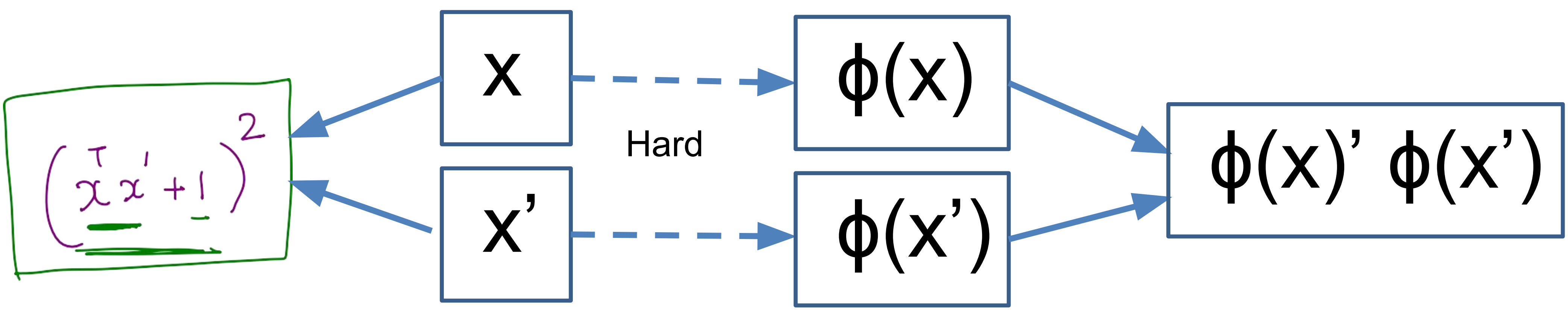
$$= \underline{f_1^2 g_1^2} + \underline{f_2^2 g_2^2} + 1 + 2 f_1 g_1 f_2 g_2$$

$$+ \underline{2 f_1 g_1} + 2 f_2 g_2$$

$$\boxed{\left(x^T x' + 1\right)^2}$$

$$= \begin{bmatrix} f_1^2 & f_2^2 & 1 & \sqrt{2} f_1 f_2 & \sqrt{2} f_1 & \sqrt{2} f_2 \end{bmatrix} \begin{bmatrix} g_1^2 \\ g_2^2 \\ 1 \\ \sqrt{2} g_1 g_2 \\ \sqrt{2} g_1 \\ \sqrt{2} g_2 \end{bmatrix}$$

$$\phi(x)^T \phi(x')$$



So far

-> To capture non-linear relationships, one can “create” non-linear functions of features.

-> But the number of features to create grows exponentially with the degree of non-linearity p that we wish to capture (d^p)

-> For d = 2 and p = 2, it appears there is a trick to get around this.

-> Is this trick general enough to be useful the general case as well? (i.e., for any d and any p?)

MORE EXAMPLES

Polynomial map

$$k(x, x') = (x^T x' + 1)^p \quad \text{for some } p \geq 1$$

→ Can be shown to be a "valid" function.

i.e., $\exists\ \phi \quad \mathbb{R}^d \rightarrow \mathbb{R}^D$ such that

$$K(x, x') = \phi(x)^T \phi(x')$$

EXERCISE

Compute ϕ for $p = 3$, $p = 4$.

MORE EXAMPLES

$$K(\underset{\in \mathbb{R}^d}{x}, \underset{\in \mathbb{R}^d}{x'}) = \exp\left(- \frac{\| x - x' \|^2}{2\sigma^2} \right)$$

for some $\sigma > 0$

→ RADIAL BASIS FUNCTION

→ Can also shown to be a "valid" map

→ Interestingly, ϕ in this case maps x to an "infinite" dimensional space

[Technicalities aside think of $\phi(x)$ as a function and dot product as integrals]

KERNEL FUNCTION

Any function $K: \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ which is a "valid" map is a kernel function

$$K(x, x') = (x^T x' + 1)^p \rightarrow \text{Polynomial kernel}$$

$$K(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right) \rightarrow \text{Gaussian kernel / Radial basis kernel}$$

Question: Given a function $k: \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, how can we say it's a valid kernel?

METHOD 1: Explicitly construct a ϕ map

[might be hard sometimes]

METHOD 2 : MERCER's THEOREM

A function $k : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ is a valid kernel

if and only if

a) k is symmetric.

b) For any dataset $\{x_1, \ldots, x_n\}$, the

matrix $K \in \mathbb{R}^{n \times n}$

$K_{ij} = k(x_i, x_j)$ is POSITIVE SEMI DEFINITE

eigenvalues of K are all non-negative.

DUAL PROBLEM

$$\max_{\boxed{\alpha \geq 0}} \quad \alpha^T 1 - \frac{1}{2} \alpha^T y^T X^T X y \alpha$$

Solving in n instead of d.

easy constraints

can be KERNELIZED!

nxn.

Instead of $X^T X$, use K for the chosen Kernel.

- Cannot recover $w = \sum \alpha_i \phi(x_i)$) – *but that's okay. Why?*

- Solve the dual problem. Obtain alphas.
- For a test point, to predict, use the following:

$$w^T \Phi(x\_test) = (\sum \alpha_i \phi(x_i))^T \Phi(x_{\_test}) = \sum \alpha_i \; \mathbf{K(x_i, x_{\_test})}$$

Idea (to deal with outliers):

Fix any w. w classifies some points correct and some incorrectly. Let the incorrect points pay "bribe" to get to the correct side.

Modified formulation

$C > 0$ [hyper parameter]

$$\min_{\boxed{w, \xi}} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i$$

$\rightarrow (w^T x_i) y_i + \underline{\xi_i} \geq 1 \leftarrow \forall i$

$\rightarrow \underline{\xi_i \geq 0} \leftarrow \forall i$

If $C = 0 \Rightarrow$ Bribes don't cost $\Rightarrow$ $w = 0$ is solution

$C \rightarrow \infty \Rightarrow$ Bribes are too costly $\Rightarrow$ Linear separable case.

$$L(w, \xi, \alpha, \beta) = \frac{1}{2}\|w\|^2 + \underbrace{C\left(\sum_{i=1}^{n} \xi_i\right)}_{\uparrow} + \sum_{i=1}^{n} \alpha_i \left(1 - (w^T x_i) y_i - \xi_i\right)$$

$$+ \sum_{i=1}^{n} \beta_i (-\xi_i)$$

Dual:

$$\max_{\substack{\alpha \geq 0 \\ \beta \geq 0}} \min_{w} L(w, \xi, \alpha, \beta)$$

$$\frac{\partial L}{\partial w} = 0 \implies w^* = \sum_{i=1}^{n} \alpha_i x_i y_i$$

$$\boxed{w^* = XY\alpha}$$

$$\frac{\partial L}{\partial \xi_i} = 0 \quad \Rightarrow \quad \boxed{C - \alpha_i - \beta_i = 0}$$

$$\boxed{\alpha_i + \beta_i = C} \quad \forall i$$

Substitute $w = XY\alpha$ in the original objective

$$\frac{1}{2}(XY\alpha)^T(XY\alpha) + \sum_{i=1}^{n}\underline{(C-\alpha_i-\beta_i)}\xi_i + \alpha^T 1 - (XY\alpha)^T(XY\alpha)$$

SOFT-MARGIN

SUPPORT

VECTOR

MACHINE

$$\max_{\substack{\alpha \geq 0 \\ \beta \geq 0 \\ \alpha_i + \beta_i = C}} \quad \alpha^T 1 - \frac{1}{2} (X y \alpha)^T (X y \alpha)$$

$$\equiv \boxed{\max_{0 \leq \alpha \leq C} \quad \alpha^T 1 - \frac{1}{2} \alpha^T y^T (X^T X) y \alpha}$$

BOX CONSTRAINT.

Summary

HARD-MARGIN SVM

PRIMAL

$$\min_{w} \frac{1}{2}\|w\|^2$$

$$\text{s.t. } \underbrace{(w^T x_i) y_i \geq 1}_{1 - w^T x_i y_i \leq 0} \quad \forall i$$

DUAL

$$\max_{\alpha \geq 0} \alpha^T 1 - \alpha^T y^T x^T x y \alpha$$

$$\alpha_i^*(1 - w^{*T} x_i y_i) = 0 \quad \forall i$$

$$\boxed{w^* = \sum_{i=1}^{n} \alpha_i^* x_i y_i}$$

SOFT-MARGIN SVM

PRIMAL ✓

$$\min_{w,\xi} \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{n} \xi_i$$

$$\text{s.t. } (w^T x_i) y_i + \xi_i \geq 1 \quad \forall i = 1, \ldots, n \quad (\alpha)$$

$$\xi_i \geq 0 \quad \forall i = 1, \ldots, n \quad (\beta)$$

DUAL ✓

$$\max_{\substack{\boxed{\alpha + \beta = C} \\ \alpha \geq 0 \\ \beta \geq 0}} \alpha^T 1 - \alpha^T y^T x^T x y \alpha$$

$$\underbrace{\alpha \geq 0, \ \beta \geq 0}_{0 \leq \alpha \leq C}$$

- Let $(\underline{w}^*, \underline{\xi}^*)$ be the primal optimal solution
- Let $(\underline{\alpha}^*, \underline{\beta}^*)$ be the dual optimal solution

COMPLEMENTARY SLACKNESS

$$\alpha_i^* \left(1 - ({w^*}^T x_i) y_i - \xi_i^* \right) = 0 \quad \forall i$$

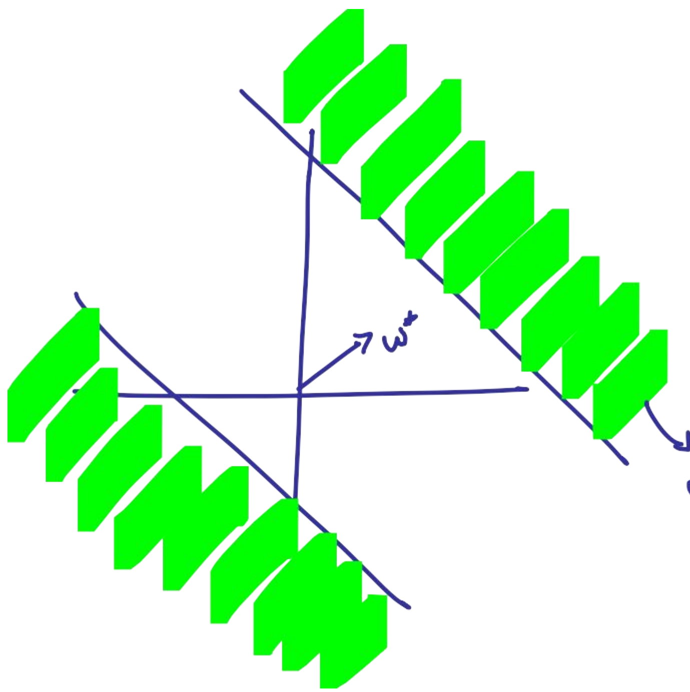
$$\beta_i^* \xi_i^* = 0 \quad \forall i$$

$$\alpha_i^* + \beta_i^* = C \quad \forall i$$

$\hookrightarrow$ (A)

Various cases possible

Case 1: $\alpha_i^* = 0$ $\overset{4}{\Rightarrow}$ $\beta_i^* = C$ $\overset{\boxed{CS}}{\Rightarrow}$ $\xi_i^* = 0$



$\{x : w^{*T} x_i y_i \geq 1\}$

$1 - (w^{*T} x_i) y_i - \underbrace{\xi_i^*}_{=0} \leq 0$ [Primal feasibility]

$\Rightarrow \quad 1 - (w^{*T} x_i) y_i \leq 0$

$\Rightarrow \quad w^{*T} x_i y_i \geq 1$

$\Rightarrow$ w^* classifies (x_i, y_i) correctly.

Case 2: $0 < \alpha_i^* < C \overset{A}{\Longrightarrow} 0 < \beta_i^* < C \overset{CS}{\Longrightarrow} \underline{\xi_i^*} = 0$

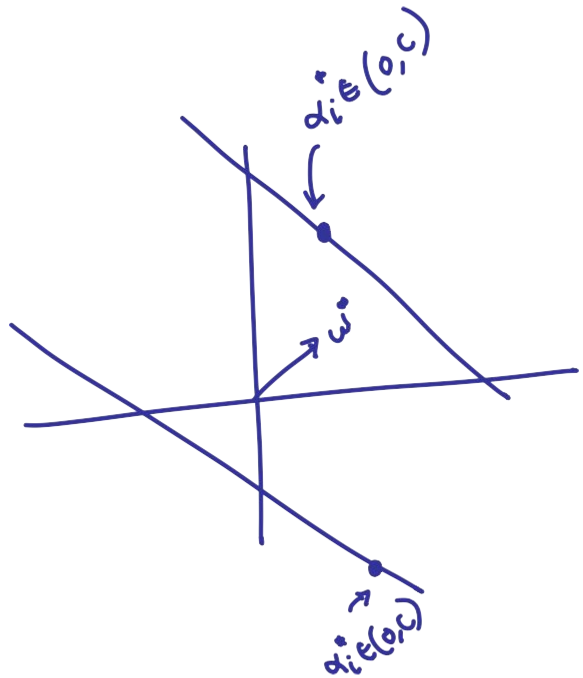
$$\Big\Downarrow \boxed{CS}$$

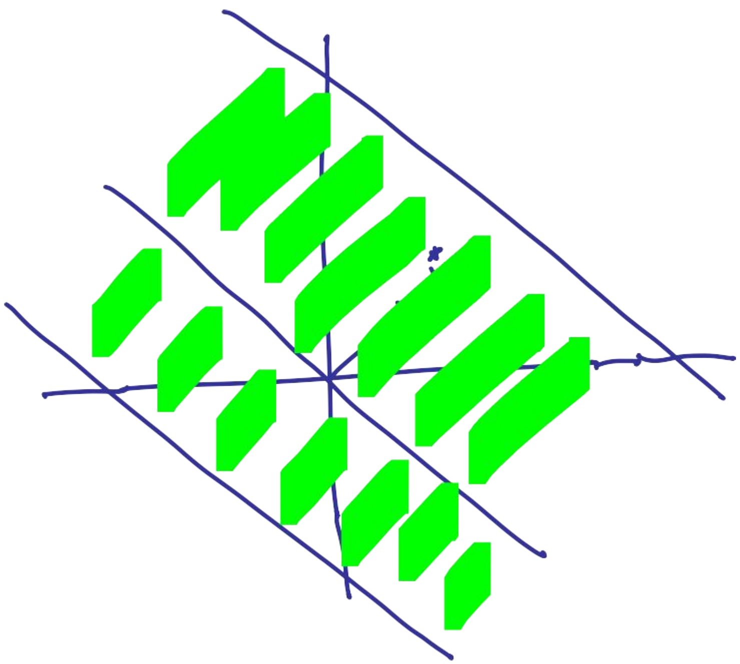
$$1 - (w^{*T} x_i) y_i - \xi_i^* = 0$$

$$\Big\Downarrow$$

$$(w^{*T} x_i) y_i = 1$$

$\Rightarrow$ (x_i, y_i) lies on the supporting hyperplane.





Case 3: $\quad \underline{\alpha_i^* = C} \Rightarrow \beta_i^* = 0 \Rightarrow \xi_i^* \geq 0$

$\Downarrow \boxed{CS}$

$$1 - w^{*T} x_i y_i - \xi_i^* = 0$$

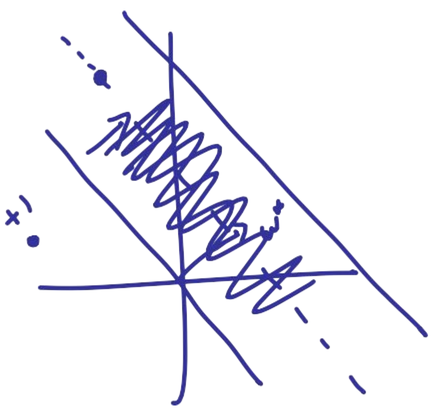
$$\xi_i^* = 1 - w^{*T} x_i y_i \geq 0$$

$$\Rightarrow \boxed{w^{*T} x_i y_i \leq 1}$$

Let's see this from P.o.v of data

CASE 1

$$\boxed{w^{*T} x_i y_i < 1}$$



$$1 - w^{*T} x_i y_i - \xi_i^* \leq 0$$

$$w^{*T} x_i y_i \geq 1 - \xi_i^*$$

$$\boxed{\xi_i^* \geq 1 - w^{*T} x_i y_i}$$

$$\Rightarrow \xi_i^* > 0 \Rightarrow \beta_i^* = 0 \Rightarrow \alpha_i^* = C$$

$$\alpha_i^* \left(1 - w^{*T} x_i y_i - \xi_i^*\right) = 0$$

$$\beta_i^* \xi_i^* = 0.$$

CASE 2: $w^{*T} x_i y_i = 1$

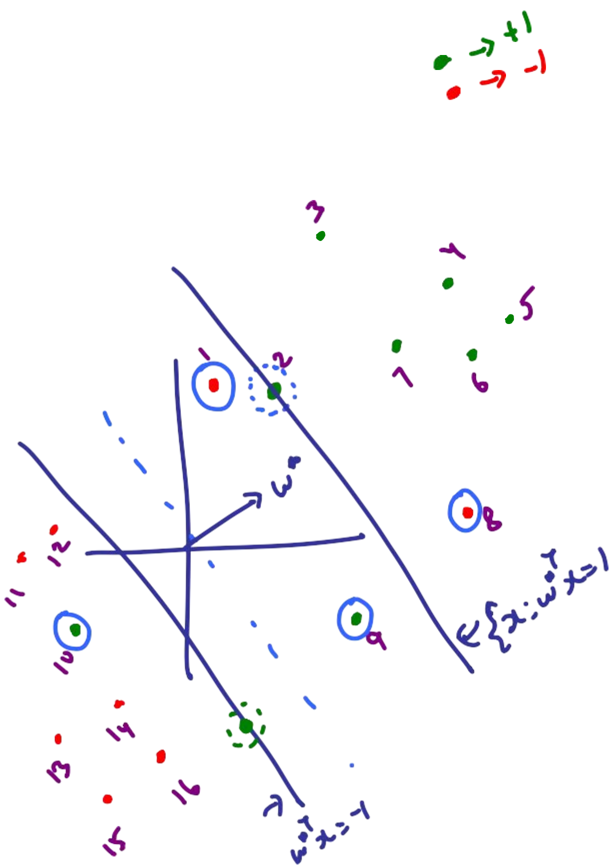
$$\xi_i^* \geq 1 - \underline{w^{*T} x_i y_i}$$

$$\Rightarrow \quad \xi_i^* \geq 0 \quad \Rightarrow \quad \alpha_i^* \in [0, C]$$

CASE 3 $\quad w^{*T} x_i y_i > \underline{1}$

$$1 - \underbrace{\underbrace{w^{*T} x_i y_i}} - \xi_i^* \leq 0 \quad \text{[Primal feasibility]}$$

$$\Rightarrow \quad 1 - w^{*T} x_i y_i - \xi_i^* < 0 \quad \xRightarrow{\boxed{C.S}} \quad \alpha_i^* = 0$$



SUMMARY

$$\alpha_i^* = 0 \quad \Rightarrow \quad w^{*T} x_i y_i \geq 1$$

$$0 < \alpha_i^* < C \quad \Rightarrow \quad w^{*T} x_i y_i = 1$$

$$\alpha_i^* = C \quad \Rightarrow \quad w^{*T} x_i y_i \leq 1$$

✓ $w^{*T} x_i y_i < 1 \quad \Rightarrow \quad \alpha_i^* = C$

→ $w^{*T} x_i y_i = 1 \quad \Rightarrow \quad \alpha_i^* \in [0, C]$

→ $w^{*T} x_i y_i > 1 \quad \Rightarrow \quad \alpha_i^* = 0.$