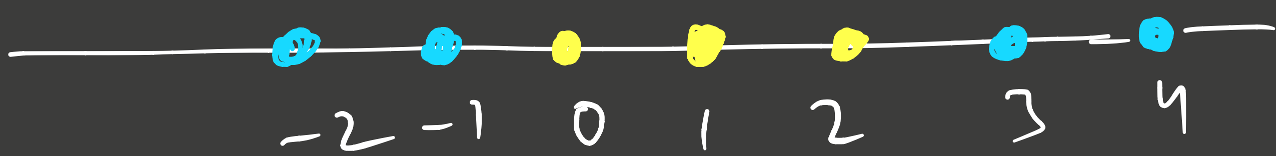
PRIMAL WAS QP, DUAL IS QP,

WHY BOTHER CONVERTING TO DUAL?

JUST WAIT FOR FEW MORE MINS

ANSWER: "KERNEL TRICK")

NON-LINEARLY SEPARABLE DATA

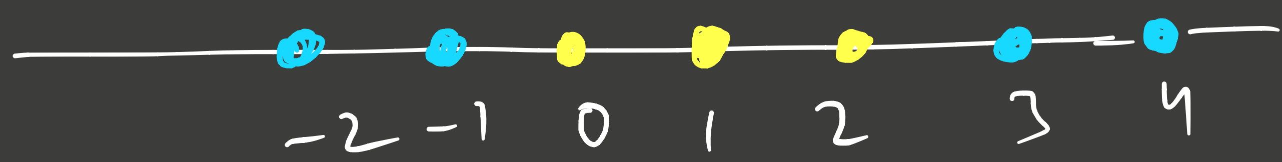


Data not separate in $\mathbb{R}$

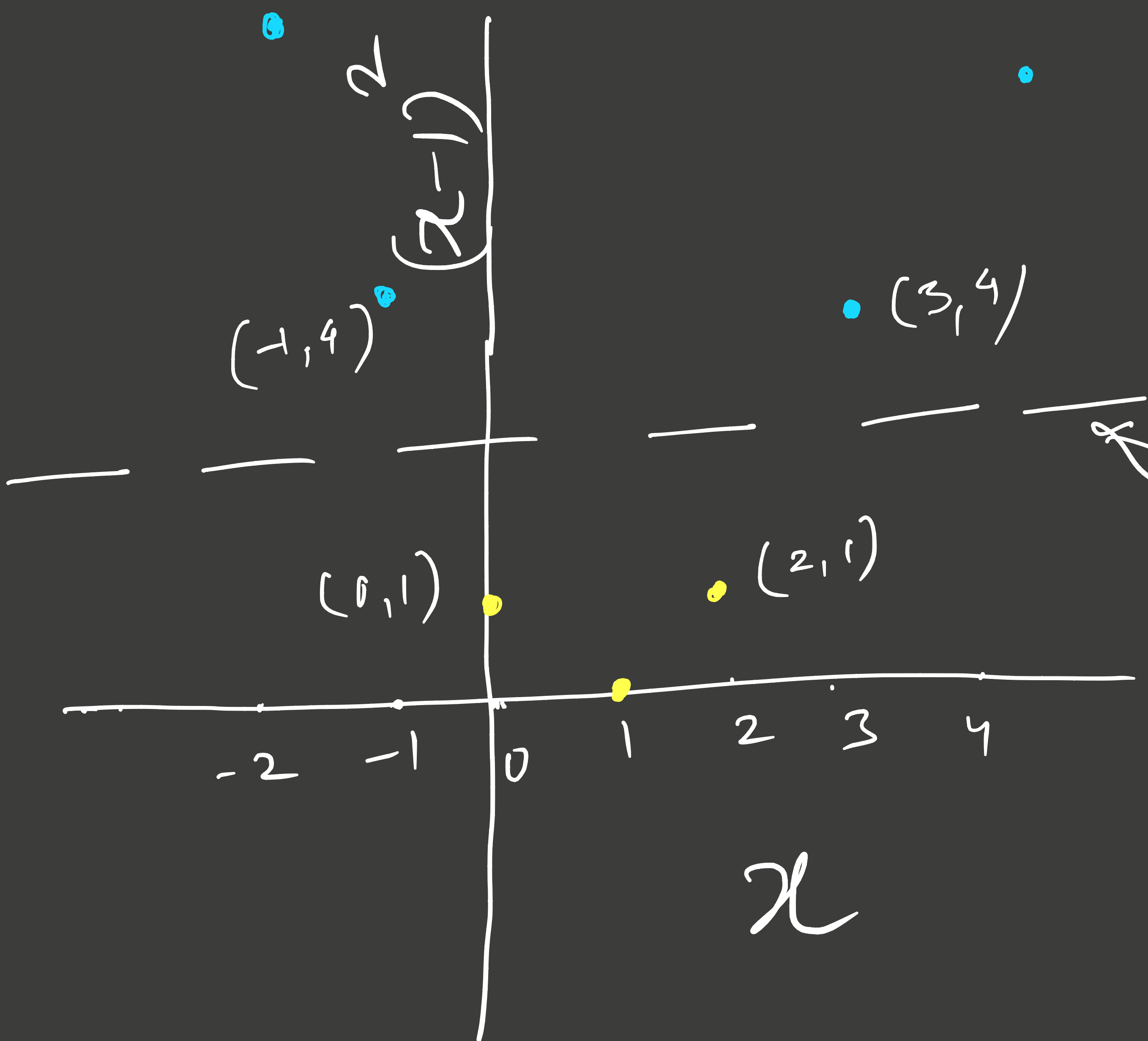
Can we still use SVM?

Yes!

How: Project data to a higher dimensional space.

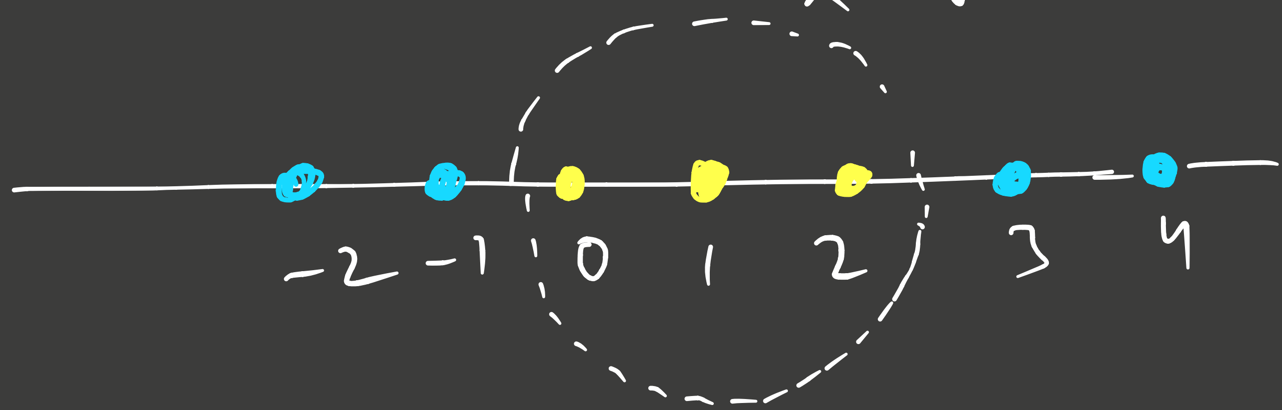


original data in R



Transformed data in R^2

Max margin separating hyperplane



← separating plane in original dimension

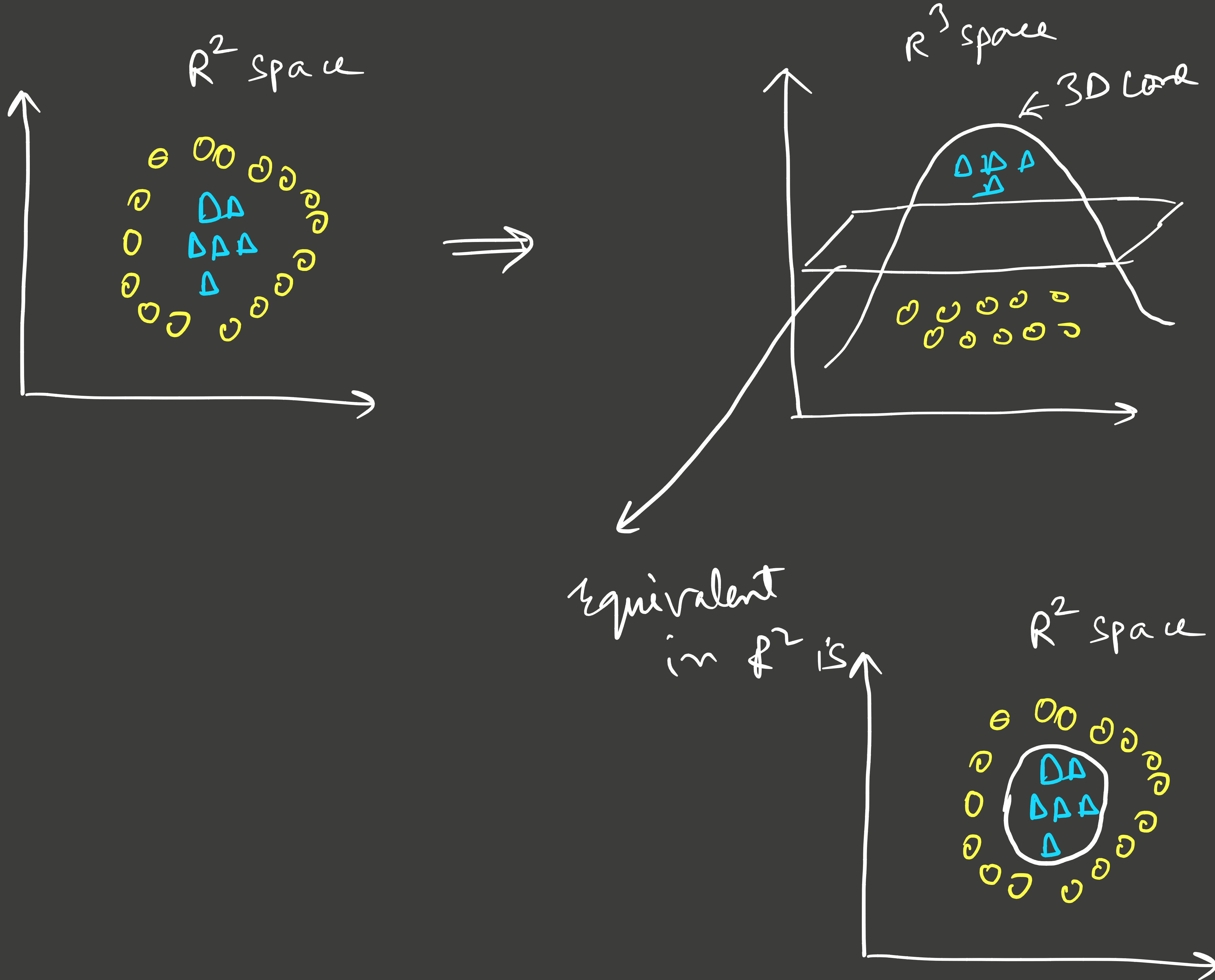
$$(x-1)^2 = \left(\sqrt{\frac{5}{2}}\right)^2$$

Circle :

Center ($x = 1$)

Radius $\sqrt{5/2}$

ANOTHER EXAMPLE TRANSFORMATION



Projection | Transformation Function

$$\phi : R^d \longrightarrow R^D$$

where $d =$ original dimension

$D =$ New dimension

In our example;

$d = 1$; $D = 2$

Linear SVM

MAXIMIZE

$$L(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \vec{x_i} \cdot \vec{x_j}$$

s.t.

CONSTRAINTS

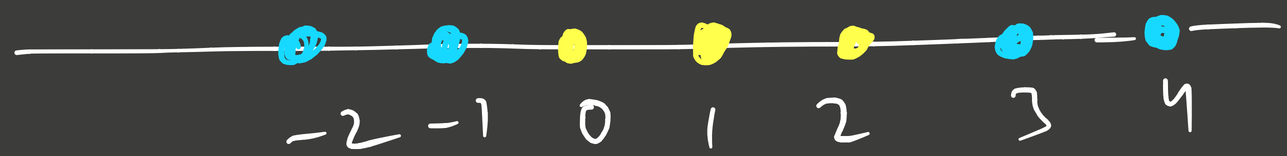
$\Downarrow$

TRANSFORMATION (ϕ)

$\Downarrow$

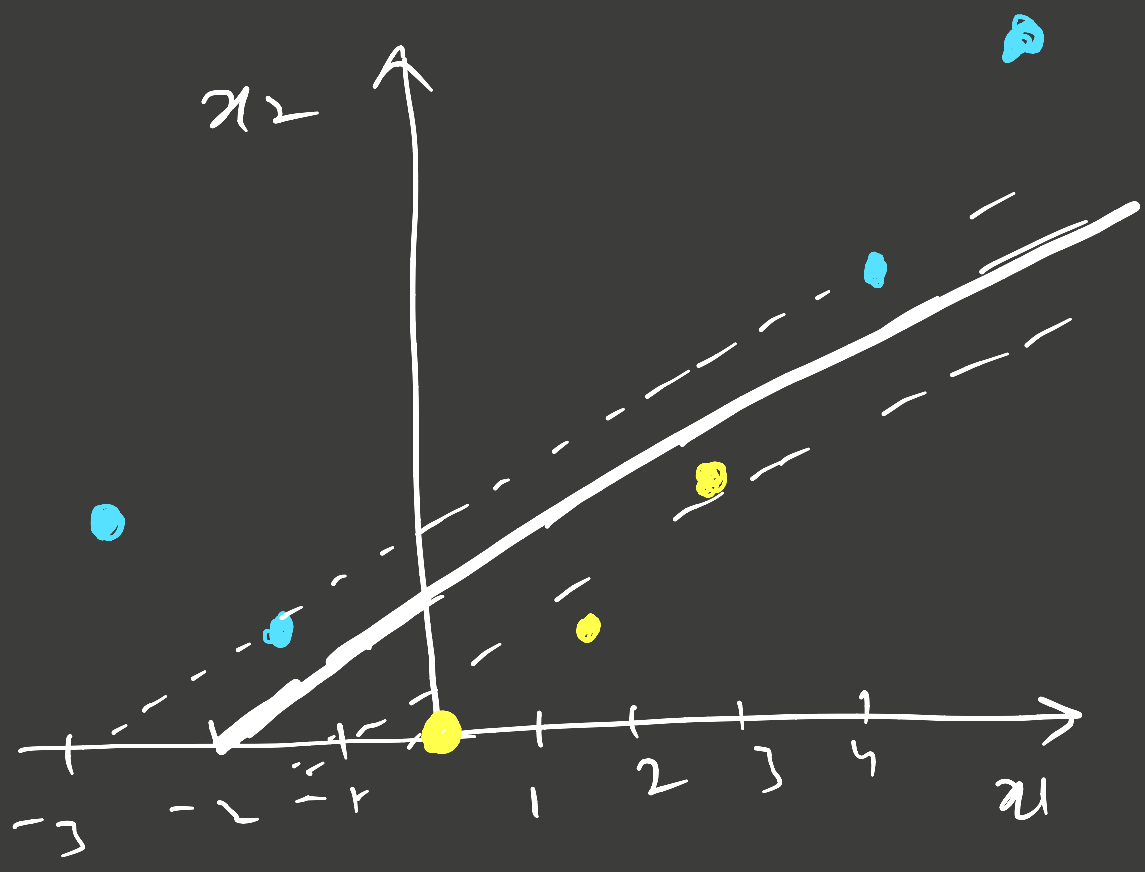
$$L(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \phi(\vec{x_i}) \cdot \phi(\vec{x_j})$$

TRIVIAL EXAMPLE (again)



Original Data $(x) \in \mathbb{R}$

Transformed Data $\left(\phi(x) = \langle \sqrt{2}x, x^2 \rangle \right)$



Steps

① Compute $\phi(x)$ for each point

$$\phi : R^d \to R^D$$

② Compute dot products over R^d space

0) If $D >> d$

Both steps are expensive!

KERNEL TRICK

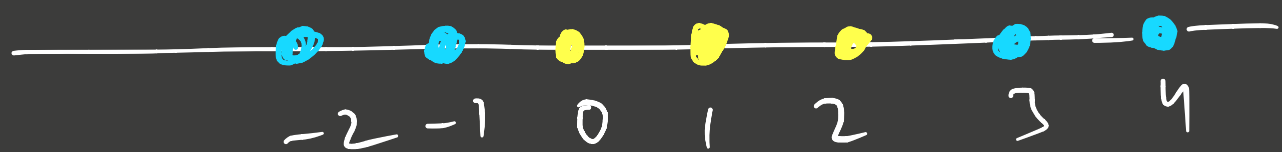
Can we compute $K(\bar{x}_i, \bar{x}_j)$

s.t.

$$K(\bar{x}_i, \bar{x}_j) = \phi(\bar{x}_i) \cdot \phi(\bar{x}_j)$$

some func^n of dot product in original dimension

Dot product in high dimensions (after transformation)



$$\phi(x) = \langle \sqrt{2}\,x, x^2 \rangle$$

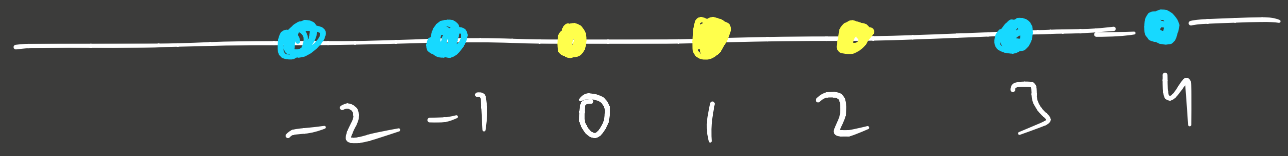
$$K(x_i, x_j) = (1 + x_i x_j)^2 - 1$$

Dot product in lower dimension

$$= \cancel{1} + 2x_i x_j + x_i^2 x_j^2 - \cancel{1}$$

$$= \langle \sqrt{2}\,x_i, x_i^2 \rangle \cdot \langle \sqrt{2}\,x_j, x_j^2 \rangle$$

$$= \phi(x_i) \cdot \phi(x_j)$$



Original Dataset

#	x	y
1	-2	-1
2	-1	-1
3	0	1
⋮	⋮	⋮

Transformed dataset

#	$\sqrt{2}x$	x^2	y
1	$-2\sqrt{2}$	4	-1
2	$-\sqrt{2}$	1	-1
3	0	0	1
⋮	⋮	⋮	⋮

$\phi(x_1) = \langle -2\sqrt{2}, 4 \rangle \,;\; \phi(x_2) = \langle -\sqrt{2}, 1 \rangle$ TRANSFORMATION

$\phi(x_1) \cdot \phi(x_2) = -2\sqrt{2} * -\sqrt{2} + 4 \times 1 = 8$ DOT PRODUCT IN 2D

$K(x_1, x_2) = \{1 + (-2) * (-1)\}^2 - 1$ DOT PRODUCT IN 1D

WHY DID WE USE DUAL FORM?

KERNELS AGAIN!!

PRIMAL FORM DOESN'T ALLOW FOR "KERNEL TRICK"

$K(\vec{x}_1, \vec{x}_2)$ in DUAL

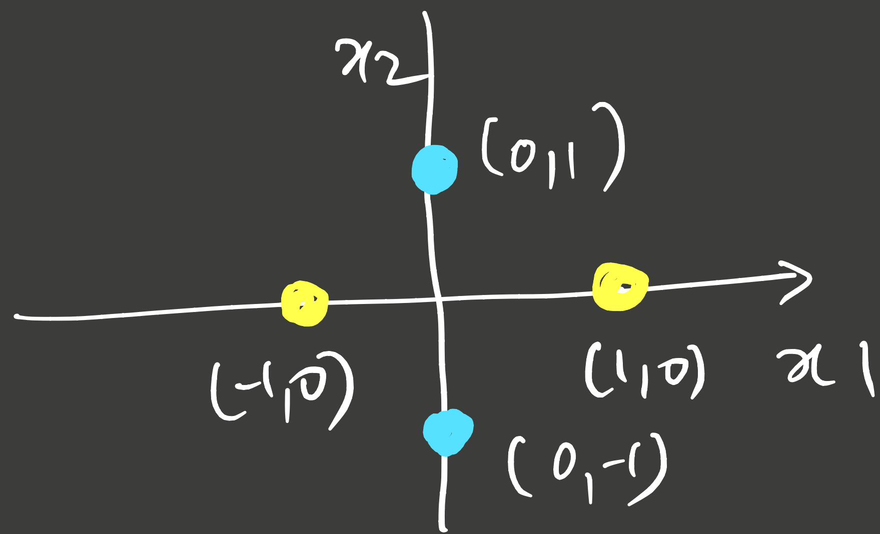
& COMPUTE $\phi(x)$ and then dot product in 'D' dimensions?

GRAM MATRIX (Positive Semi-Definite)

$$K(x_i, x_j) = (1 + x_i \cdot x_j)^2 - 1$$

	x_1	x_2	x_3	x_4	x_5	x_6	x_7
x_1	24	8	0	0	8	24	48
x_2	8	1	0	-1	0		
x_3	0	..	..	..	..		
x_4	0						
x_5	8						
x_6	24						
x_7	48						

ANOTHER EXAMPLE



$$K(\bar{x}, \bar{x}') = (\vec{x}^T \cdot \vec{x}')^2$$

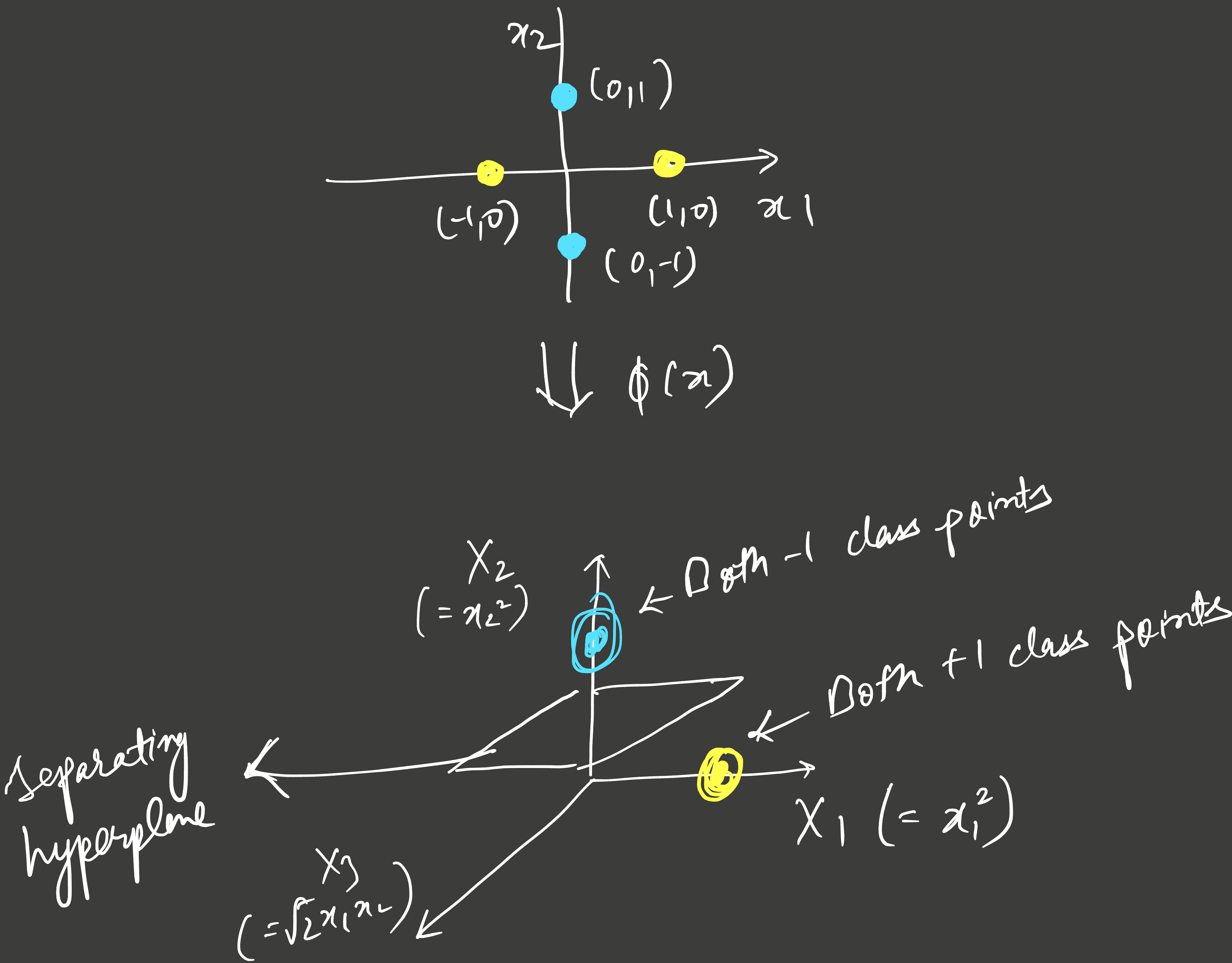
$$\bar{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} ; \quad x' = \begin{bmatrix} x_1' \\ x_2' \end{bmatrix}$$

Q: What is $\phi(x)$?

$$K(\bar{x}, \vec{x}') = \phi(\vec{x}) \cdot \phi(\vec{x}')$$

$$K(\vec{x}, \vec{x}') = \left\{ \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} x_1' \\ x_2' \end{bmatrix} \right\}^2 = (x_1 x_1' + x_2 x_2')^2$$

$$\Rightarrow \boxed{\phi(x) = \langle x_1^2, \sqrt{2} x_1 x_2, x_2^2 \rangle} = x_1^2 x_1'^2 + x_2^2 x_2'^2 + 2 x_1 x_1' x_2 x_2'$$



x_2
$(0,1)$
$(-1,0)$
$(1,0)$
x_1
$(0,-1)$
$\Downarrow \phi(x)$
X_2
$(= x_2^2)$
Both -1 class points
Both +1 class points
Separating hyperplane
X_1 $(= x_1^2)$
X_3
$(= \sqrt{2}x_1x_2)$

SOME KERNELS

① Linear : $K(\bar{x}_1, \bar{x}_2) = \bar{x}_1 \cdot \bar{x}_2$

② Polynomial: $K(\bar{x}_1, \bar{x}_2) = (p + \vec{x}_1 \cdot \vec{x}_2)^q$

③ Gaussian : $K(\bar{x}_1, \bar{x}_2) = e^{-\gamma \|\bar{x}_1 - \bar{x}_2\|^2}$

ALSO CALLED RADIAL BASIS FUNCTION (RBF)

$$\gamma = \frac{1}{2\sigma^2}$$

Q) For $\bar{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ what space does

Kernel $K(\bar{x}, \vec{x}') = (1 + \bar{x} \cdot \vec{x}')^3$ belong to?

or

$\vec{x} \in R^2$

$\phi(\vec{x}) \in R^?$

$$K(x, z) = (1 + x_1 z_1 + x_2 z_2)^3$$

$$= \ldots$$

$$= \langle 1, x_1, x_2, x_1^2, x_2^2, x_1^2 x_2, x_1 x_2^2, x_1^3, x_2^3 \rangle$$

10 dimensional?

Q) For $\bar{x} = x$; what space does RBF kernel lie in?

$$K(x,z) = e^{-\gamma \|x-z\|^2}$$

$$= e^{-\gamma (x-z)^2}$$

Now; $e^{\alpha} = \sum_{n=0}^{\infty} \frac{\alpha^n}{n!}$

$\therefore e^{-\gamma (x-z)^2}$ is ∞ dimensional!!

Q) Is SVM parametric or non-parametric?

Q) Is SVM parametric or non-parametric?

Yes and No

↓ Linear kernel / Polynomial kernel (form fixed)

↓ RBF (form changes with data)

RBF is Non Parametric

$$\hat{y}(x_{TEST}) = \text{SIGN}(\vec{w} \cdot \vec{x}_{TEST} + b)$$

$$= \text{SIGN}\left(\sum_{j=1}^{N_{SV}} \alpha_j y_j \vec{x}_j \cdot \vec{x}_{Test} + b\right)$$

$\Downarrow$ Kernelized.

$$\hat{y}(x_{Test}) = \text{SIGN}\left(\sum_{j=1}^{N} \alpha_j y_j K(\vec{x}_j, \vec{x}_{Test}) + b\right)$$

$\alpha_j = 0$ where $j \neq$ S.V.

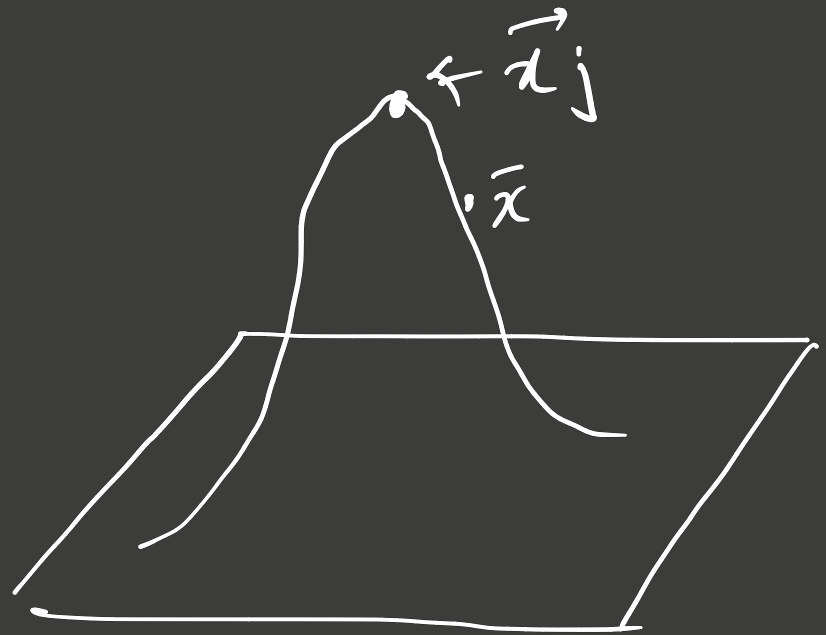
Now $K(\vec{x}_j, \vec{x}_{Test})$ for RBF is:

$$e^{-\gamma \|\underline{\underline{\vec{x}_j}} - \vec{x}_{Test}\|^2}$$

$\therefore$ Hypothesis is a $\underline{\underline{\text{function}}}$ of "All" train points.

$\downarrow$

What kind of?



Closer $\vec{x}$ is to $\vec{x}_N$, more is it influencing $\underline{\underline{\hat{y}}}(\vec{x})$

$\uparrow$ Hypothesis function

Now if we add a point to dataset

$\Downarrow$

Functional form can adapt (similar to KNN)

$\therefore$ SVM with RBF Kernel is Non-Parametric

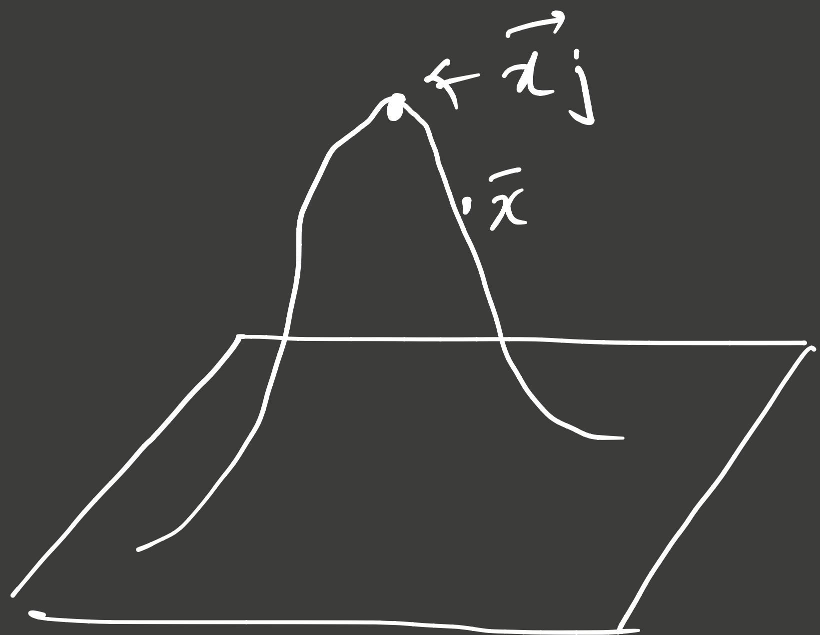
Interpretation of RBF

$$\hat{y}(x) = \text{SIGN}\left(\sum \underbrace{\alpha_i y_i}_{\text{Activation}} \underbrace{e^{-\|x - x_i\|^2}}_{\text{Basis}} + b \right)$$

Radial

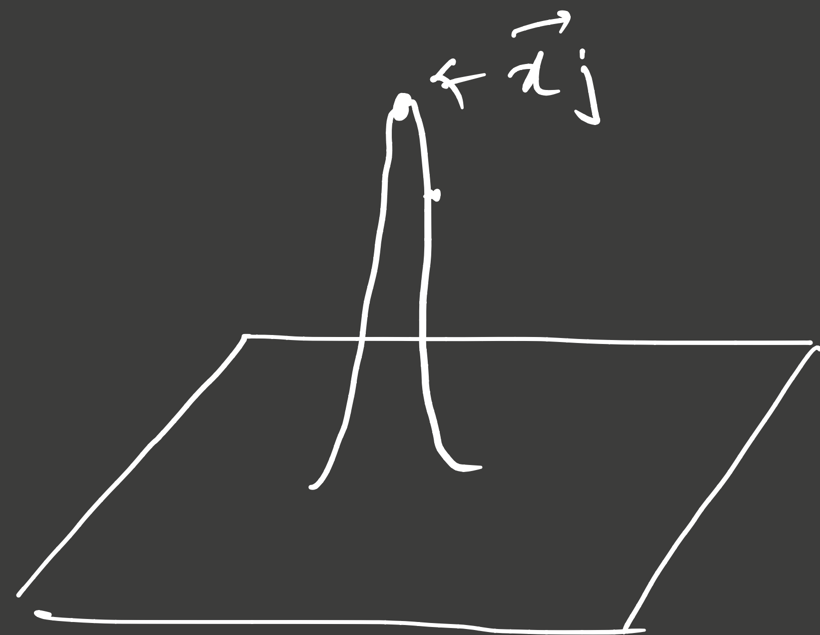
RBF : Effect of γ

γ: How far is the influence of a single training sample



γ = Low

High influence of $\bar{x}_j$

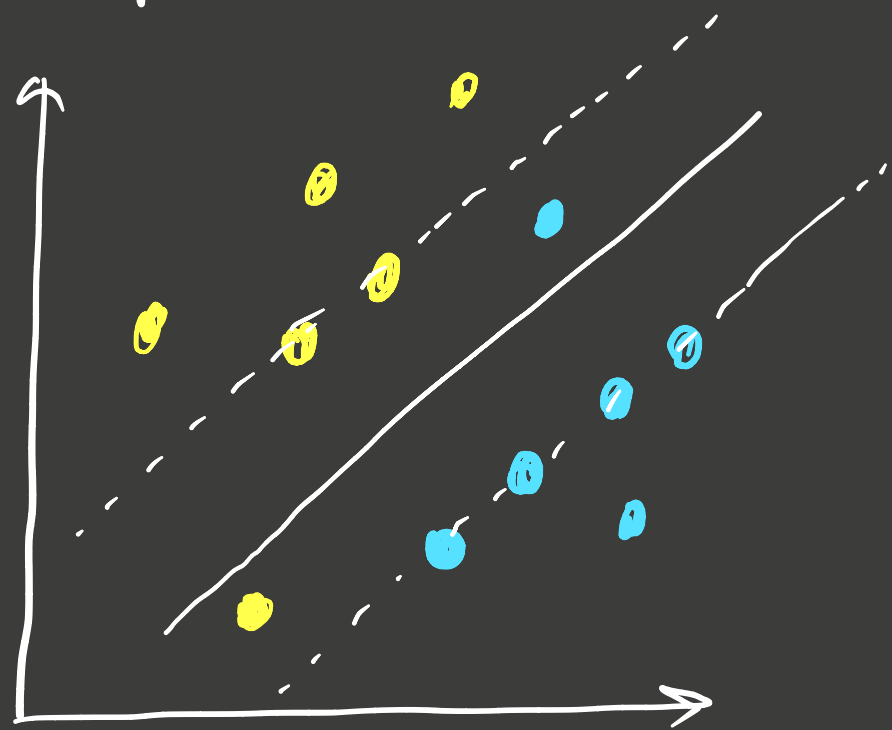


γ = High

low influence of $\bar{x}_j$

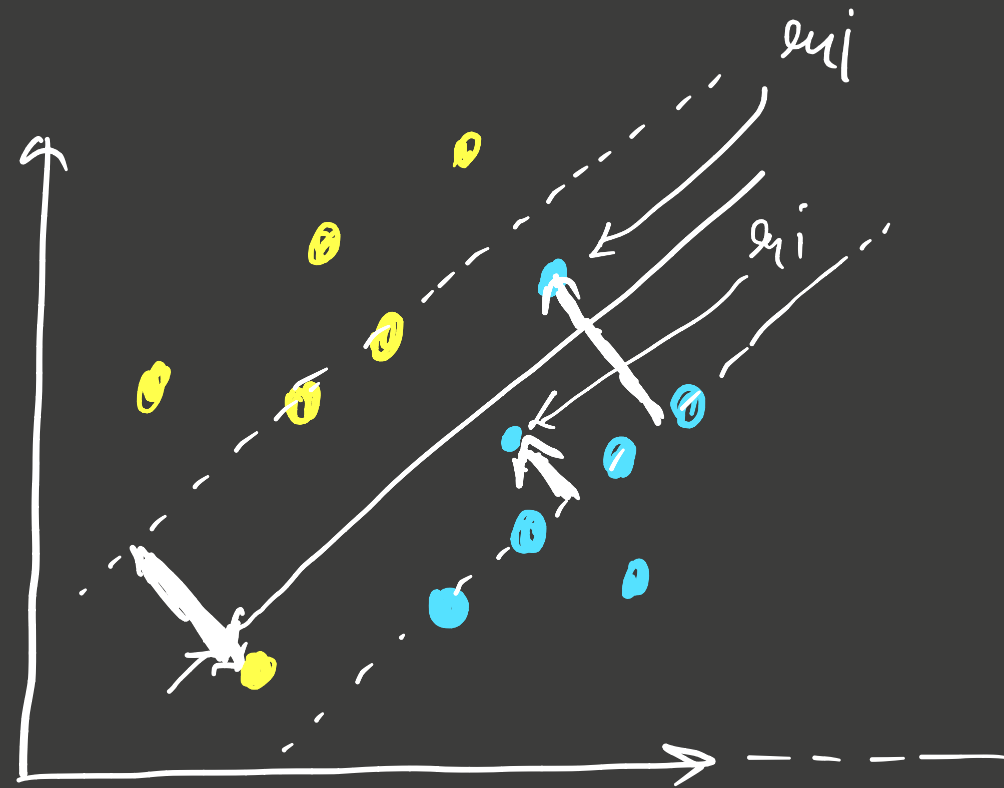
SOFT MARGIN SVM

Q:) Can we learn SVM for "slightly" non separable data without projecting to a higher space?



"Slightly" non separable data.

SOFT MARGIN SVM



SLACK VARIABLE

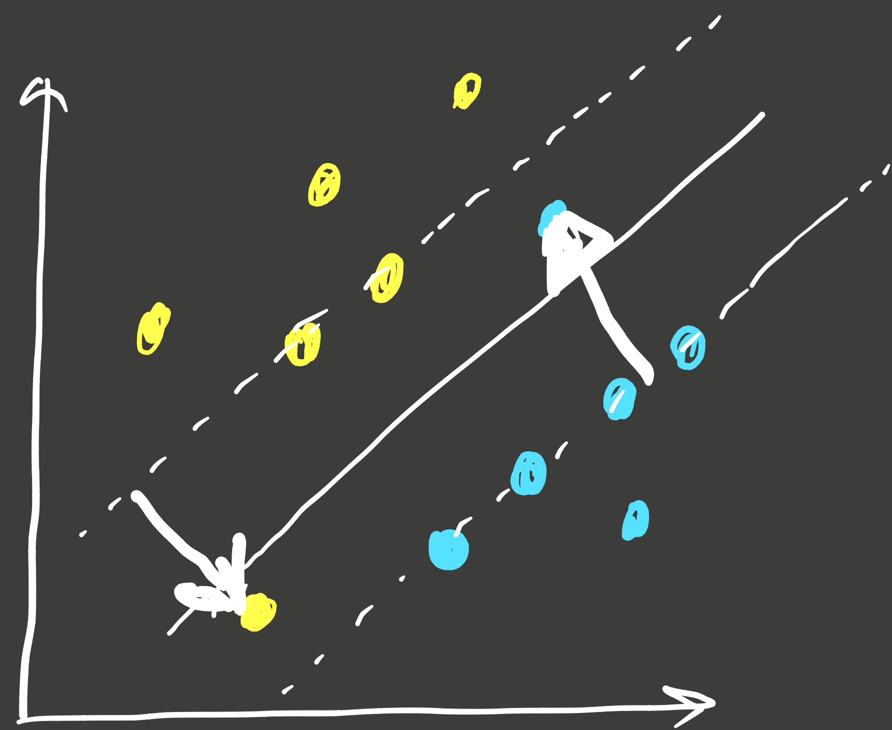
$$\xi_i = \begin{cases} 0 & \text{iif point on correct side of margin} \\ \text{Distance from margin} & \text{o/w.} \end{cases}$$

Change objective

$$\min \frac{1}{2}\|\vec{w}\|^2 + C \sum_{i=1}^{n} \xi_i$$

s.t. $\quad y_i\,(\bar{w}\cdot\bar{x}_i + b) \geq 1 - \xi_i \,;\; \xi_i \geq 0 \,;$

SOFT MARGIN SVM



Change objective

$$\min \frac{1}{2}\|\vec{w}\|^2 + C\sum_{i=1}^{n} \xi_i$$

s.t. $\quad y_i(\bar{w}\cdot\bar{x}+b) \geq 1-\xi_i$

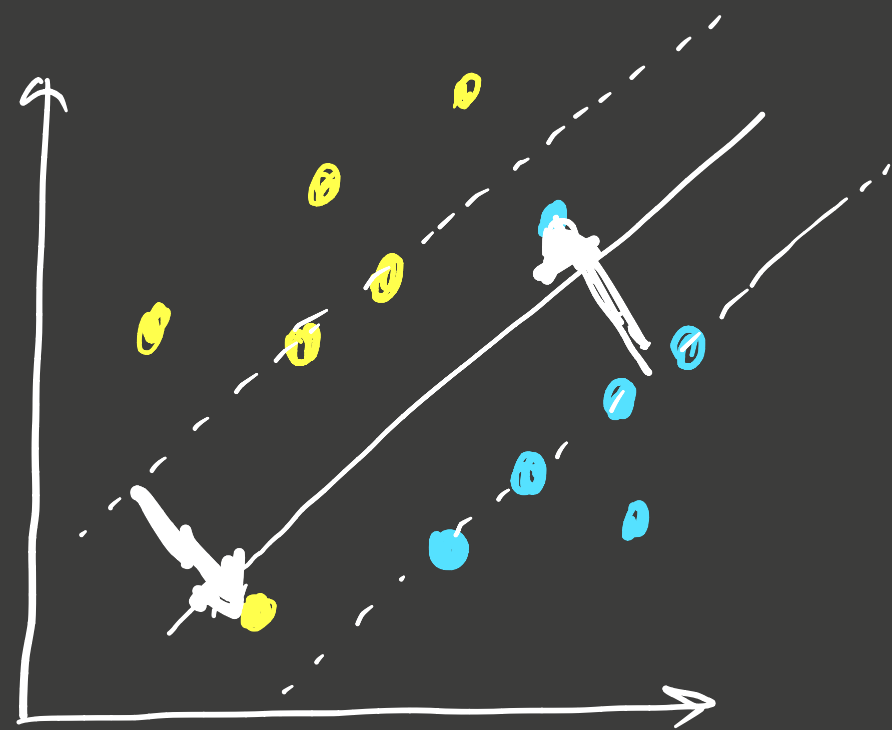
in Dual

Maximize $\quad \sum_{i=1}^{n}\alpha_i - \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \bar{x}_i\cdot\bar{x}_j$

s.t.

$$0 \leq \alpha_i \leq C \quad \& \quad \sum_{i=1}^{n}\alpha_i y_i = 0$$

SOFT MARGIN SVM



Change objective

$$\min \frac{1}{2}\|\vec{w}\|^2 + C\sum_{i=1}^{n} \xi_i$$

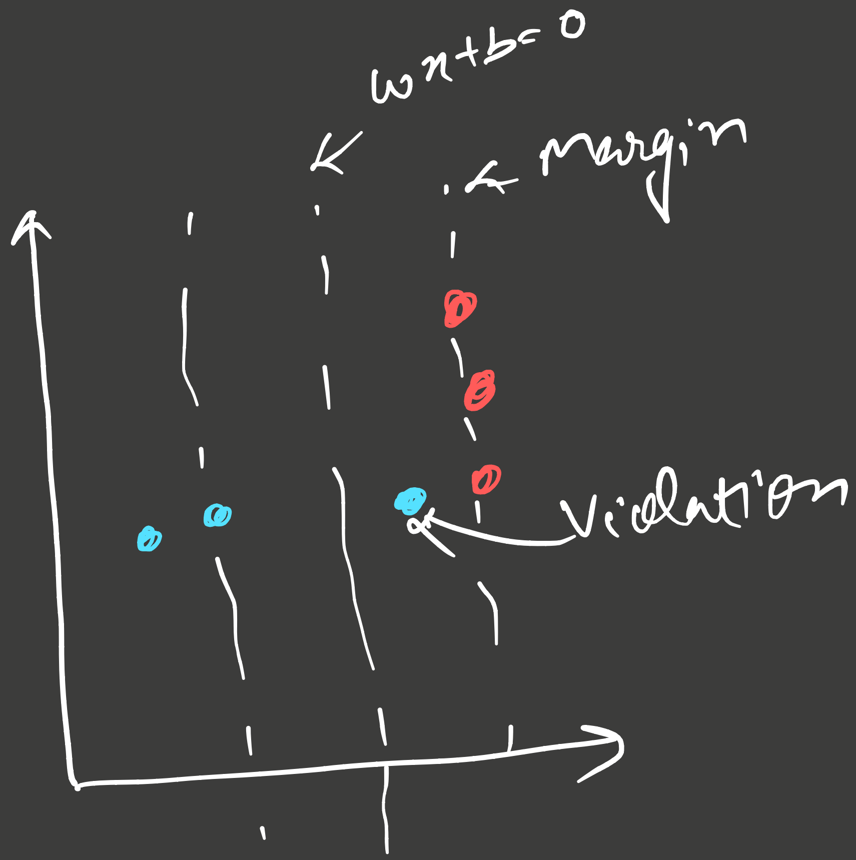
$$\text{s.t.} \quad y_i(\bar{w}\cdot\bar{x} + b) \geq 1 - \xi_i$$

$C \to 0$: Larger margin

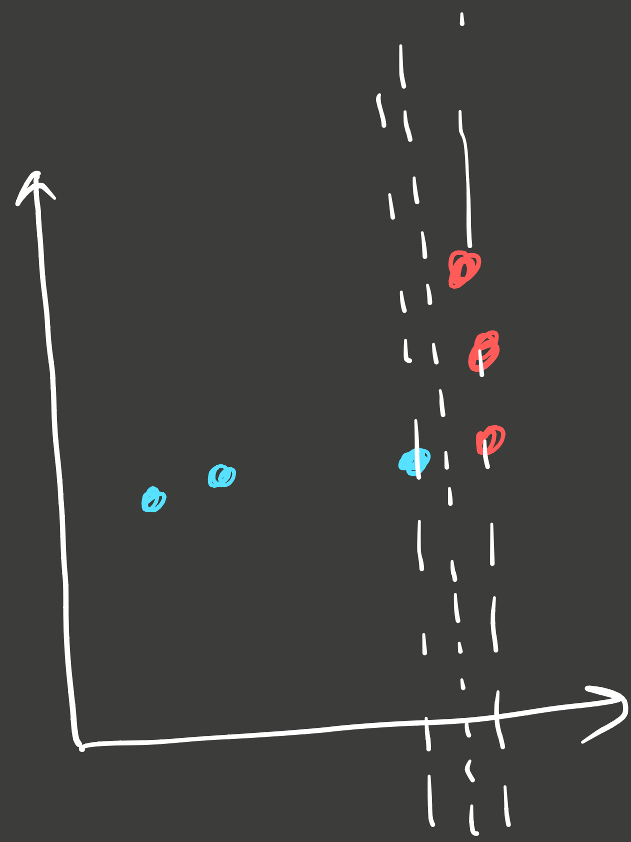
$C \to \infty$: smaller margin

$\boxed{\text{why?}}$

Notebook : svm-soft-margin



Low value of C



High value of C

If $C \to 0$

Objective $\to$ Min $\frac{1}{2}\|\vec{w}\|^2$

$\Rightarrow$ Choose large margin (without worrying for ξ_i's)

[Recall : Margin $= \frac{2}{\|\vec{w}\|}$]

If $C \to \infty$ (or very large)

Objective $\to$ Min $C \sum \xi_i$ or choose 'w', 'b', s.t. ξ_i is small!

Q) What is equivalent of hard margin?

a) $C \to 0$

b) $C \to \infty$

Q) what is equivalent of hard margin?

a) $C \to 0$

$\boxed{\text{b)} \quad C \to \infty}$

$\Downarrow$

No violations!!

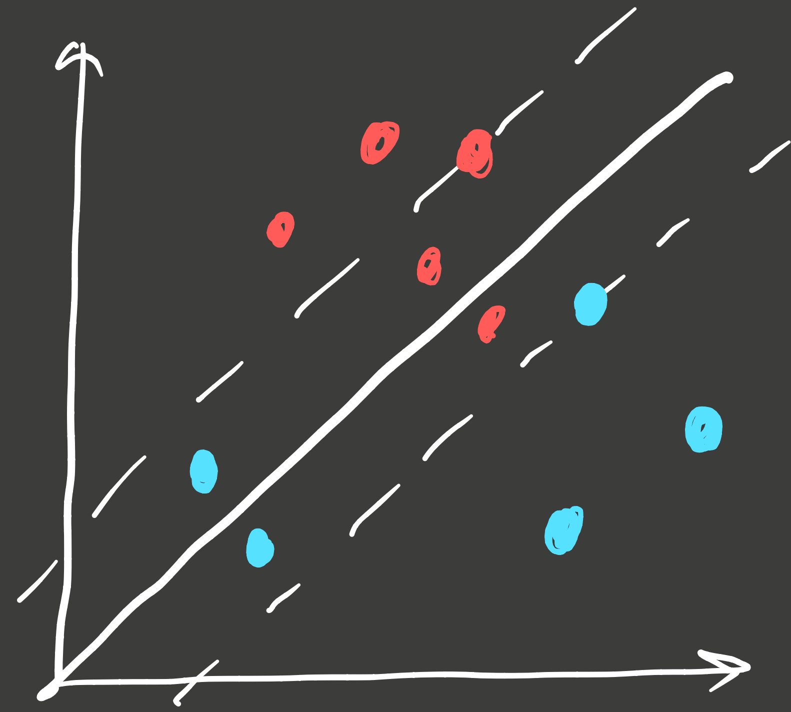
BIAS VARIANCE TRADE-OFF FOR SOFT-MARGIN SVM

Low C $\Rightarrow$ Higher train error
(Higher bias)

High C $\Rightarrow$ Very sensitive to dataset
(High variance)

SOFT MARGIN SVM

Types of support vectors



Zone 1 $y_i(\bar{w}\cdot\bar{x}_i+b)=1$

Zone 2 $0<\xi_i<1$ (correctly classified)

Zone 3 $\xi_i>1$ (Misclassified)

$\therefore$ As C increases, # support Vectors decreases

Notebook : svm-soft-margin

SVM FORMULATION IN LOSS + PENALTY FORM

Objective:

$$\min \ \frac{1}{2}\|\bar{w}\|^2 + C\sum_{i=1}^{N}\xi_i$$

s.t. $y_i(\bar{w}\cdot\bar{x}_i + b) \geq 1 - \xi_i \ ; \ \xi_i \geq 0$

Now: $y_i(\bar{w}\cdot\bar{x}_i + b) \geq 1 - \xi_i$

$$\xi_i \geq 1 - y_i(\bar{w}\cdot\bar{x}_i + b)$$

But $\xi_i \geq 0$

$$\therefore \ \xi_i = \text{MAX}\left[0, 1 - y_i(\bar{w}\cdot\bar{x}_i + b)\right]$$

∴ Objective is:

$$\text{Min} \quad C \sum \xi_i + \frac{1}{2}\|\vec{w}\|^2$$

Or

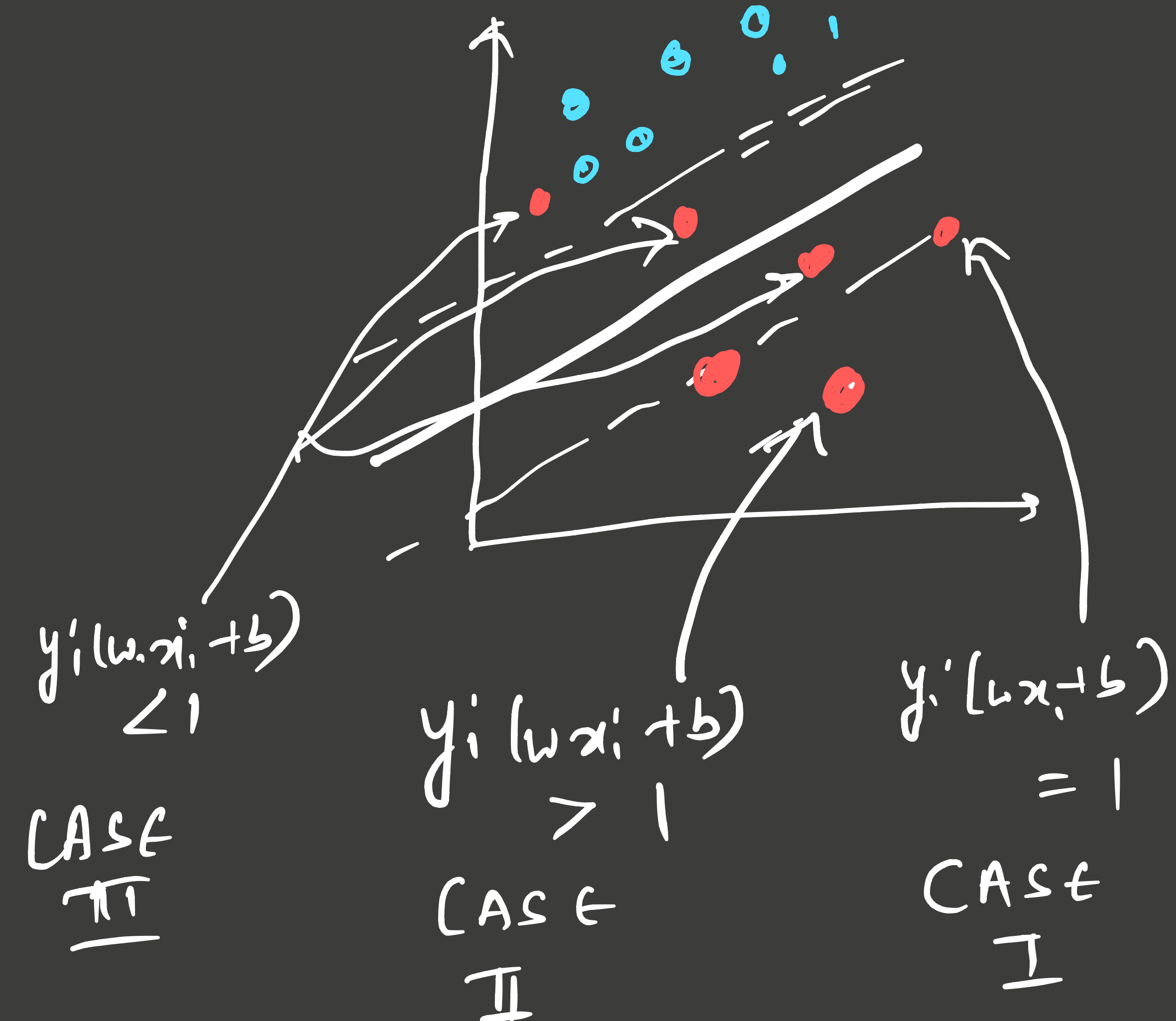
$$\text{Min} \quad C \sum_{i=1}^{N} \text{MAX}\left[0,\ 1 - y_i(\vec{w}\cdot\vec{x}_i + b)\right] + \frac{1}{2}\|\vec{w}\|^2$$

Or

$$\text{Min} \quad \underbrace{\sum_{i=1}^{N} \text{MAX}\left[0,\ 1 - y_i(\vec{w}\cdot\vec{x}_i + b)\right]}_{\text{LOSS}} + \underbrace{\frac{1}{2C}\|\vec{w}\|^2}_{\text{REGULARISATION}}$$

LOSS FUNCTION FOR SVM (HINGE LOSS)

LOSS FUNCTION IS: $\sum_{i=1}^{N} \text{MAX}\left[0,\ 1 - y_i \cdot [\bar{w} \cdot \bar{x}_i + b]\right]$



__Case I__

$$y_i(\vec{w} \cdot \vec{x}_i + b) = 1$$

$\boxed{\text{LIES ON MARGIN}}$ $\text{Loss}_i = 0$

__Case II__

$$y_i(\bar{w} \cdot \bar{x}_i + b) > 1$$

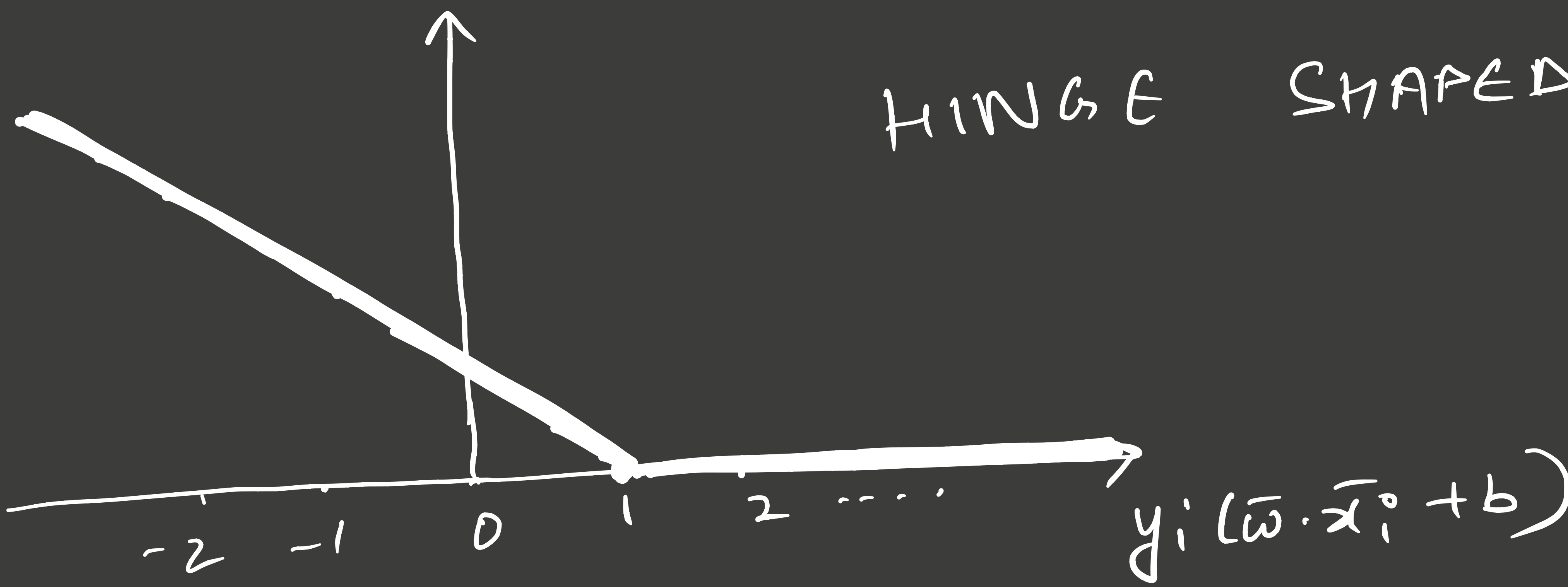
$\text{Loss}_i = 0$

__Case III__

$$y_i(\bar{w} \cdot \bar{x}_i + b) < 1$$

$\text{Loss}_i \neq 0$

HINGE LOSS CONTINUED



HINGE SHAPED

Q) IS HINGE LOSS CONVEX & DIFFERENTIABLE?

CONVEX: ✓

DIFFERENTIABLE: X

SUBGRADIENT: ✓

SVM LOSS IS CONVEX

HINGE LOSS $\Sigma(\text{MAX}\,[0, (1 - y_i(\bar{w} \cdot x_i + b)]$

IS CONVEX

PENALTY $\frac{1}{2} \|\bar{w}\|^2$

IS CONVEX

$\therefore$ SVM LOSS IS CONVEX