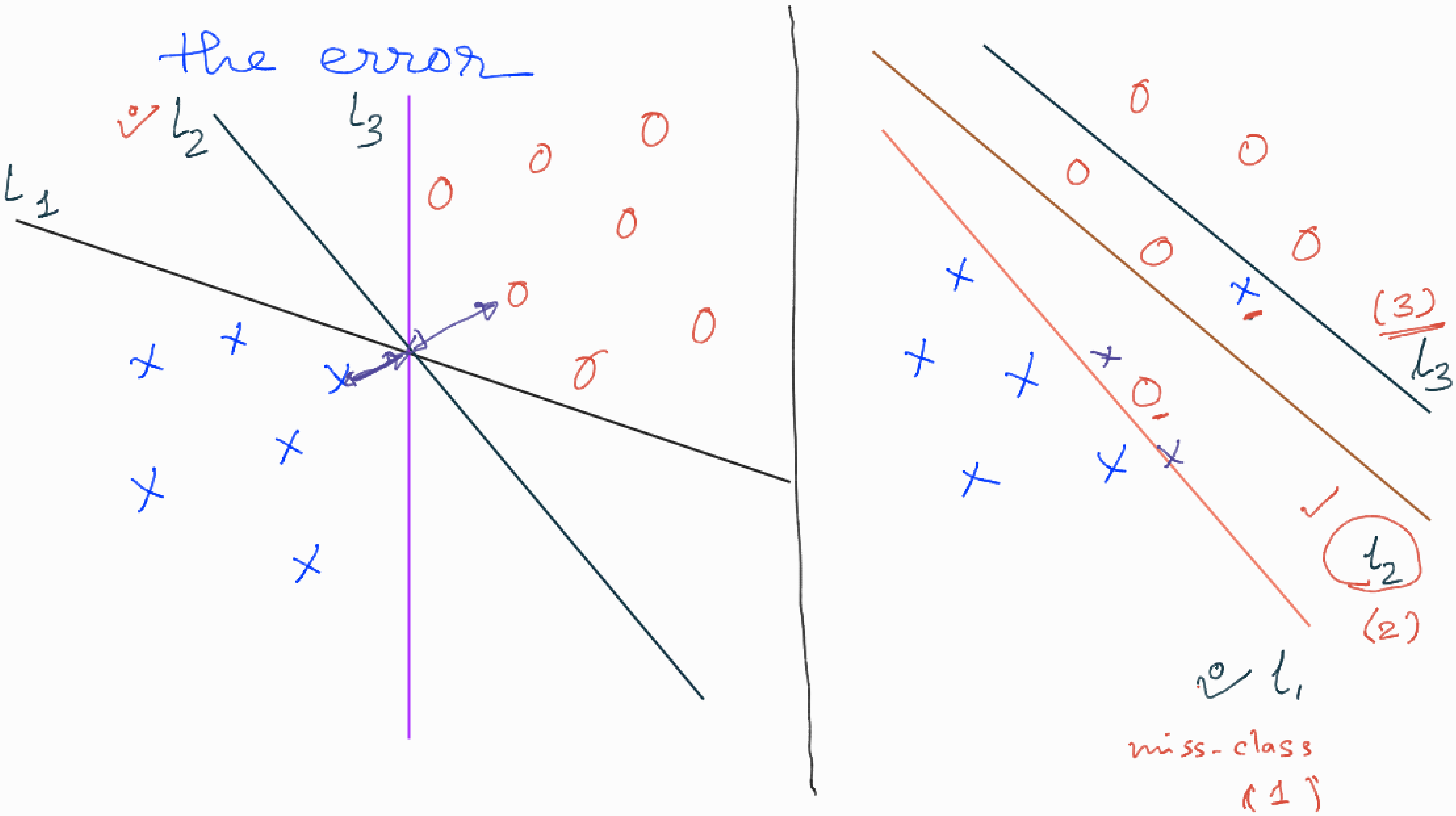
Linear classifier: (Error Correction)

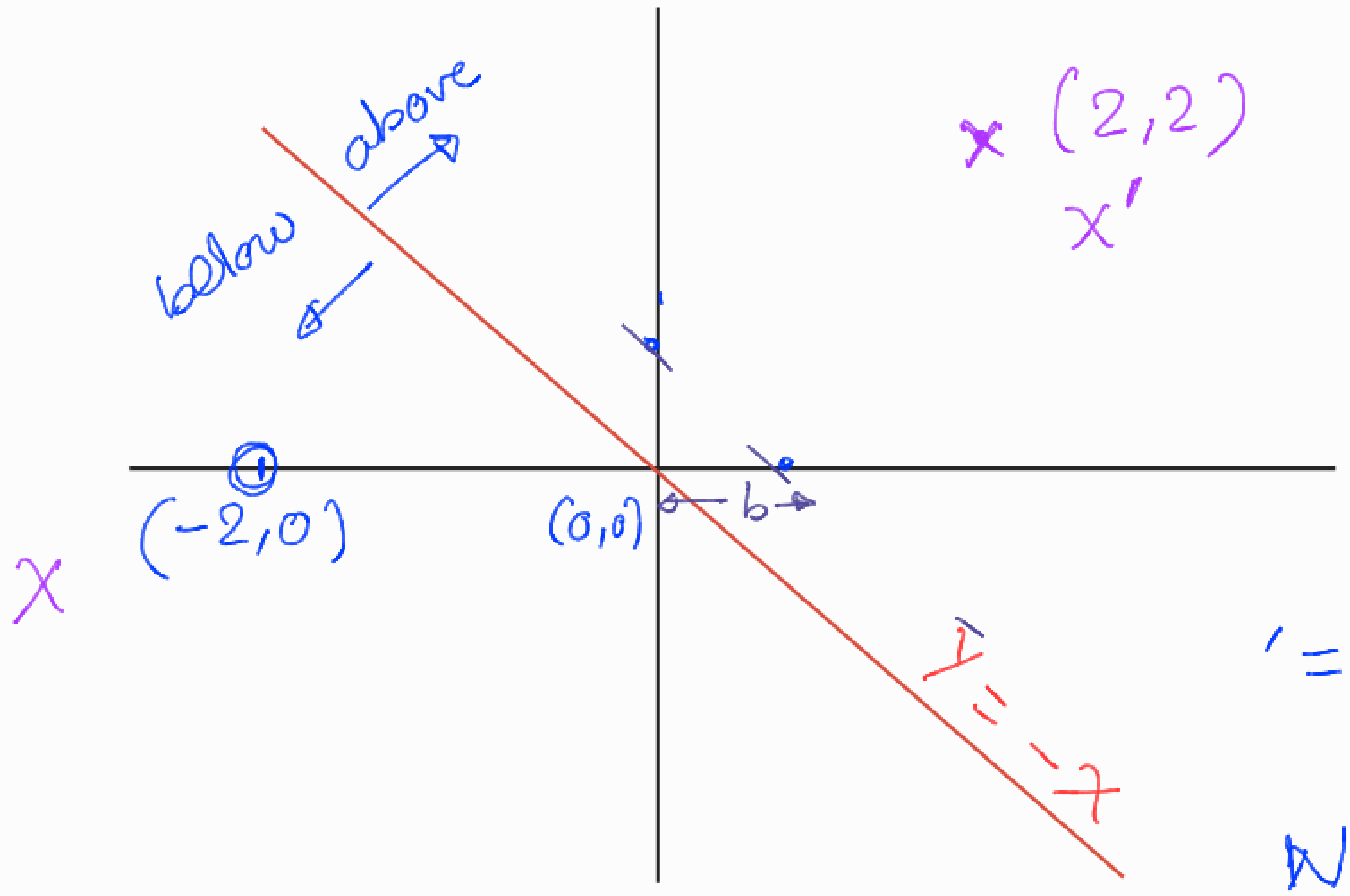
- Start with initial parameters
- Calculate errors
- Update parameters to minimize the error



Support Vector Machine (SVM)

- Best separating hyperplan with maximum distance from class boun- -daries.

- It models the problem of classification as an optimization problem which can be solved by quadratic programming



$' = \underline{m}x + \underline{c}$

$W' = [m, c]$

$= [-1, 0]$

$X = (-2, 0)$

$$W^T \cdot X = \begin{bmatrix} -1 \\ 0 \end{bmatrix} [-2, 0]$$

$$= \underline{\underline{2}} > 0$$

$X' = (2, 2)$

$$W^T \cdot X' = \begin{bmatrix} -1 \\ 0 \end{bmatrix} [2, 2]$$

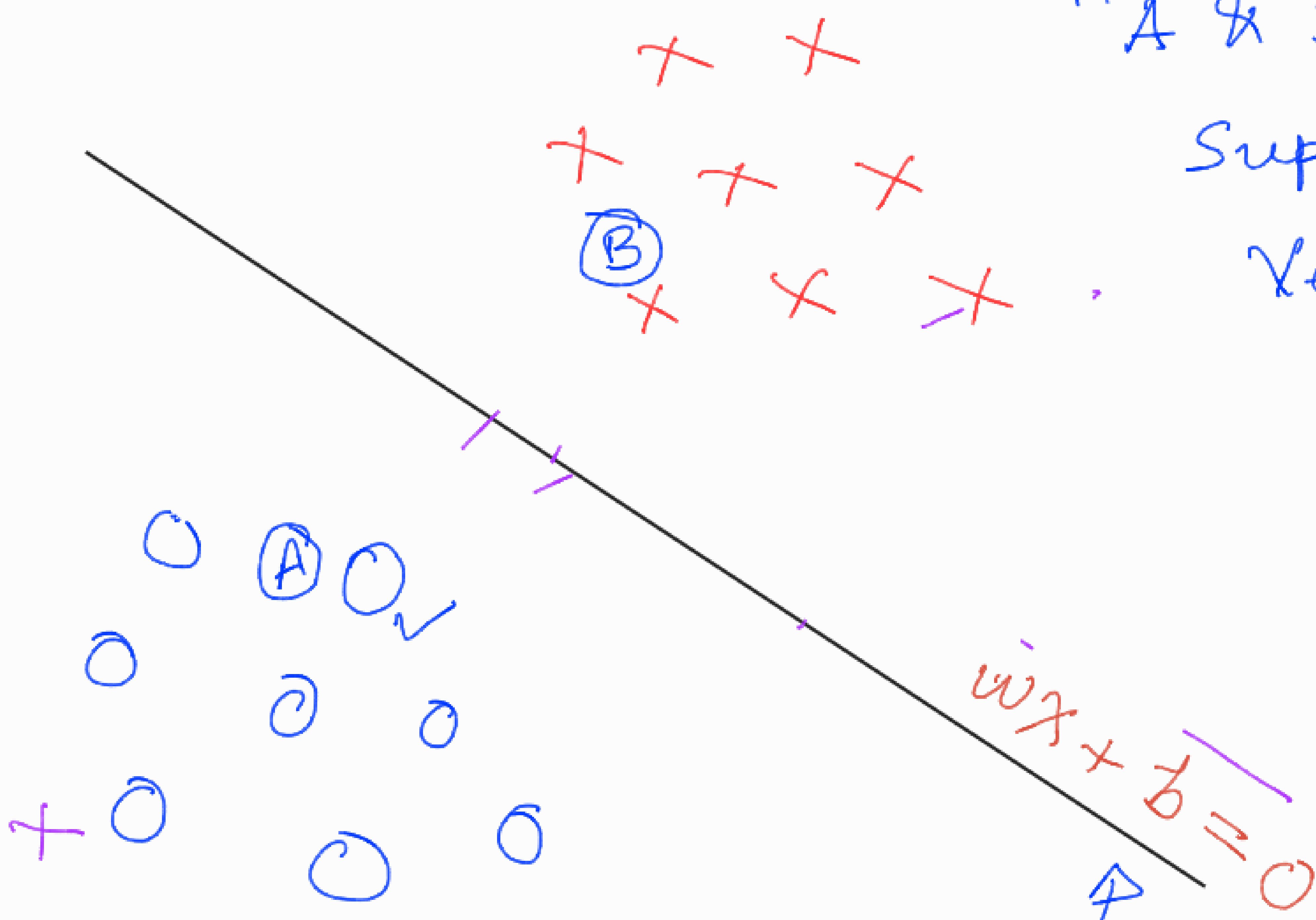
$$= -2 < 0$$

All datapoints for which

$w^T x < 0 \Rightarrow$ above

$w^T x > 0 \Rightarrow$ below

$w^T x = 0$



"A & B are Support Vectors"

"Farthest from the nearest"

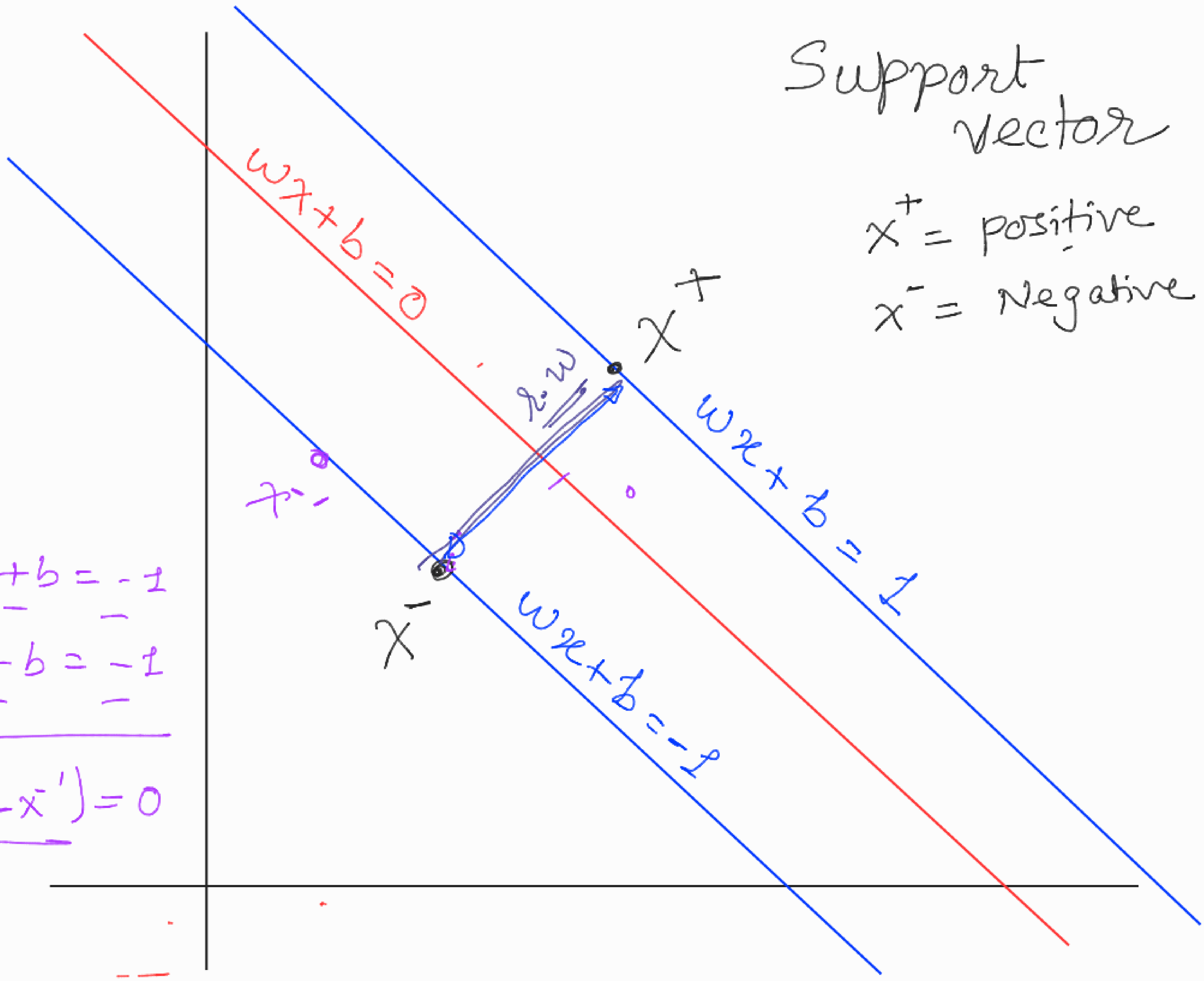
"This line should be as far as possible from both the classes"

- Find out support vectors & a separting line (hyperplane)

Support vector

x^+ = positive

x^- = Negative



$$wx^- + b = -1$$
$$wx' + b = -1$$
$$w.(x^- - x') = 0$$

$$wx^+ + b = 1 \quad (1) \qquad x^+ = r\cdot w + x^- \quad (3)$$

$$wx^- + b = -1 \quad (2)$$

Sub x^+ from (3) to (1)

$$w\,(r\cdot w + x^-) + b = 1$$

$$r\cdot\|w\|^2 + \underbrace{w\cdot x^- + b}_{(2)} = 1$$

$$r\cdot\|w\|^2 + (-1) = 1$$

$$r \cdot \|w\|^2 = 2$$

$$\boxed{r = \frac{2}{\|w\|^2}}$$

Optimization problem

max. $r = \dfrac{2}{\|w\|^2}$

given constraints,

$wx + b \geq 1$ if $y = 1$

$wx + b \leq -1$ if $y = -1$

$$\max \frac{2}{\|w\|^2} \approx \min \frac{\|w\|^2}{2} \quad \text{— (1)}$$

$$\left.\begin{array}{l} wx + b \geq 1 \text{ if } y = 1 \\ wx + b \leq -1 \text{ if } y = -1 \end{array}\right\} \Rightarrow y(wx + b) \geq 1 \quad \text{— (2)}$$

Restate the problem

$$\boxed{\begin{array}{l} \text{opti : } \min \dfrac{\|w\|^2}{2} \\ \text{Costraint : } \sum_i y_i(wx_i + b) - 1 \geq 0 \end{array}}$$

An Example:

$$= \frac{2}{3} \times \frac{1}{\sqrt{3}} = \boxed{\frac{2}{3\sqrt{3}}}$$

max. $f(x, y) = x^2 y$

Constraint: $g(x, y) = \underline{x^2 + y^2 - 1 = 0}$

$$\frac{\partial f(x,y)}{\partial x} = 2xy \propto \frac{\partial g(x,y)}{\partial x} = 2x$$

$$\frac{\partial f(x,y)}{\partial y} = x^2 \propto \frac{\partial g(x,y)}{\partial y} = 2y$$

Lagrange Multiplier

$$2xy = \underline{\lambda}\, 2x$$

some constant

Lagrange Multiplier

$$\boxed{\begin{aligned} \frac{\partial f(x,y)}{\partial x} &= \lambda \frac{\partial g(x,y)}{\partial x} \\ \frac{\partial f(x,y)}{\partial y} &= \lambda \frac{\partial g(x,y)}{\partial y} \end{aligned}}$$

$\therefore \quad 2xy = \lambda 2x$

$\therefore \quad \boxed{\lambda = y}$ ①

$x^2 = \lambda 2y$

$\boxed{x^2 = 2y^2}$ ②

$x^2 + y^2 = 1$

$2y^2 + y^2 = 1$

$\therefore \quad \boxed{y = 1/\sqrt{3}} \qquad \boxed{x = \sqrt{2/3}}$

Way-1

$f(x)$ optimized ; $g(x)$ constraint

$$\frac{\partial f(x)}{\partial x} = \lambda \frac{\partial g(x)}{\partial x}$$

$$\boxed{\frac{\partial f(x)}{\partial(x)} - \lambda \frac{\partial g(x)}{\partial x} = 0}$$

Way-2

$$L(x) = \underbrace{f(x) - \lambda g(x)}$$

$$\boxed{\frac{\partial L(x)}{\partial x} = 0}$$

Single equation

- Convert optimization problem in Lagrange Format,

$$L(x, y \ldots) = f(x, y \ldots) - \lambda g(x, y \ldots)$$

- Take $\frac{\partial L(x, y \ldots)}{\partial x / y / \ldots} = 0$ and find values of parameters.

SVM

$$\boxed{\min \frac{\|w\|^2}{2}, \quad \text{given} \sum_i \frac{y_i(wx_i+b) - 1}{}}$$

Convex

$$\boxed{L(w, b) = \frac{\|w\|^2}{2} - \sum_i \alpha_i [y_i(wx_i + b) - 1]}$$

① ② ③

$$L(w, b) = \frac{\|w\|^2}{2} - \sum_i (\alpha_i y_i w x_i + \alpha_i y_i b - \alpha_i) \quad \text{— Ⓐ}$$

$$\frac{\partial L(w, b)}{\partial w} = w - \sum_i \alpha_i y_i x_i = 0$$

$$\boxed{w = \sum_i \alpha_i y_i x_i} \quad ①'$$

$$\frac{\partial L(w,b)}{\partial b} = \boxed{-\sum_i \alpha_i y_i = 0} \quad ②$$

Plug ① & ② in Ⓐ

$$L(w,b) = \frac{1}{2}\left(\sum_i \alpha_i y_i x_i\right)\left(\sum_j \alpha_j y_j x_j\right)$$

$$- \sum_i \alpha_i y_i x_i \cdot \left(\sum \alpha_j y_j x_j\right)$$

$$\underbrace{- \sum_i \alpha_i y_i b}_{=0} + \sum_i \alpha_i$$

$$L(w,b) = \frac{1}{2}\left(\sum_i \alpha_i y_i x_i\right)\left(\sum_j \alpha_j y_j x_j\right)$$
$$- \left(\sum_i \alpha_i y_i x_i\right)\left(\sum_j \alpha_j y_j x_j\right) + \sum_i \alpha_i$$

$$L(w,b) = -\frac{1}{2}\left(\sum_i \alpha_i y_i x_i\right)\left(\sum_j \alpha_j y_j x_j\right) + \sum \alpha_i$$

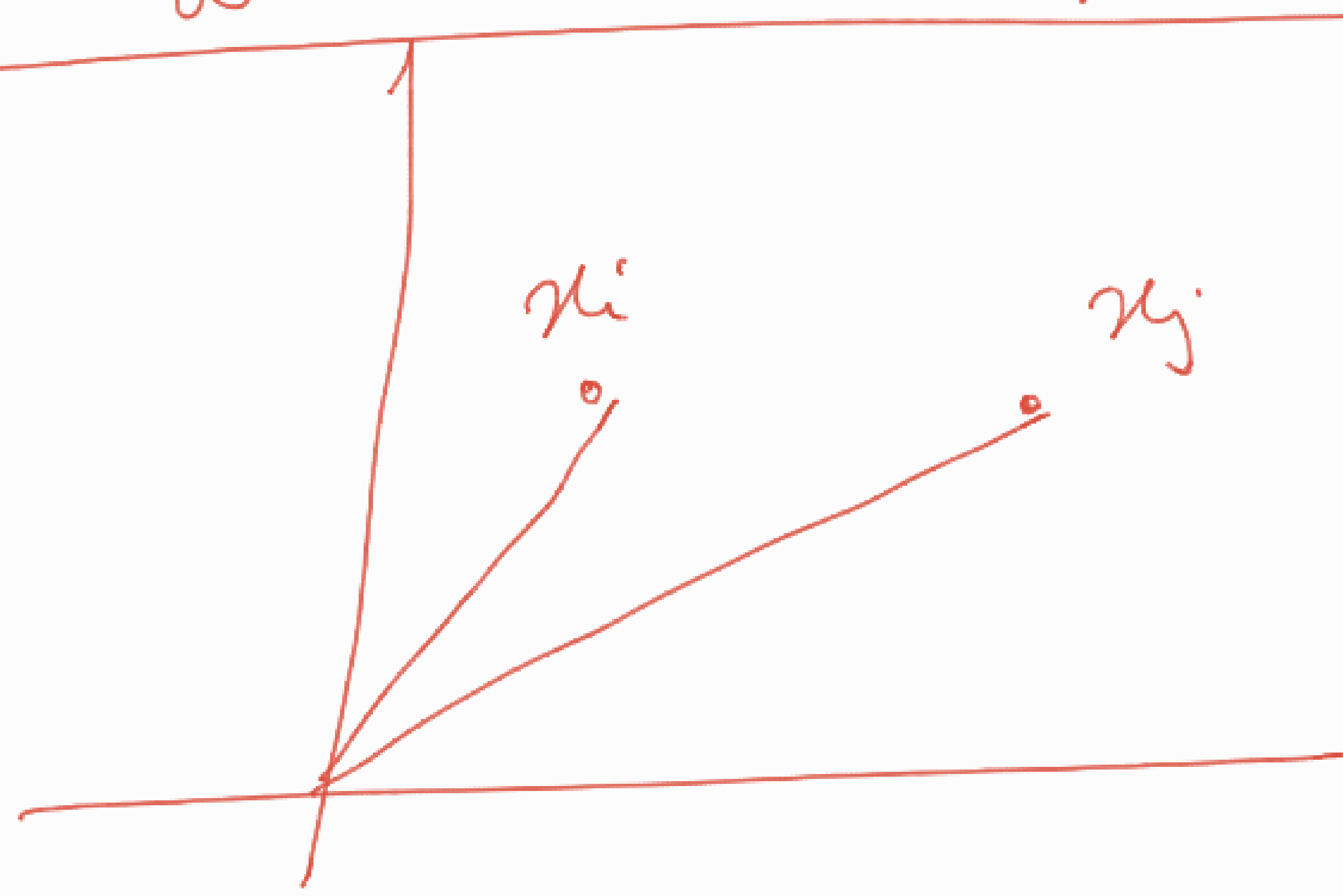
$K(x_i, x_j)$

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

$$L(w,b) = -\frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y_i y_j \boxed{x_i \cdot x_j} + \sum \alpha_i$$

Every pair of example

$\alpha_i = 0$ for datapoints which are not support vectors



$\vec{x_i} \cdot \vec{x_j}$

$\Rightarrow$ $\vec{x_i} \cdot \vec{x_j}$ helps in finding best separating hyperplane.

$\Rightarrow$ Convex optimization, $\Rightarrow$ The minimum or maximum that we find are actually global min/max

Kernel

$$\Rightarrow \underline{K(x_i, x_j)} = \underline{\phi(x_i) \cdot \phi(x_j)}$$

Tranformation

if we know K, we don't need to define/know ϕ.

Primal form

$$L(w, b) = \frac{\|w\|^2}{2} - \sum_i \alpha_i [y_i(wx_i + b) - 1]$$

Dual

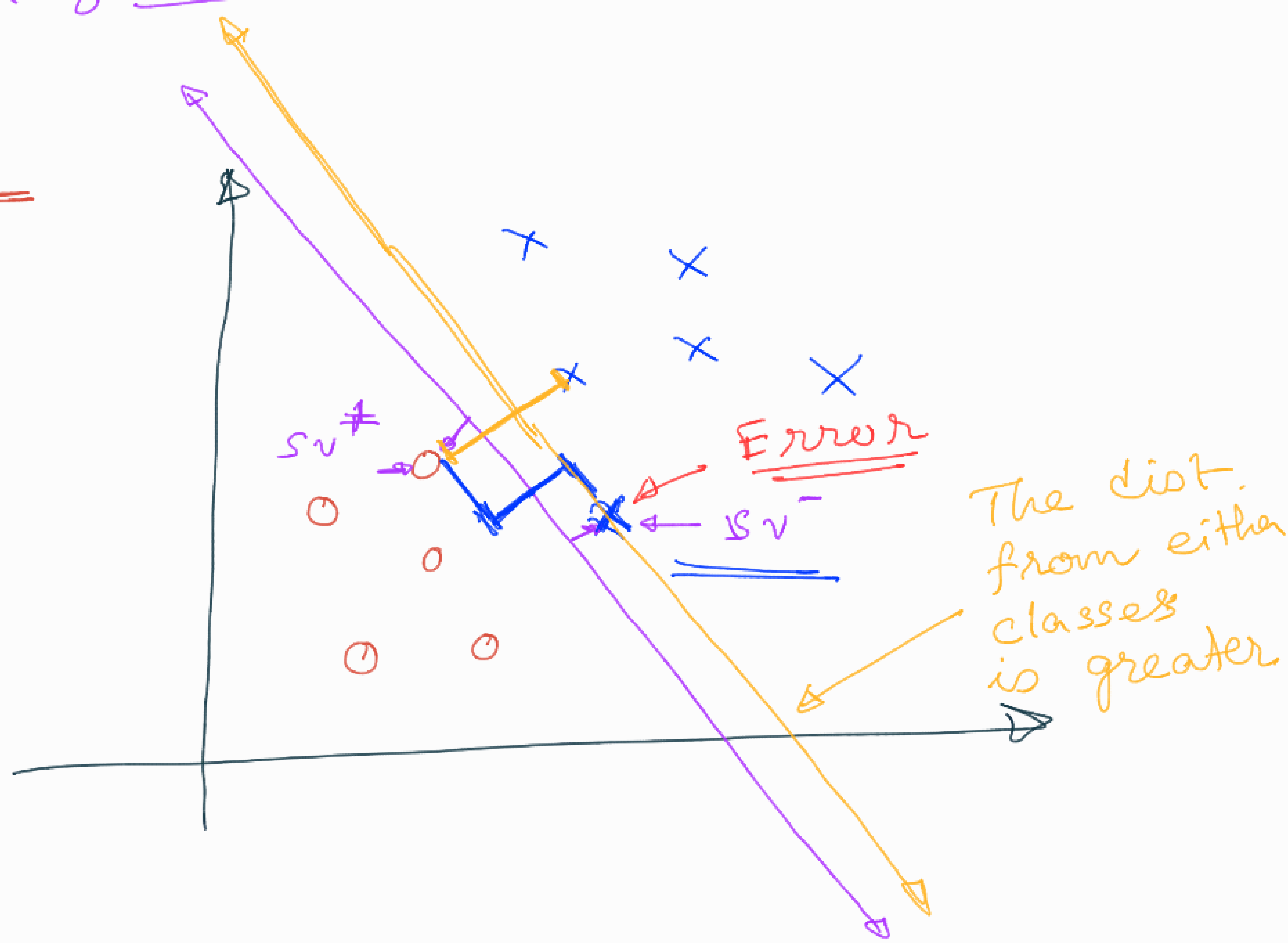
$$L(\cancel{w, b})_{\alpha} = -\frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j \underline{K(x_i, x_j)} + \sum \alpha_i$$

Lagrange multi-

Finding w, b $\Rightarrow$ finding α_i

- How do we learn α? ✓ Algorithm
- What about errors? ✓
- What about linearly inseparable data? ✓

Errors



"Yellow line separates red & blue data points better than the purple line."

$\Rightarrow$ Allow some error to find better separating hyperplanes.

Regularization parameter

Error

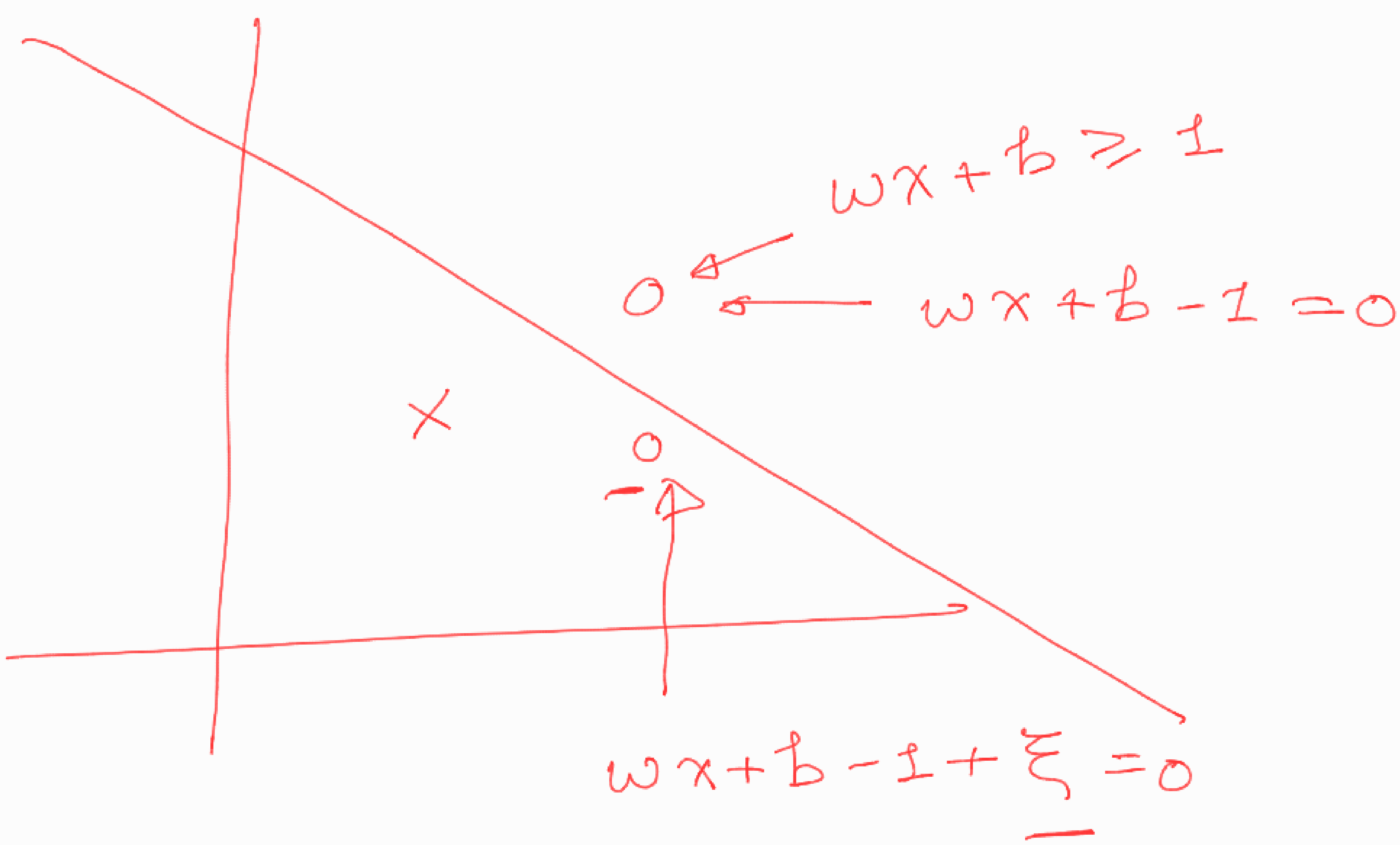
$$\min \left(\frac{\|w\|^2}{2} + C \sum_{i=1}^{m} \xi_i \right)$$

For each example

$$\text{s.t.} \quad y(wx+b) \geq 1$$

$$L = \underbrace{\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m} \xi_i}_{\text{Equation with error parameter}} - \sum_{i=1}^{m} \alpha_i \left(y(wx+b) - 1 + \xi_i \right) - \sum_{i=1}^{m} r_i \xi_i$$

r_i Lagrange



$$\sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y_i y_j \alpha_i \alpha_j \boxed{(x_i \cdot x_j)}$$

$$\sum_{i=1}^{m} \alpha_i y_i = 0$$

Learning Algorithm

- Quadratic Programming
- Sequential Minimal Optimization (SMO), Platt's algo.

- start will all α initialize to zero.
- Iteratively change values of pairs of α, to optimize the function.

why pair of α ?

$\boxed{\text{KKT condition}}$

$$\sum \alpha_i y_i = 0$$

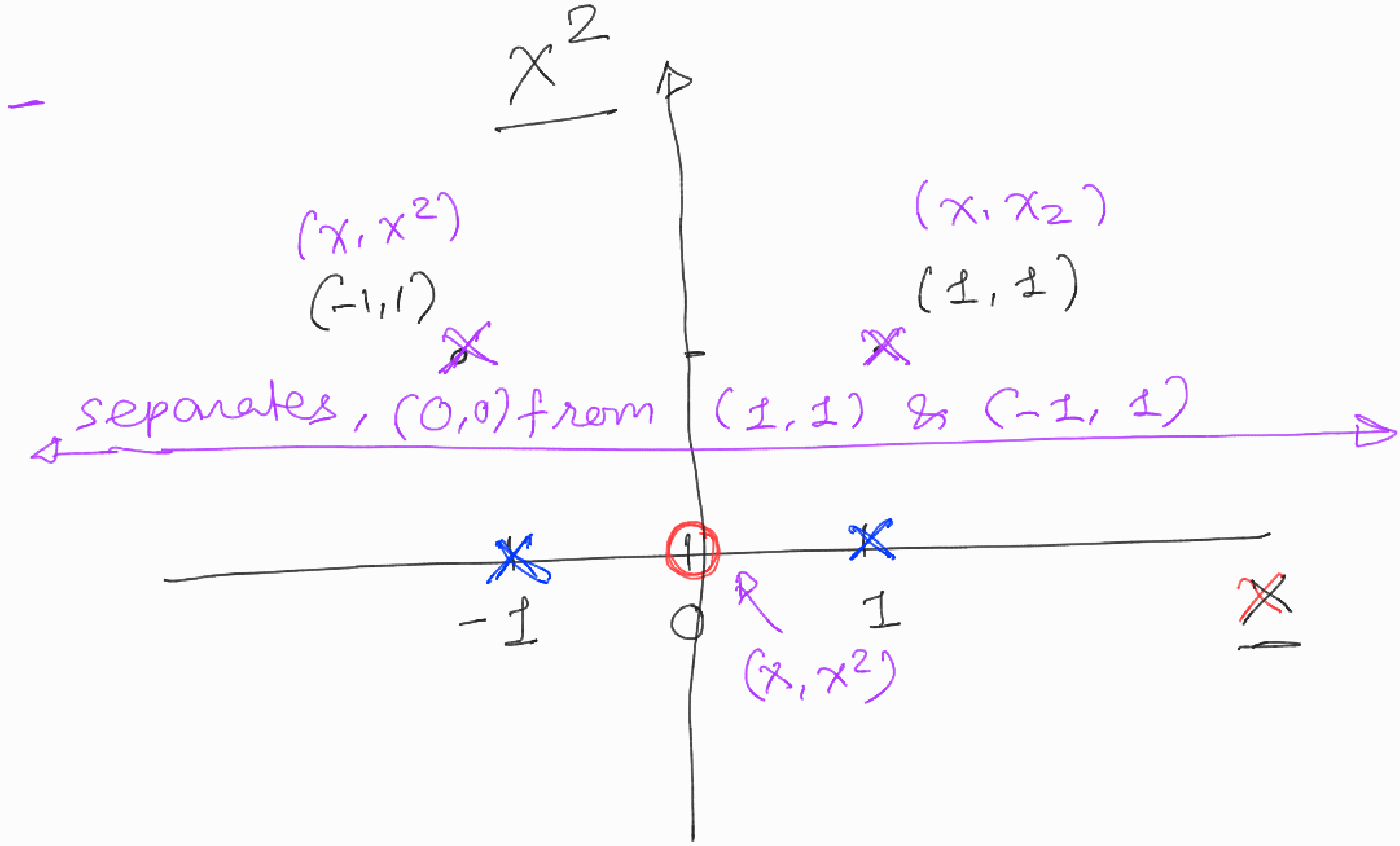
$$\therefore \underline{\alpha_m y_m} + \sum \alpha_{i/\alpha_m} \cdot y_{i/y_m} = 0$$

Exact answer, given all other α_is.

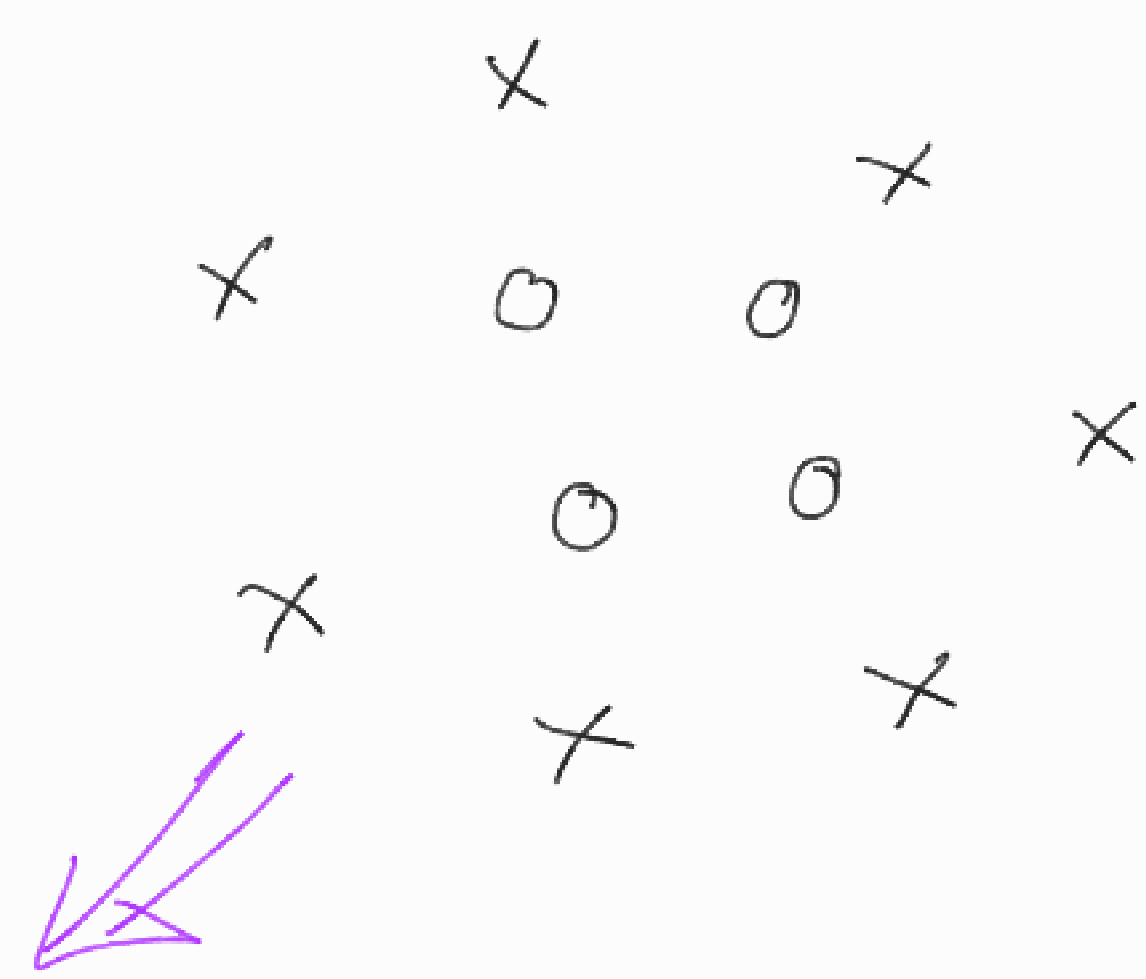
$$\underline{\alpha_m} y_m = - \sum \alpha_{i/\alpha_m} \cdot y_{i/y_m}$$

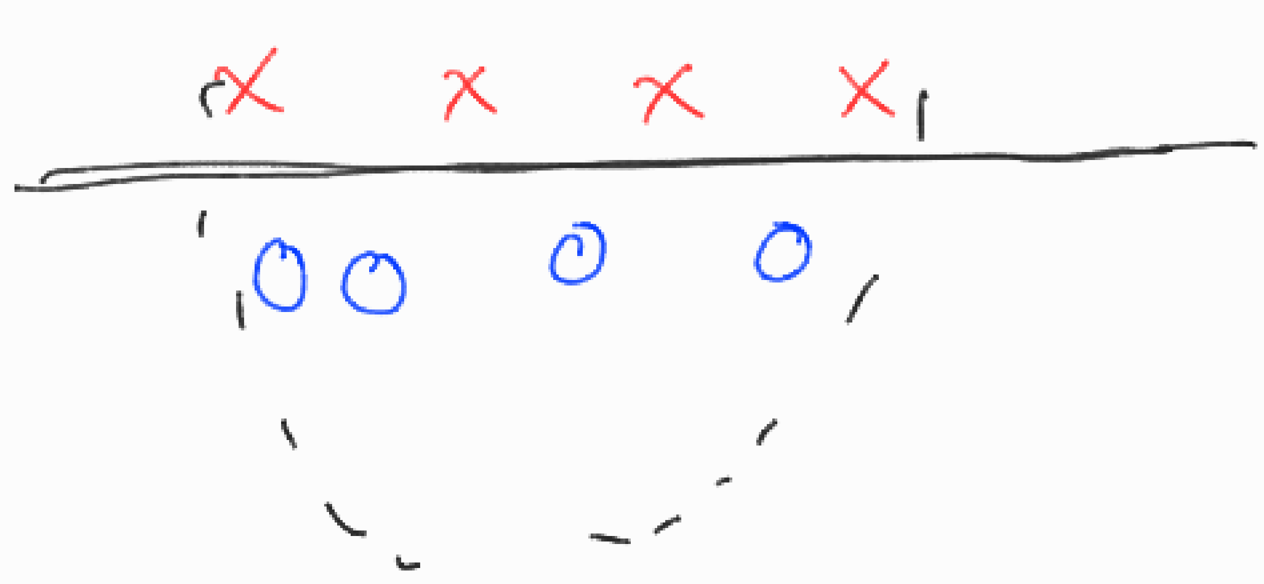
$$\alpha_m y_m + \alpha_n y_n + \sum_i \alpha_{i/\alpha_m, \alpha_n} y_{i/y_m y_n}$$

- How to allow some error.
- Learning process.
- Linearly non-separable data



"We have moved data to higher dimension to find better separating hyper plane."

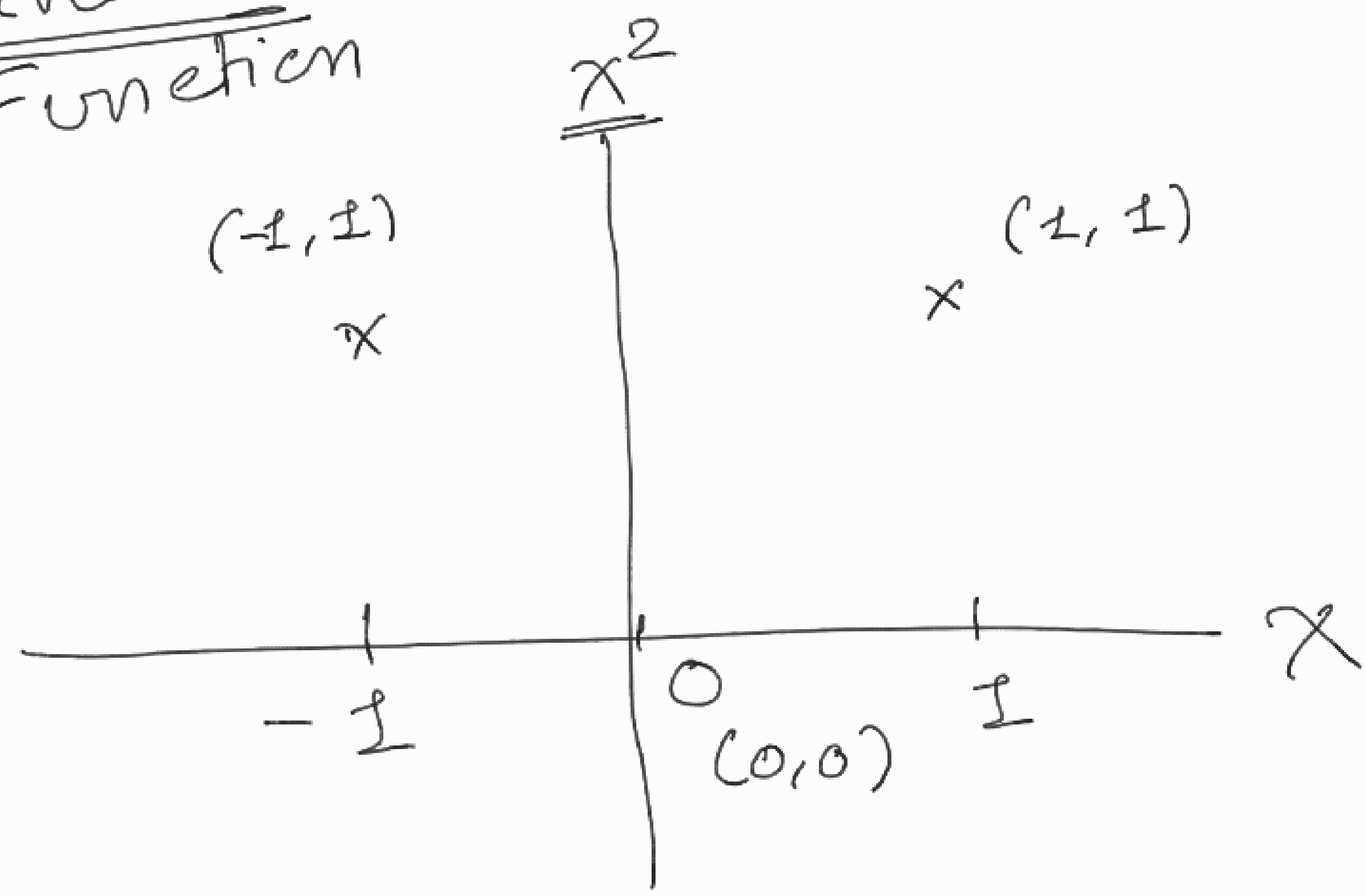




$$k(x, y) = \phi(x) \cdot \phi(y)$$

"Kernal" Function

Transformation on x.



Polynomial kernal $\left(a \cdot b + \frac{1}{2}\right)^2$

$$a^2 \cdot b^2 + a \cdot b + 1/4$$

$(x^2 \cdot y^2)$

"Transformed"

$\boxed{x \cdot y}$

Original before the transfor-mation

Polynomial kernel

$$K(x_i, x_j) = \left(x_i \cdot x_j + 1/2\right)^2$$

Gaussian Kernel:

$$K(x_i, x_j) = \exp\left(-\frac{|x_i - x_j|^2}{2\underline{\sigma}^2}\right)$$

Radial Basis Function (RBF)

$$K(x_i, x_j) = \exp\left(-\underline{\gamma}|x_i - x_j|^2\right)$$

Sigmoid, hyperbolic Tangent

SVM parameters

Regu. $\rightarrow C$, $\gamma \leftarrow$ kernel, parameters of diff. kernels.