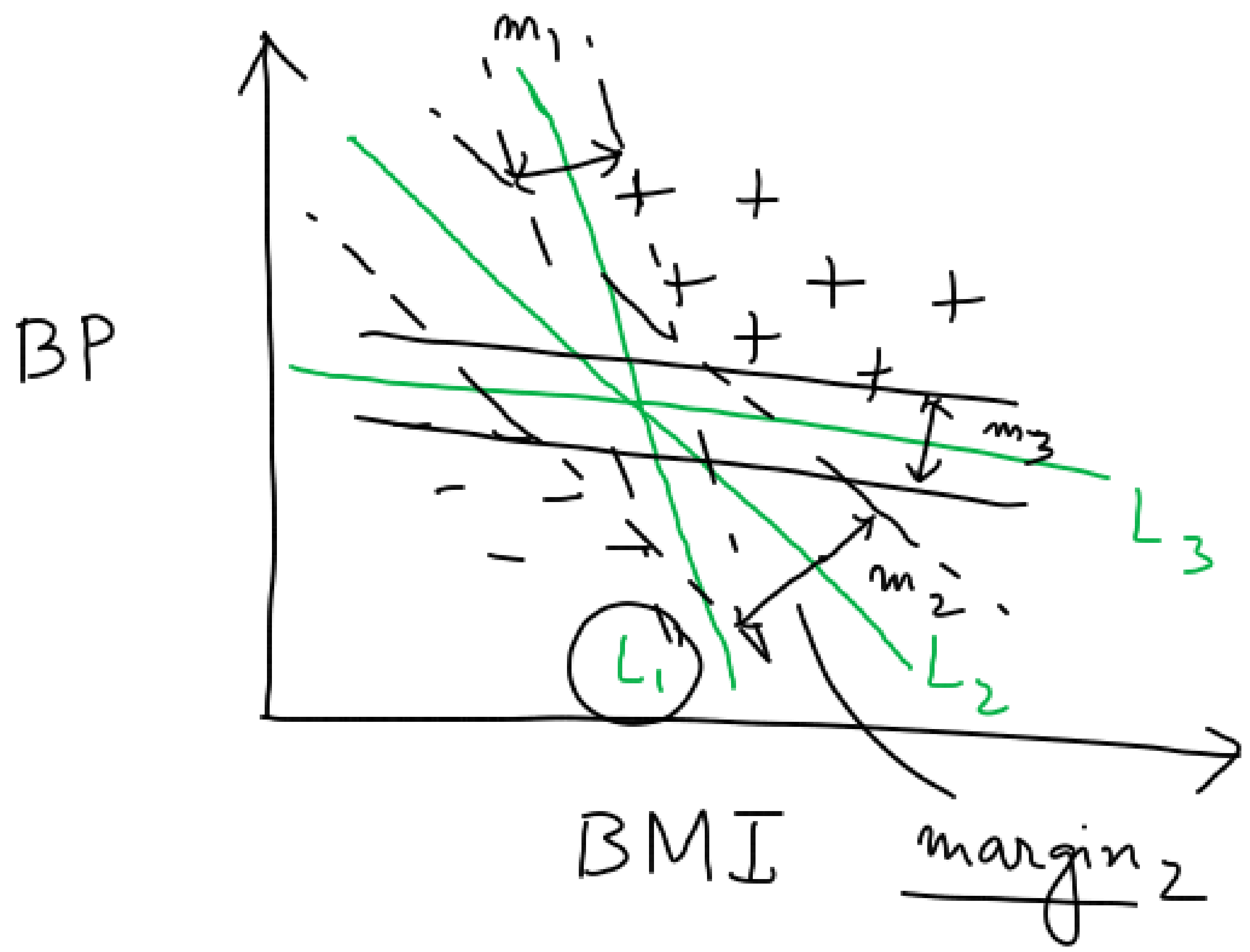
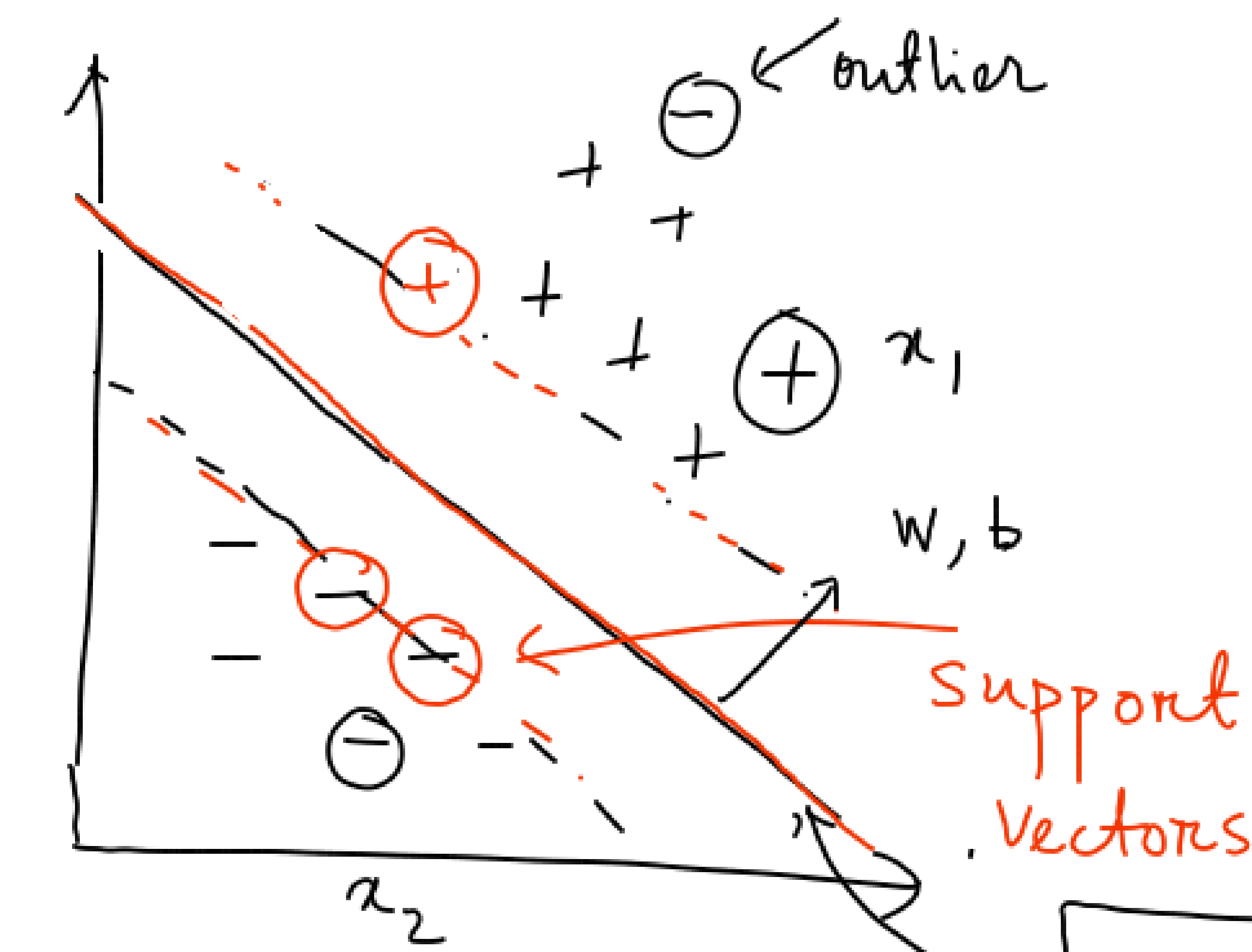
Lec 15: Support Vector Machines (SVM)



margin-based classifier.

$m_2 > m_1, m_3$



maximizes the margin

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

$$w^T x_1 + b > 0$$

$$w^T x_2 + b < 0$$

$$D = \{(\underline{x}_1, y_1), (\underline{x}_2, y_2), \ldots (\underline{x}_n, y_n)\}$$

$$x_i \in \mathbb{R}^d$$

$$d >> n$$

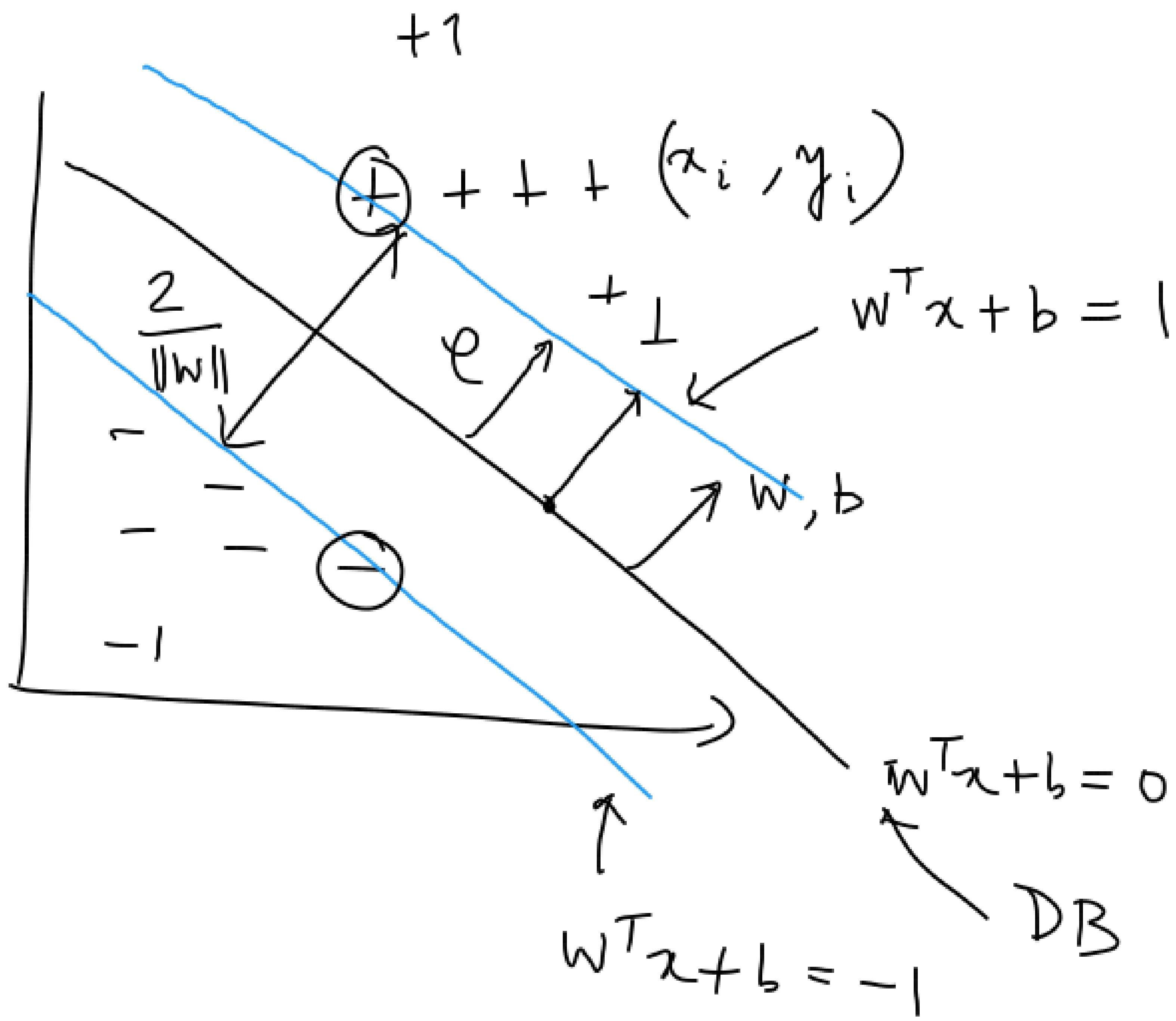
$$y_i \in \{+1, -1\}$$

Two variants of SVMs:

1. Hard margin SVMs: allows no misclassification.

2. Soft margin SVMs: misclassifications allowed but to a "limited" extent.

Formal description



$\frac{W}{m} \to W, \quad \frac{b}{m} \to b$

Pick a point x on DB : $W^T x + b = 0$

$$W^T\left(x + \rho \frac{W}{\|W\|}\right) + b = 1$$

$$\Rightarrow \underbrace{W^T x + b}_{=0} + \rho \|W\| = 1 \quad \Rightarrow \rho = \frac{1}{\|W\|}$$

Convex optimization

$$\max \frac{2}{\|W\|} \Rightarrow \min \frac{1}{2}\|W\|^2$$

$$\text{s.t.} \quad y_i(W^T x_i + b) \geq 1 \quad \forall i = 1, \dots, n.$$

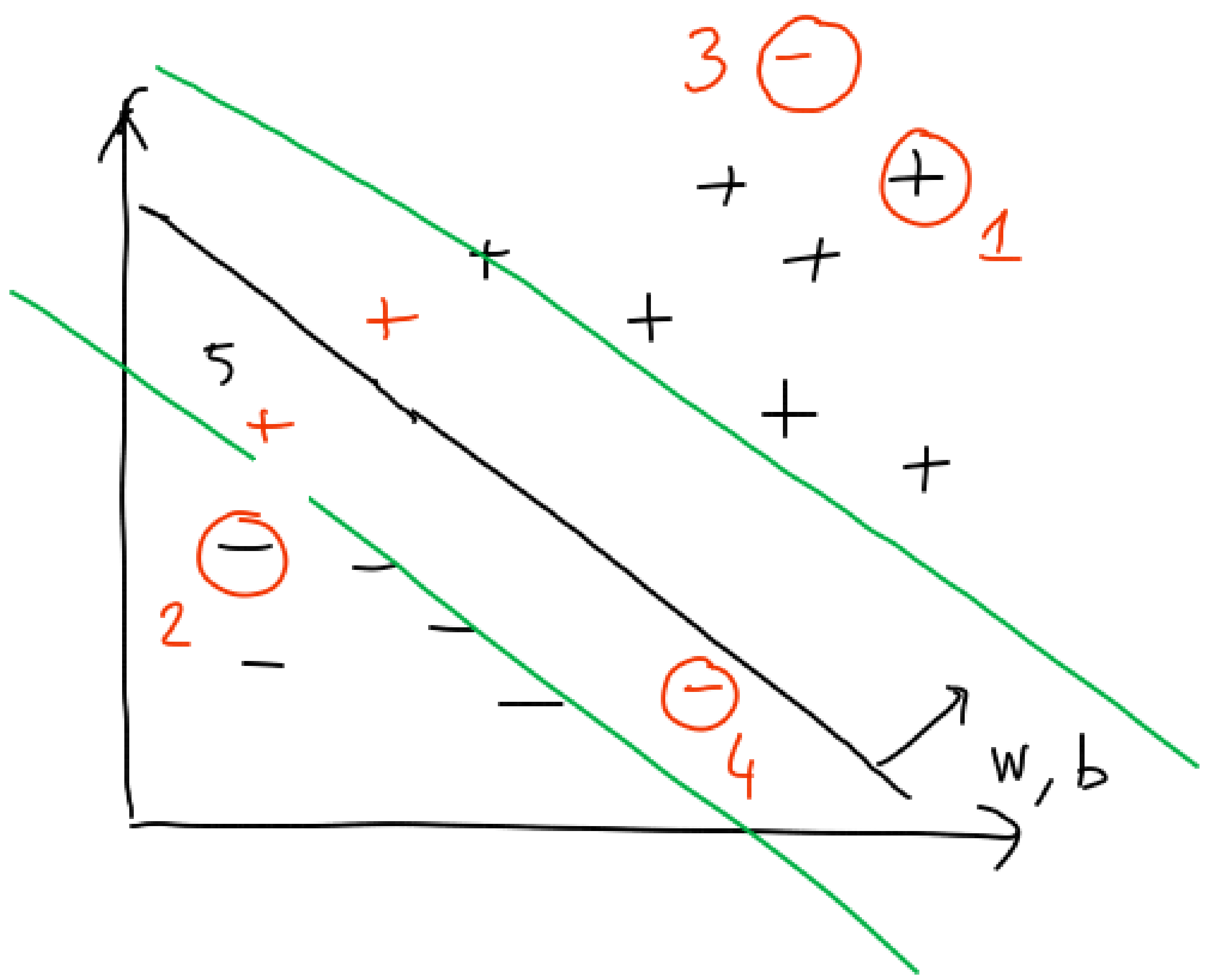
Constraints:

$$\left.\begin{array}{l} y_i = +1, \Rightarrow W^T x_i + b \geq 1 \\ y_j = -1, \Rightarrow W^T x_j + b \leq -1 \end{array}\right\}$$

$$y_i(W^T x_i + b) \geq 1 \quad \forall i = 1, \dots, n.$$

Remark: for $i \in$ support vectors

$$y_i(W^T x_i + b) = 1.$$



Soft margin SVMs

Data points are not linearly separable

Q: How to capture the degree of misclassification?

A: via a loss function.

Correct classification: $y_i(w^T x_i + b) \geq 1$

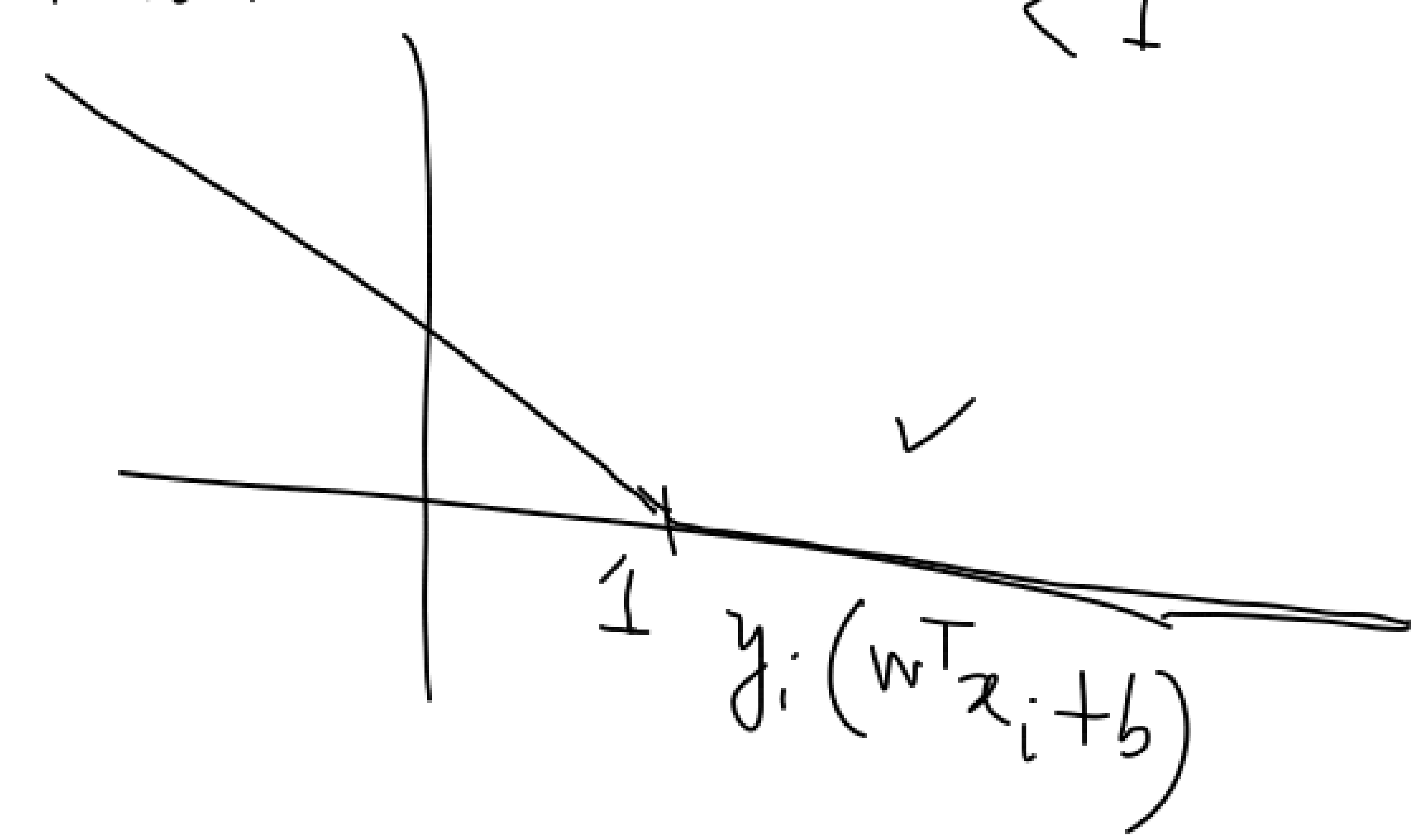
misclassification: < 1

$$\max\{0, \underbrace{1 - y_i(w^T x_i + b)}\} = L_i(w, b)$$

≤ 0 for correct

> 0 for misclassification

(hinge loss)



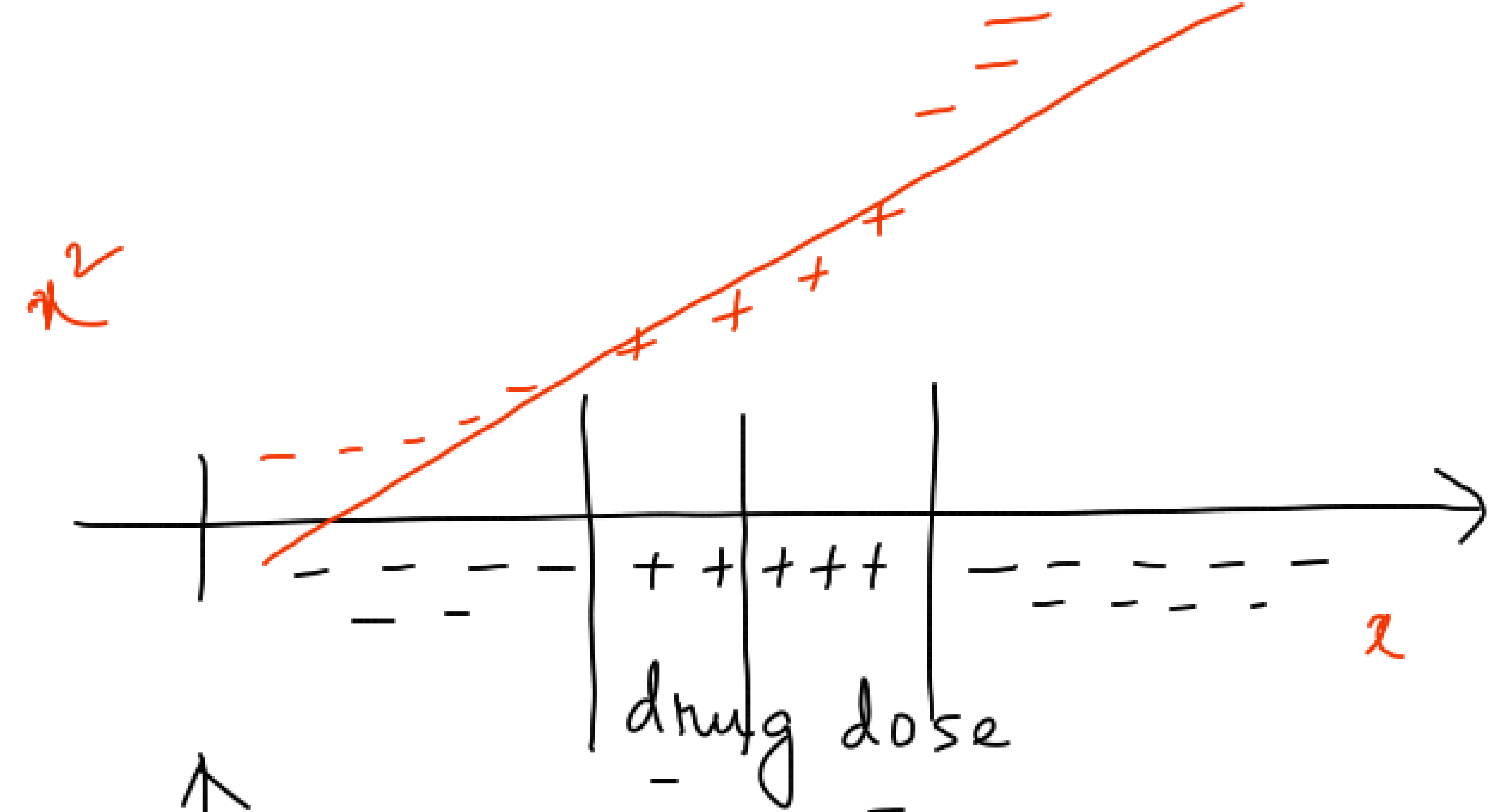
$$\min_{w,b} \frac{1}{n}\sum_{i=1}^{n} L_i(w,b) + \lambda \|w\|^2$$

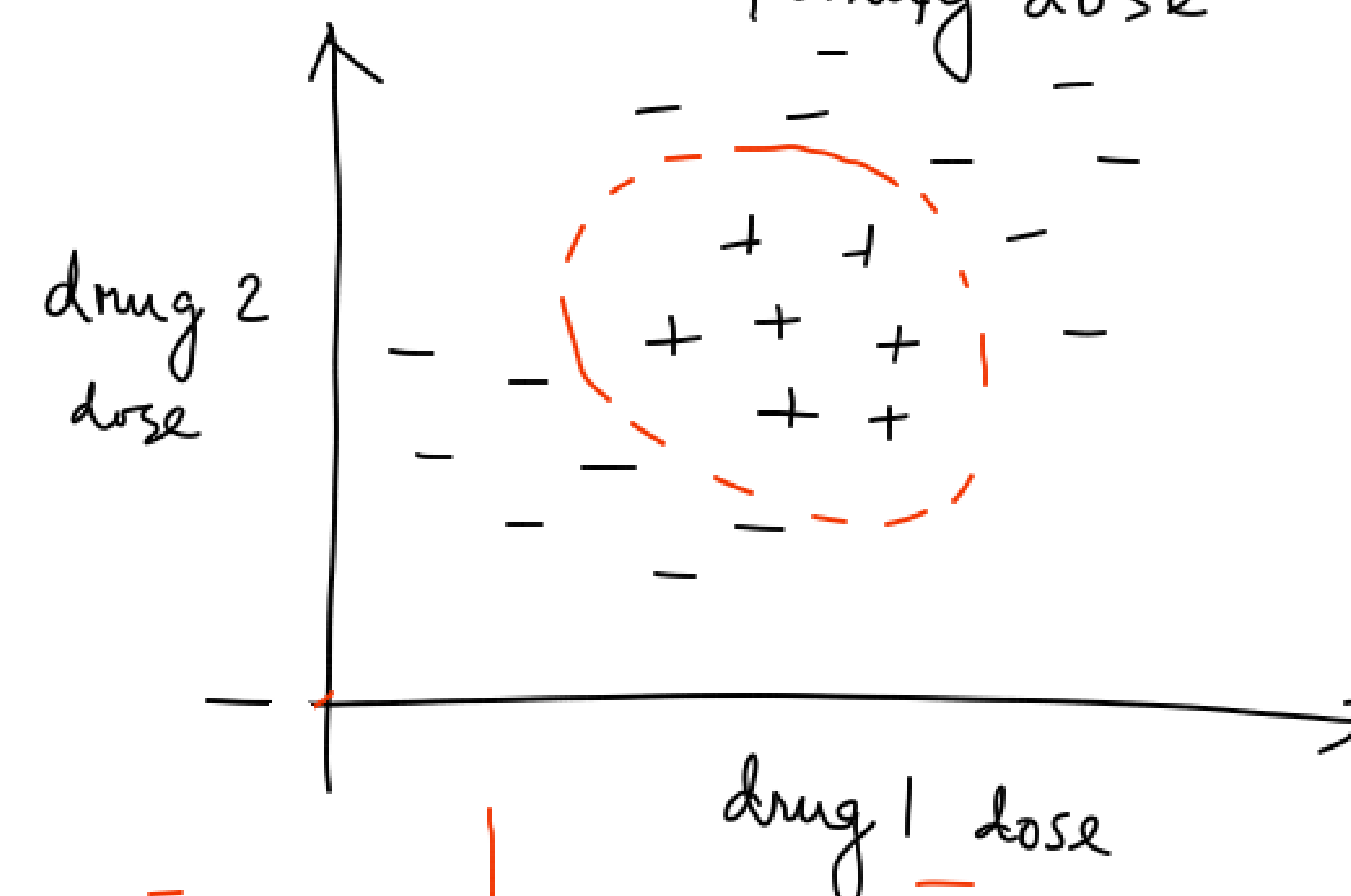
hyperparameter

Soft margin SVM.

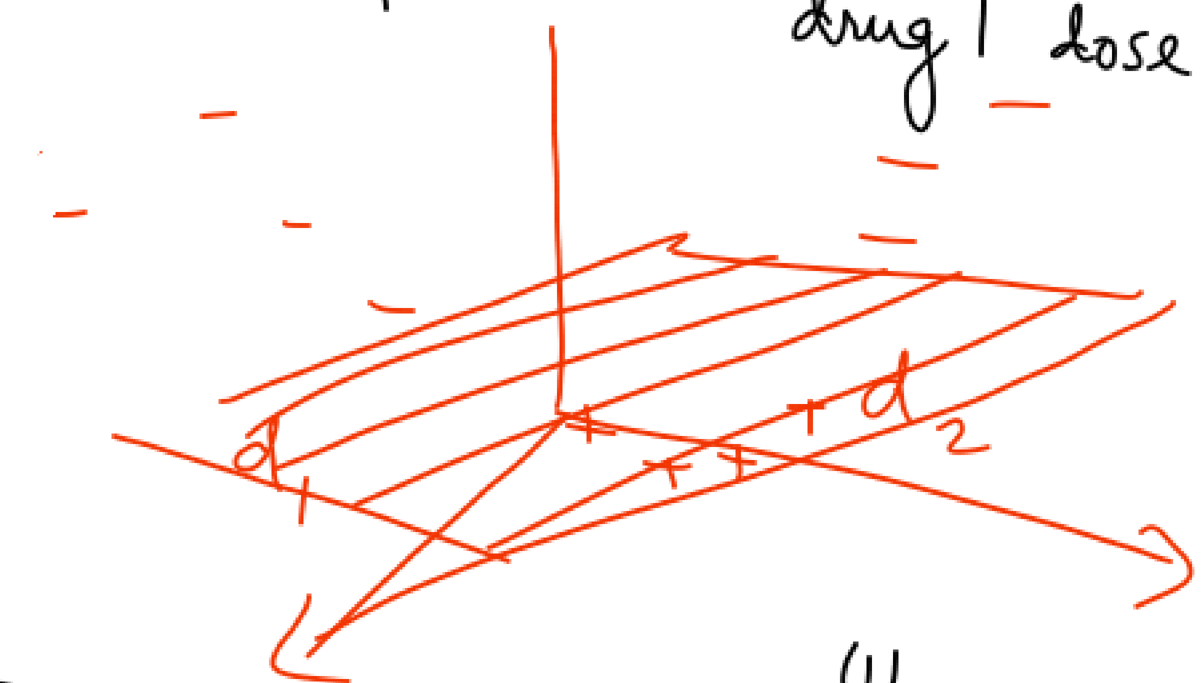
λ determines the tradeoff between misclassification and margin maximization.

Kernelization: a method to calculate the higher dimension trans. in a computationally efficient manner.





$x_i \longmapsto \phi(x_i)$



$x_i \to \mathbb{R}^d$

$d \gg n$

$$\begin{matrix} x_i^{(1)} \\ x_i^{(2)} \end{matrix} \to \begin{matrix} 1 \\ x_i^{(1)} \\ \vdots \\ x_i^{(1)} x_i^{(2)} \end{matrix}$$

Background of Kernelization

Hard margin SVM:

$$\min \frac{1}{2}\|W\|^2$$

$$\text{s.t.} \quad y_i(W^T x_i + b) \geq 1$$

$$\forall i = 1, \dots, n.$$

$$\mathcal{L}(W, b, \lambda)$$

$$= \frac{1}{2}\|W\|^2 - \sum \lambda_i \left(y_i(W^T x_i + b) - 1\right)$$

$$\lambda_i \geq 0$$

In general

$$\min f_0(x) \qquad \text{PRIMAL}$$

$$\text{s.t.} \quad f_i(x) \leq 0, \quad i = 1, \dots, m \leftarrow \lambda_i$$

$$h_i(x) = 0, \quad i = 1, \dots, p \leftarrow \gamma_i$$

domain D where the feasible x's live

$$D = \{x : f_i(x) \leq 0, \forall i, \; h_j(x) = 0, \forall j\}$$

Lagrangian

$$\mathcal{L}(x,\lambda,\gamma) = \underline{f_0(x)} + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{j=1}^{p} \gamma_j \underline{\underline{h_j(x)}}$$

$$\lambda_i \geq 0$$

Consider the problem

$$\underbrace{\max_{\lambda \geq 0, \gamma} \left[\min_{x \in D} \mathcal{L}(x,\lambda,\gamma) \right]}_{\text{EQUIVALENT to PRIMAL.}}$$

Lagrange dual

$$\underline{g(\lambda,\gamma)} = \min_{x \in D} \mathcal{L}(x,\lambda,\gamma)$$

$$\leq p^*$$

primal optimal

Dual problem

$$\boxed{\max_{\lambda \geq 0, \gamma} g(\lambda,\gamma)}$$

Convex opt:
dual opt $= p^*$

$$g(\lambda) = \min_{W,b} \mathcal{L}(W, b, \lambda)$$

$$\frac{\partial \mathcal{L}}{\partial W} = 0 \Rightarrow W = \sum_{i=1}^{n} \underline{\lambda_i}\, y_i\, x_i \quad \text{-- } ①$$

$$\frac{\partial \mathcal{L}}{\partial b} = 0 \Rightarrow \sum_{i=1}^{n} \lambda_i y_i = 0 \quad \text{-- } ②$$

use ① and ② to simplify

$$\max_{\lambda_i \geq 0\ \forall i} g(\lambda) = \sum \lambda_i - \frac{1}{2} \sum_i \sum_j \lambda_i \lambda_j y_i y_j \underline{x}_i^T \underline{x}_j$$

$\phi(x_i)^T \phi(x_j)$

$n \times n$

$d >> n$

$nd >> n^2$

Why dual?

① comp eff

$nd >> n^2$

② ~~Kernel friendly~~.

$$\boxed{\begin{array}{l} \text{s.t.} \\ \sum_i \lambda_i y_i = 0 \end{array}}$$

missed this constraint in the class