Foundations of Machine Learning (ECE 5984)

- Support Vector Machine -

Eunbyung Park

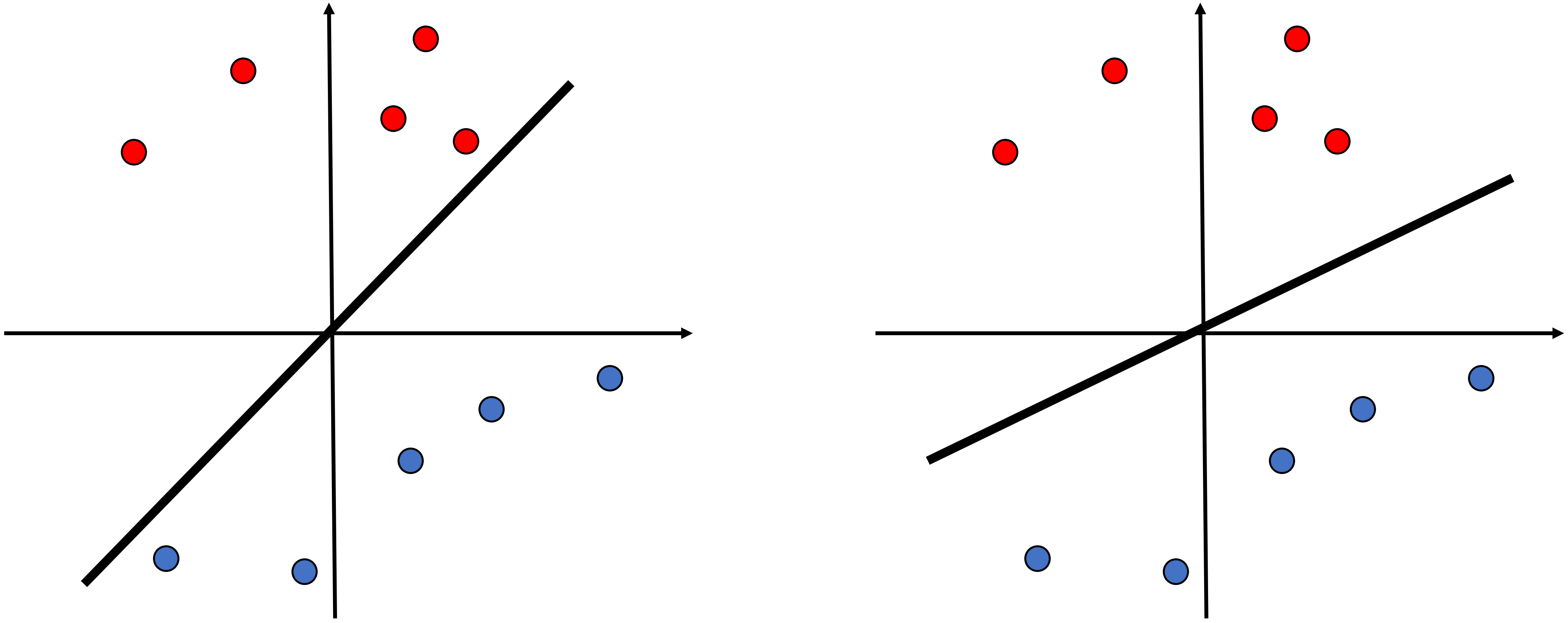
Assistant Professor

School of Electronic and Electrical Engineering

Eunbyung Park (silverbottlep.github.io)

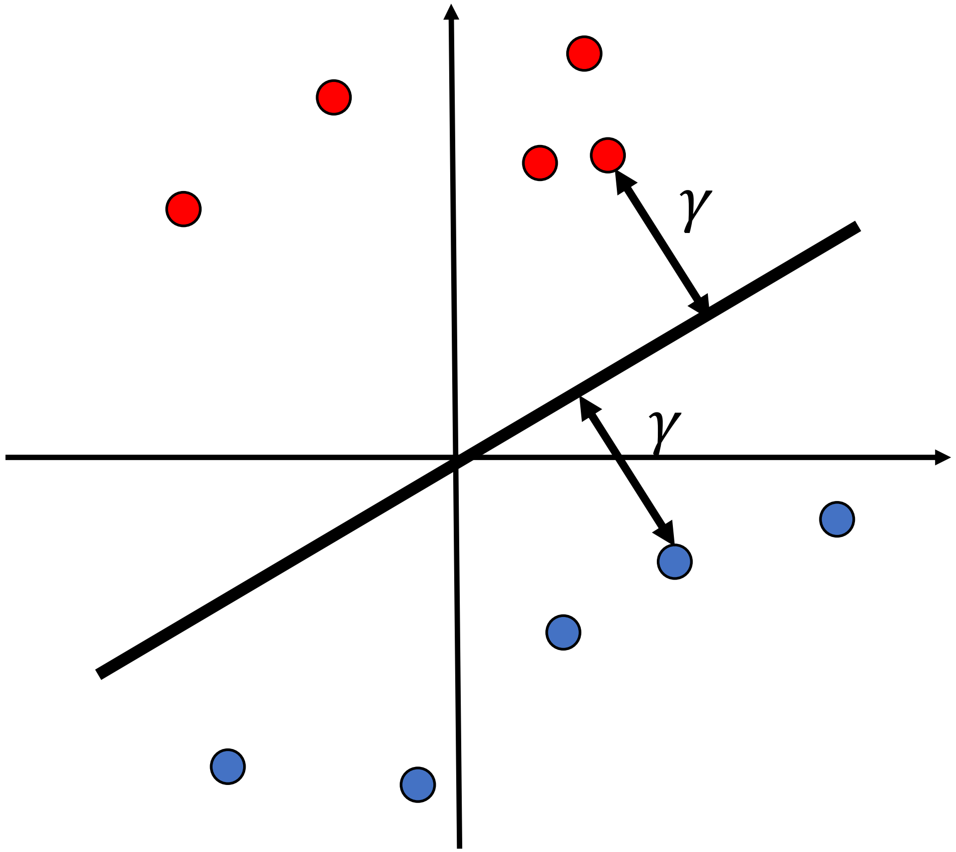
Maximum Margin Classifier

- What is the best separating hyperplane?



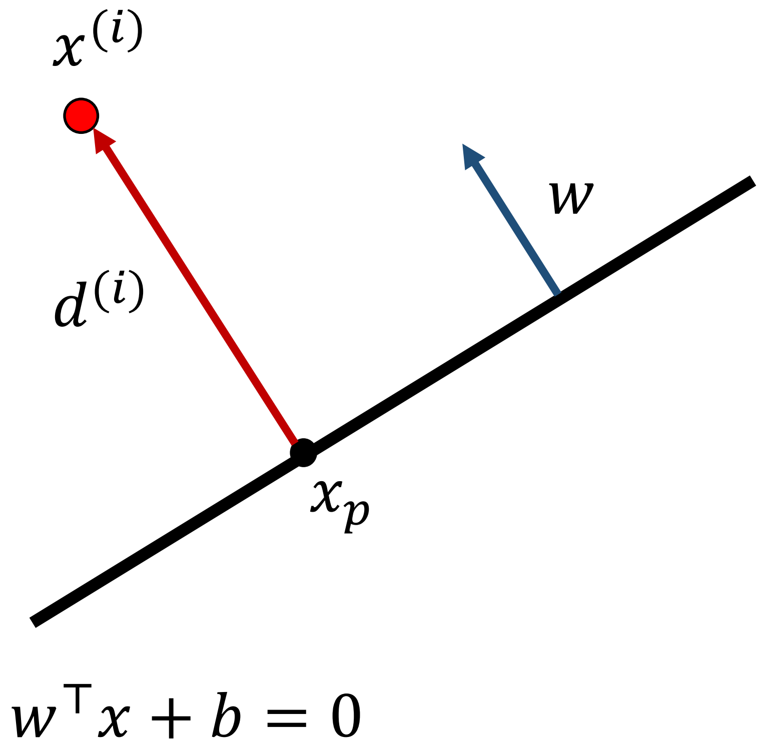
Maximum Margin Classifier

- Maximize the distance to the closest data points from both classes



Margin

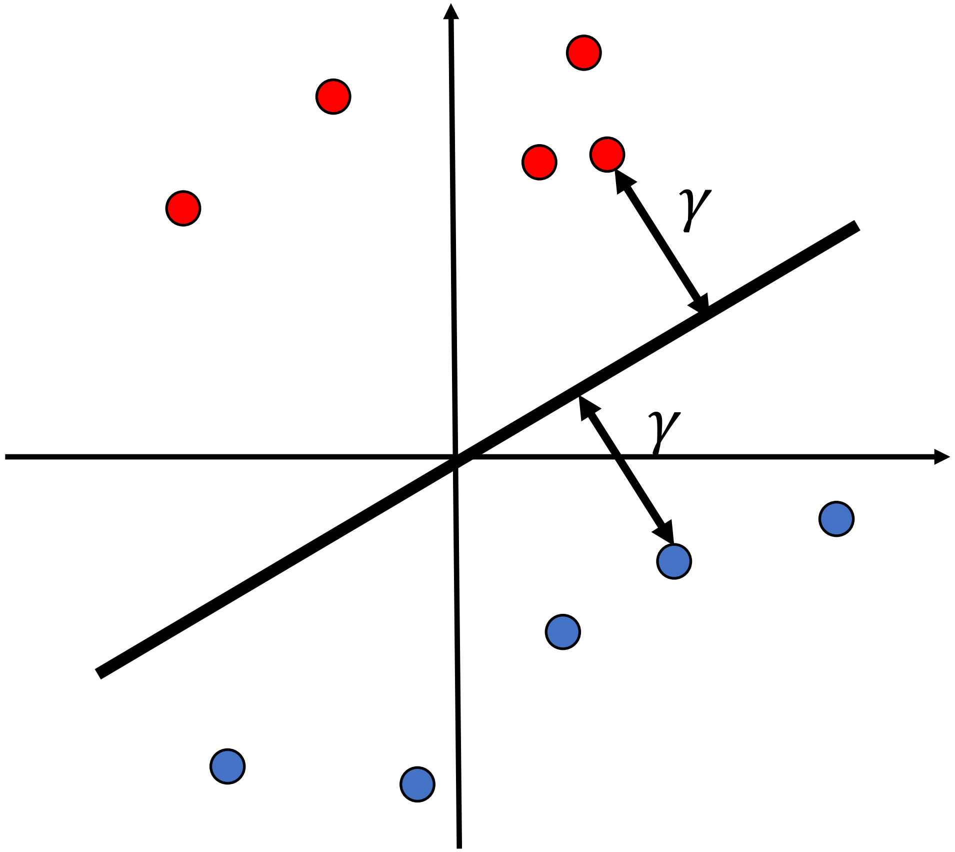
- Margin: the smallest distance across all points in dataset



$$\gamma(w, b) = \min_i ||d^{(i)}||_2$$

Maximum Margin Classifier

- Maximize the distance to the closest data points from both classes

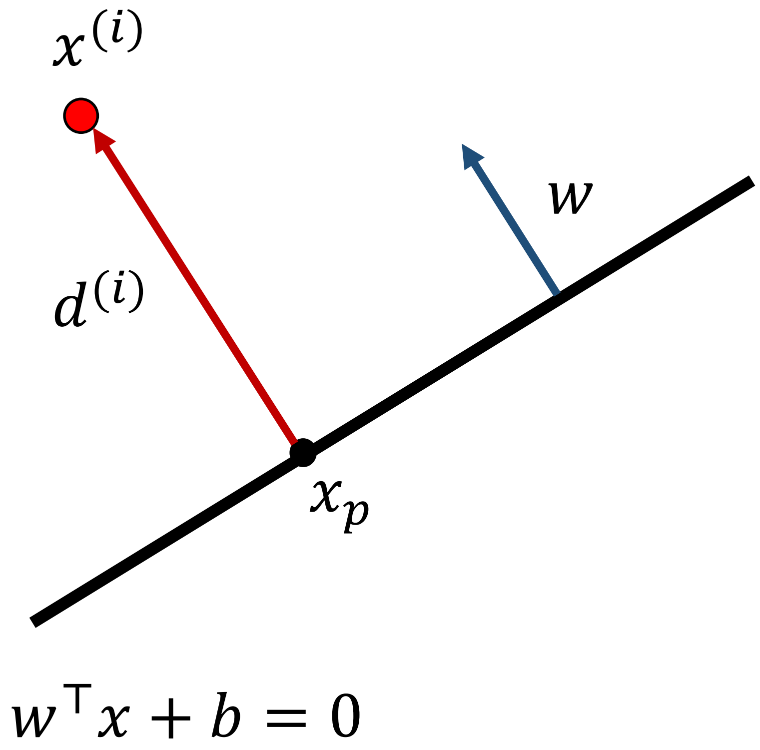


$y^{(i)} \in \{-1,1\}, x \in \mathbb{R}^d$

$$\max_{w,b} \gamma(w,b) = \max_{w,b} \min_i ||d^{(i)}||_2$$

$$\forall i, \qquad y^{(i)}\left(w^\top x^{(i)} + b\right) \geq 0$$

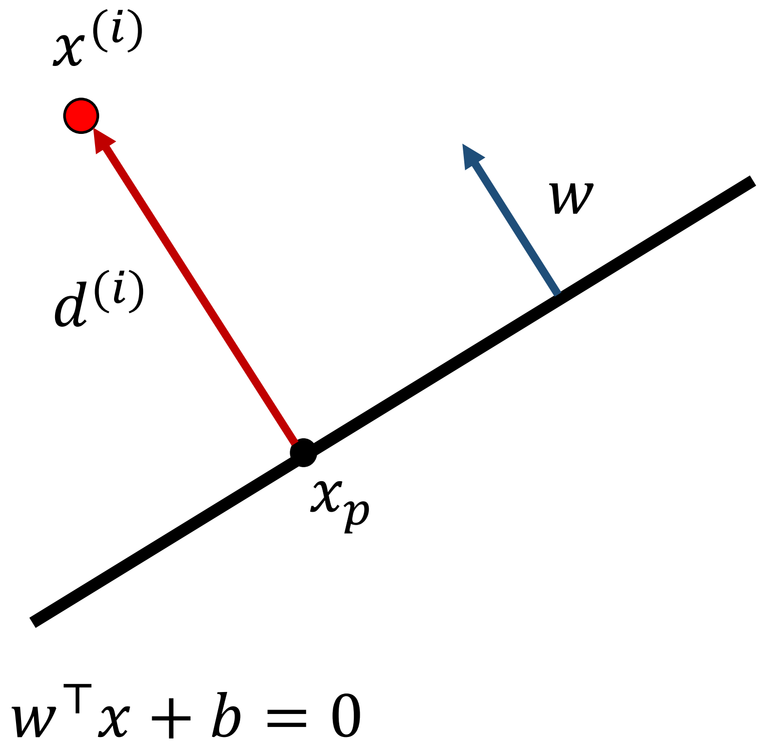
Margin



$$d^{(i)} = \alpha w$$

$$||d^{(i)}||_2 = \sqrt{{d^{(i)}}^\top d^{(i)}} = \sqrt{\alpha^2 w^\top w} = |\alpha|\sqrt{w^\top w}$$

Margin



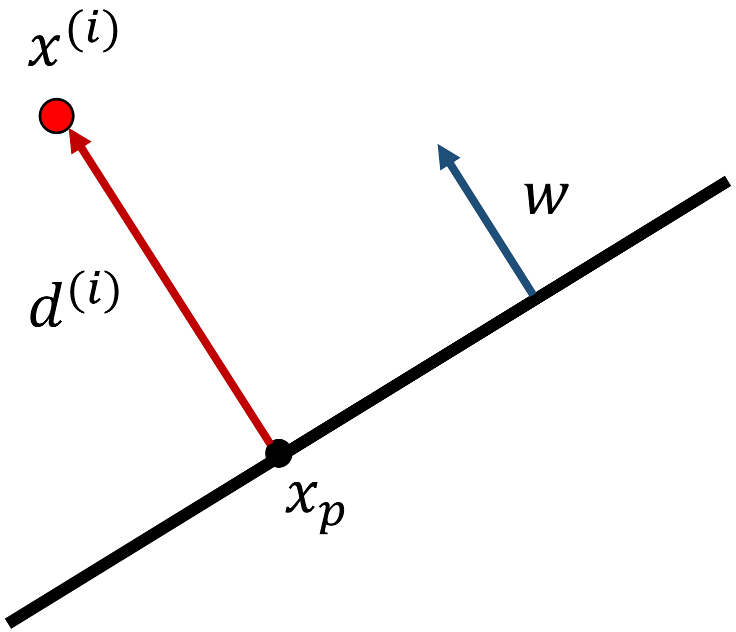
$$d^{(i)} = \alpha w$$

$$x_p = x^{(i)} - d^{(i)}$$

$$w^\top x_p + b = w^\top\left(x^{(i)} - \alpha w\right) + b = 0$$

$$\alpha = \frac{w^\top x^{(i)} + b}{w^\top w}$$

Margin



$w^\top x + b = 0$

$$d^{(i)} = \alpha w$$

$$x_p = x^{(i)} - d^{(i)}$$

$$w^\top x_p + b = w^\top \left(x^{(i)} - \alpha w\right) + b = 0$$

$$\alpha = \frac{w^\top x^{(i)} + b}{w^\top w}$$

$$||d^{(i)}||_2 = \sqrt{{d^{(i)}}^\top d^{(i)}} = \sqrt{\alpha^2 w^\top w} = |\alpha| \sqrt{w^\top w}$$

$$= \frac{|w^\top x^{(i)} + b|}{w^\top w} \sqrt{w^\top w} = \frac{|w^\top x^{(i)} + b|}{||w||_2}$$

Maximum Margin Classifier

$$\max_{w,b} \gamma(w,b) = \max_{w,b} \min_i \frac{|w^\top x^{(i)} + b|}{||w||_2} \qquad s.t. \quad \forall i, \quad y^{(i)}\left(w^\top x^{(i)} + b\right) \geq 0$$

$$= \max_{w,b} \frac{1}{||w||_2} \min_i |w^\top x^{(i)} + b| \qquad s.t. \quad \forall i, \quad y^{(i)}\left(w^\top x^{(i)} + b\right) \geq 0$$

Maximum Margin Classifier

$$\max_{w,b} \gamma(w,b) = \max_{w,b} \min_i \frac{|w^\top x^{(i)} + b|}{||w||_2} \qquad s.t. \quad \forall i, \quad y^{(i)}\left(w^\top x^{(i)} + b\right) \geq 0$$

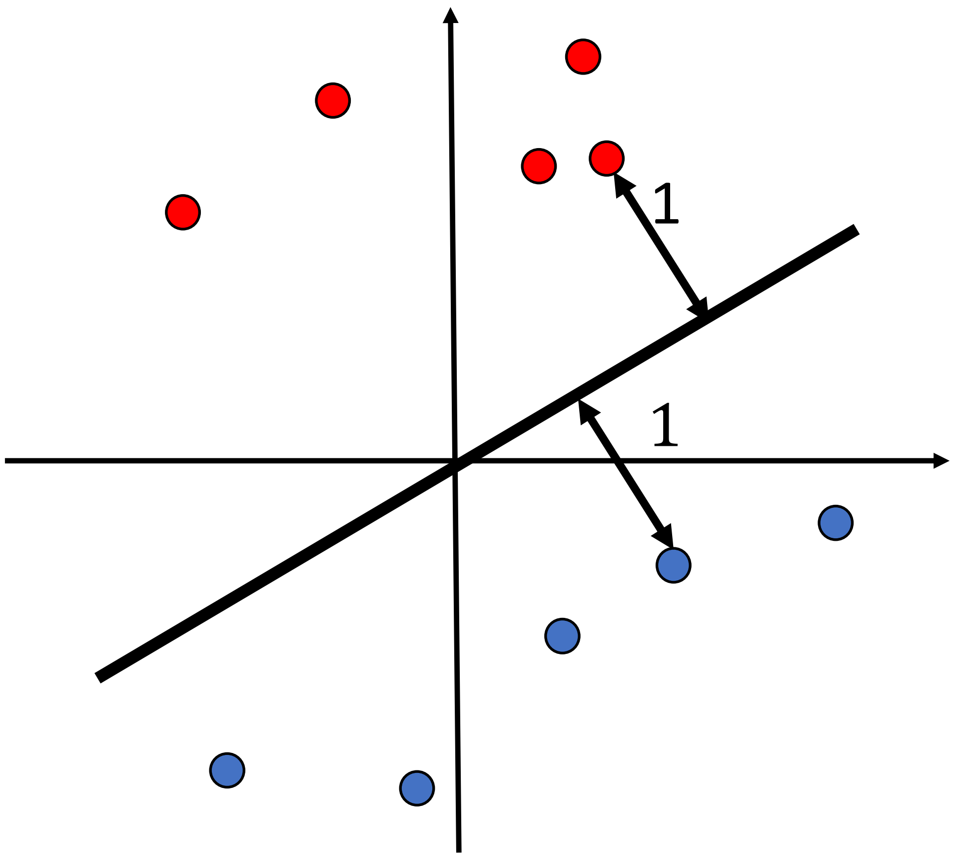
$$= \max_{w,b} \frac{1}{||w||_2} \min_i |w^\top x^{(i)} + b| \qquad s.t. \quad \forall i, \quad y^{(i)}\left(w^\top x^{(i)} + b\right) \geq 0$$

$$\rightarrow \max_{w,b} \frac{1}{||w||_2} \qquad s.t \quad \min_i \left|w^\top x^{(i)} + b\right| = 1$$

(not same w and q, but will be 'same' decision boundary)

Scale Invariant Property of Hyperplane

- Multiply something both side does not change hyperplane!
- We are going to find w,b that makes the margin 1!



$$w^\top x + b = 0$$

$$a \cdot (w^\top x + b) = a \cdot 0$$

Maximum Margin Classifier

$$\max_{w,b} \gamma(w,b) = \max_{w,b} \min_i \frac{|w^\top x^{(i)} + b|}{||w||_2} \qquad s.t. \quad \forall i, \quad y^{(i)}\left(w^\top x^{(i)} + b\right) \geq 0$$

$$= \max_{w,b} \frac{1}{||w||_2} \min_i |w^\top x^{(i)} + b| \qquad s.t. \quad \forall i, \quad y^{(i)}\left(w^\top x^{(i)} + b\right) \geq 0$$

$$\rightarrow \max_{w,b} \frac{1}{||w||_2} \qquad s.t \quad \min_i \left|w^\top x^{(i)} + b\right| = 1$$

(not same w and q, but will be ‘same’ decision boundary)

$$\rightarrow \min_{w,b} ||w||_2 \qquad s.t \quad \min_i \left|w^\top x^{(i)} + b\right| = 1$$

$$\rightarrow \min_{w,b} w^\top w \qquad s.t \quad \min_i \left|w^\top x^{(i)} + b\right| = 1$$

Maximum Margin Classifier

- The new objective

$$\min_{w,b} w^\top w \qquad \begin{array}{ll} \forall i, & y^{(i)}\left(w^\top x^{(i)} + b\right) \geq 0 \\ s.t & \min_i \left|w^\top x^{(i)} + b\right| = 1 \end{array}$$

Maximum Margin Classifier

- The new objective

(a) $$\min_{w,b} w^\top w \qquad \begin{array}{l} \forall i, \quad y^{(i)}\left(w^\top x^{(i)} + b\right) \geq 0 \\ s.t \quad \min_i \left|w^\top x^{(i)} + b\right| = 1 \end{array}$$

$$\Downarrow$$

(b) $$\min_{w,b} w^\top w \qquad \forall i, \quad y^{(i)}\left(w^\top x^{(i)} + b\right) \geq 1$$

(a)->(b): It's kind of obvious, $\left|w^\top x^{(i)} + b\right| \geq 1,\; y^{(i)} \in \{-1,1\}$

(b)->(a): ??

Maximum Margin Classifier

- The new objective

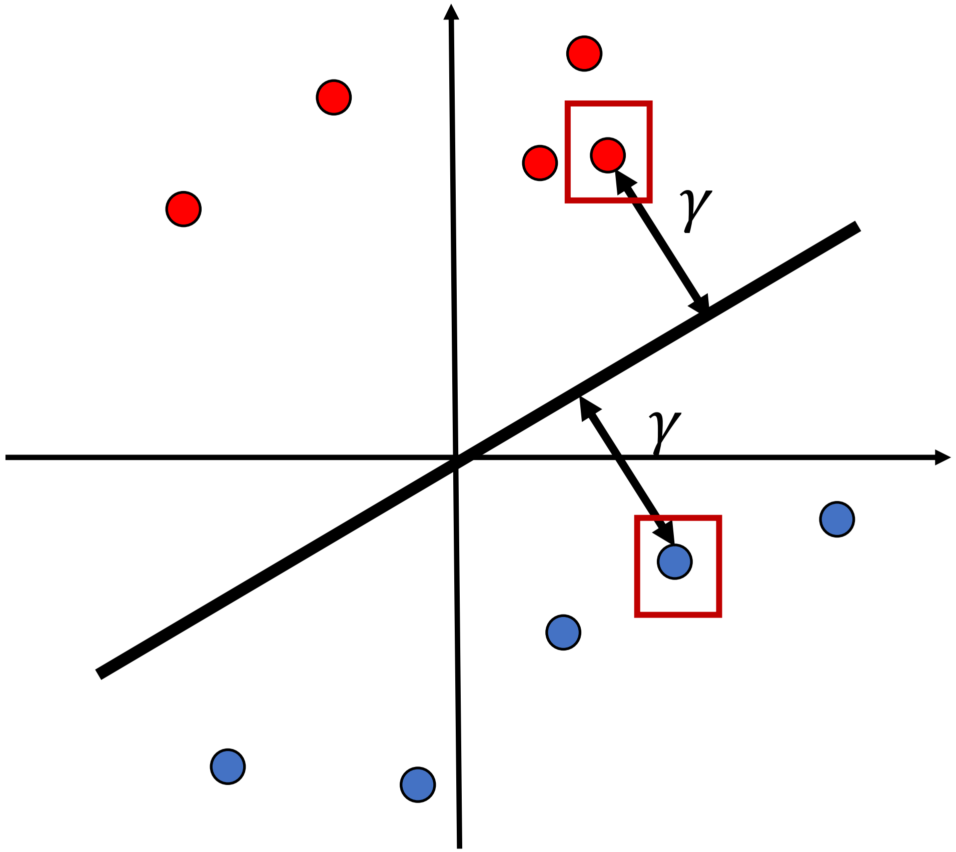
$$\min_{w,b} w^\top w \qquad \forall i, \qquad y^{(i)}\left(w^\top x^{(i)} + b\right) \geq 1$$

Quadratic objective

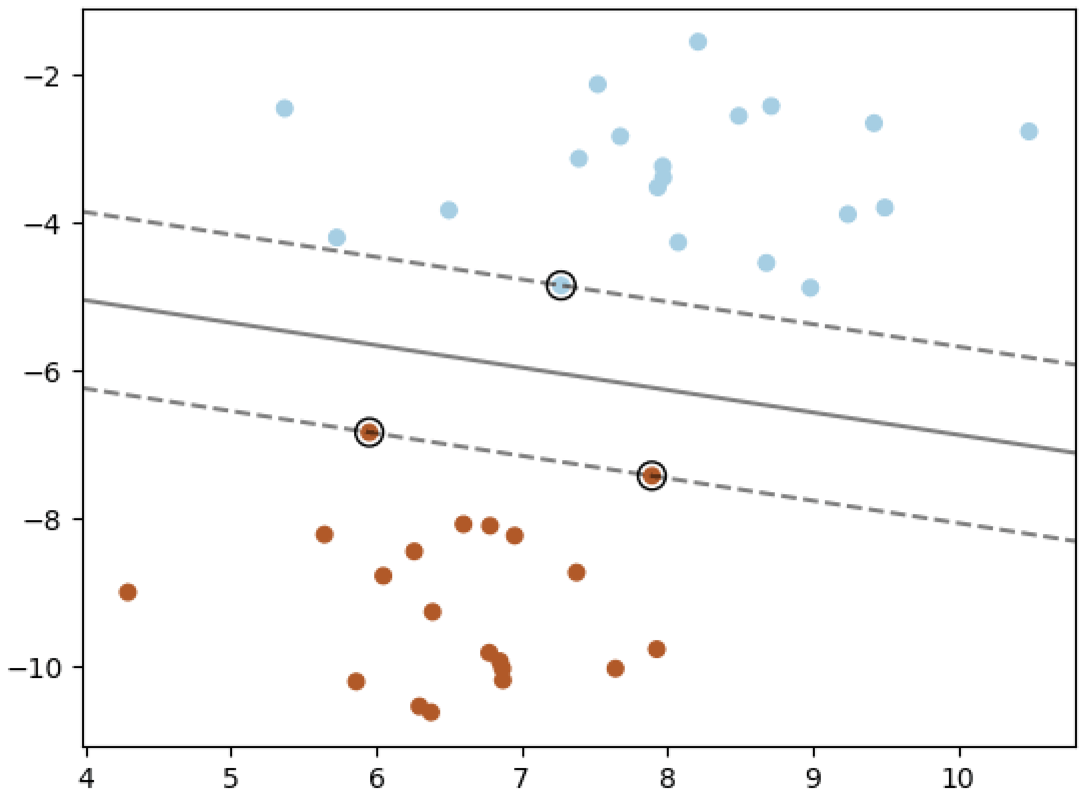
Linear constraints

1. Convex
2. We can use Quadratic Programing
   - very well established methods and softwares

Support Vectors



Decision Boundary



SVM with Soft Constraints

Non-Separable Cases

- The objective with slack variables

$$\min_{w,b} w^\top w + C \sum_{i=1}^{N} \xi^{(i)} \qquad \begin{aligned} &\forall i, \qquad y^{(i)}\left(w^\top x^{(i)} + b\right) \geq 1 - \xi^{(i)} \\ &\forall i, \qquad \xi^{(i)} \geq 0 \end{aligned}$$

Unconstrained Formulation

Margin violation

$$\xi^{(i)} = \begin{cases} 1 - y^{(i)}\left(w^\top x^{(i)} + b\right), & if\ y^{(i)}\left(w^\top x^{(i)} + b\right) < 1 \\ 0, & if\ y^{(i)}\left(w^\top x^{(i)} + b\right) \geq 1 \end{cases}$$

Out of margin

$$\Downarrow$$

$$\xi^{(i)} = \max\left(1 - y^{(i)}\left(w^\top x^{(i)} + b\right), 0\right)$$

Unconstrained Formulation

$$\min_{w,b} w^\top w + C \sum_{i=1}^{N} \max\left(1 - y^{(i)}\left(w^\top x^{(i)} + b\right), 0\right)$$

1. We can use gradient descent!
2. Is it convex?

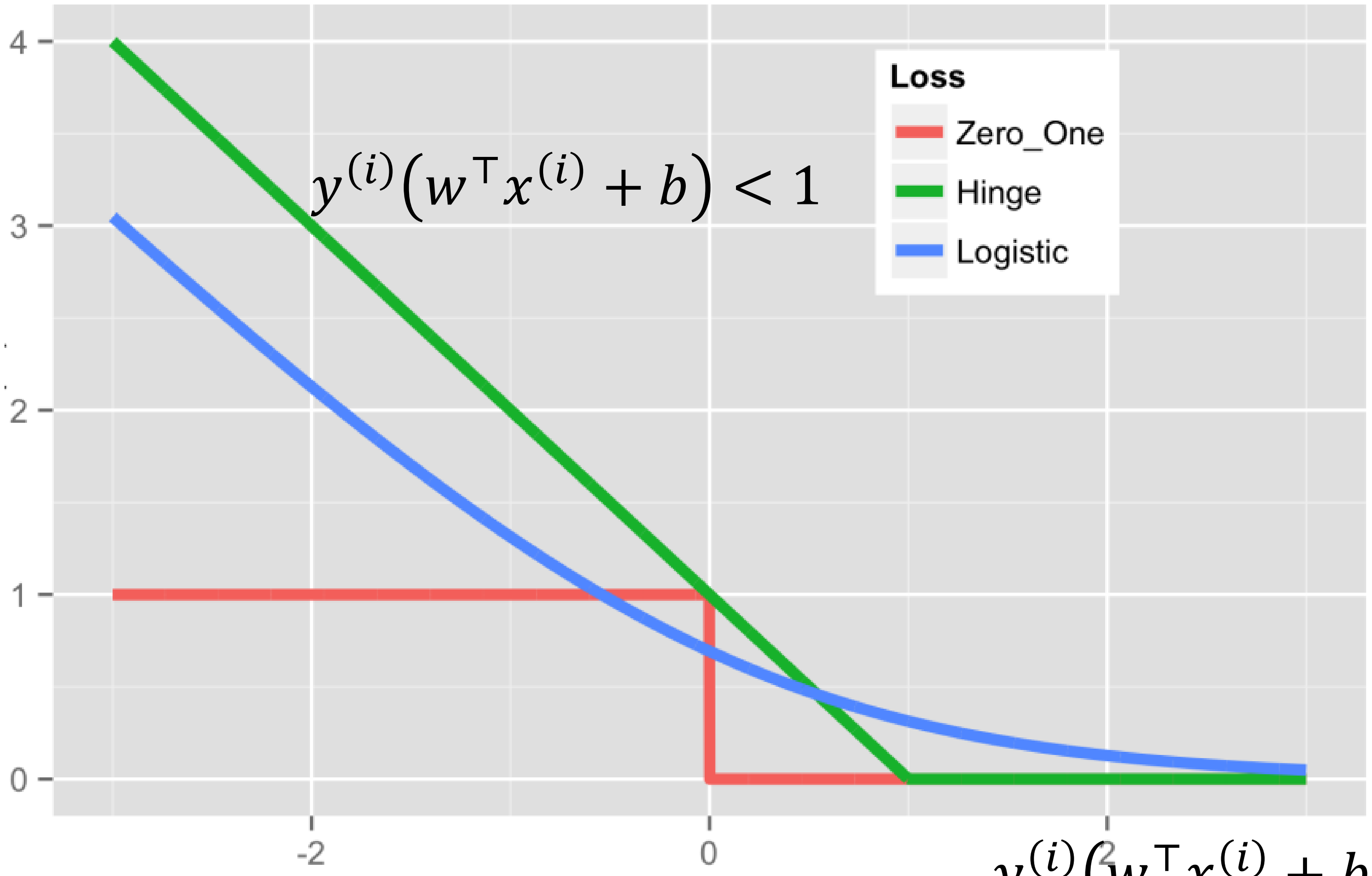
Regularization Perspective

$$\min_{w} C \sum_{i=1}^{N} \max\left(1 - y^{(i)}\left(w^\top x^{(i)} + b\right), 0\right) + w^\top w$$

Hinge Loss

L2 regularization

Hinge-Loss vs Log-Loss



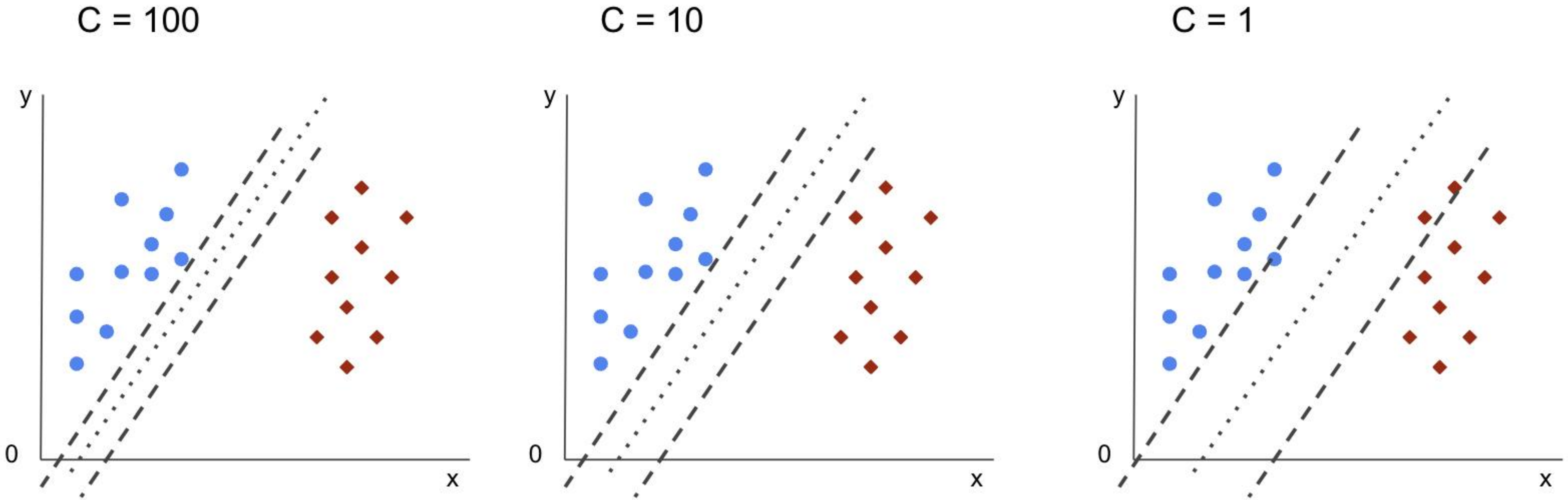
$$y^{(i)}\left(w^\top x^{(i)} + b\right) < 1$$

$$y^{(i)}\left(w^\top x^{(i)} + b\right) \geq 1$$

Hinge-Loss vs Log-Loss

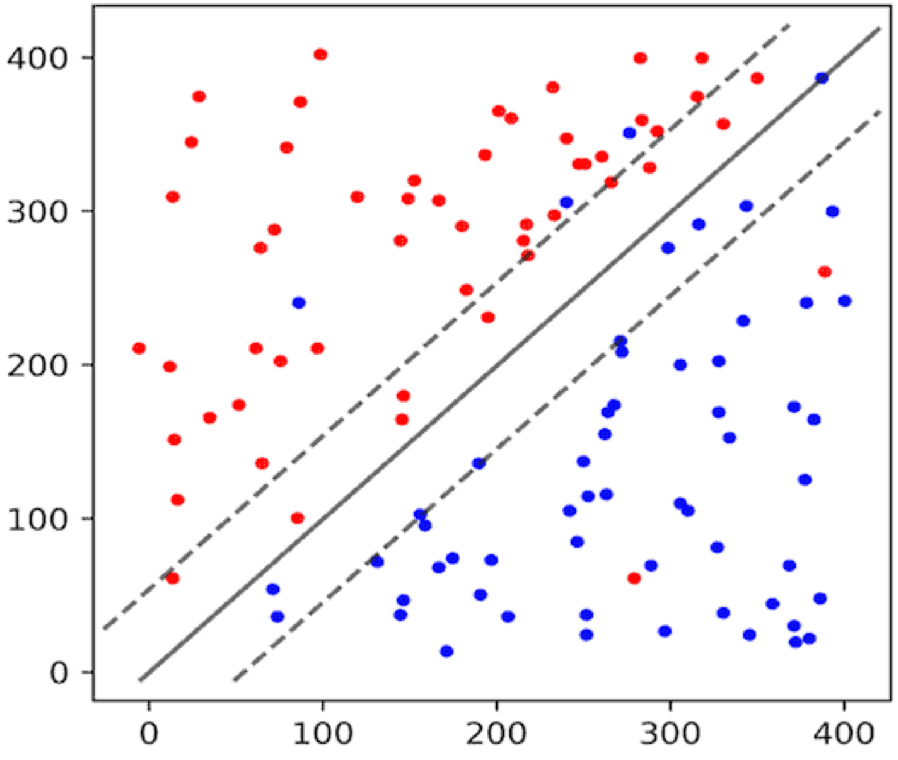
- When the distance from the boundary is greater than or equal to 1, the loss is 0
- If the distance from the boundary is less than 1, there is a loss. At 0 distance, the loss is 1
- Correctly classified points that are outside of the margin will not affect the decision boundary
- Hinge-Loss often incur sparsity

The C Hyperparameter

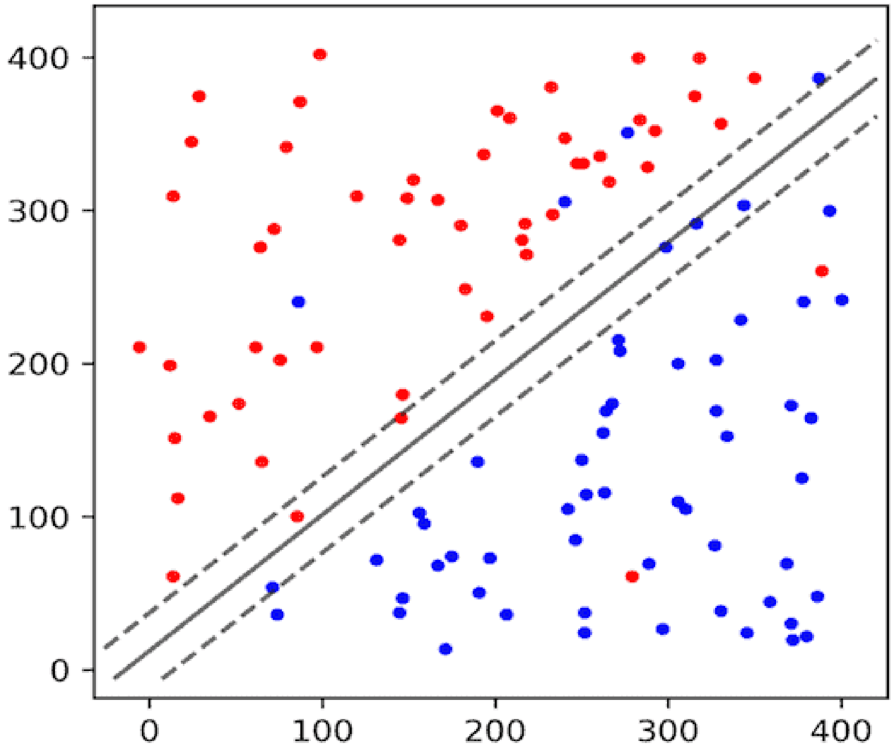


[Understanding SVM Hyperparameters (stackabuse.com)](https://stackabuse.com)

The C Hyperparameter



C = 1



C = 100

[Understanding SVM Hyperparameters (stackabuse.com)](https://stackabuse.com)

Optimization

Subgradient Descent

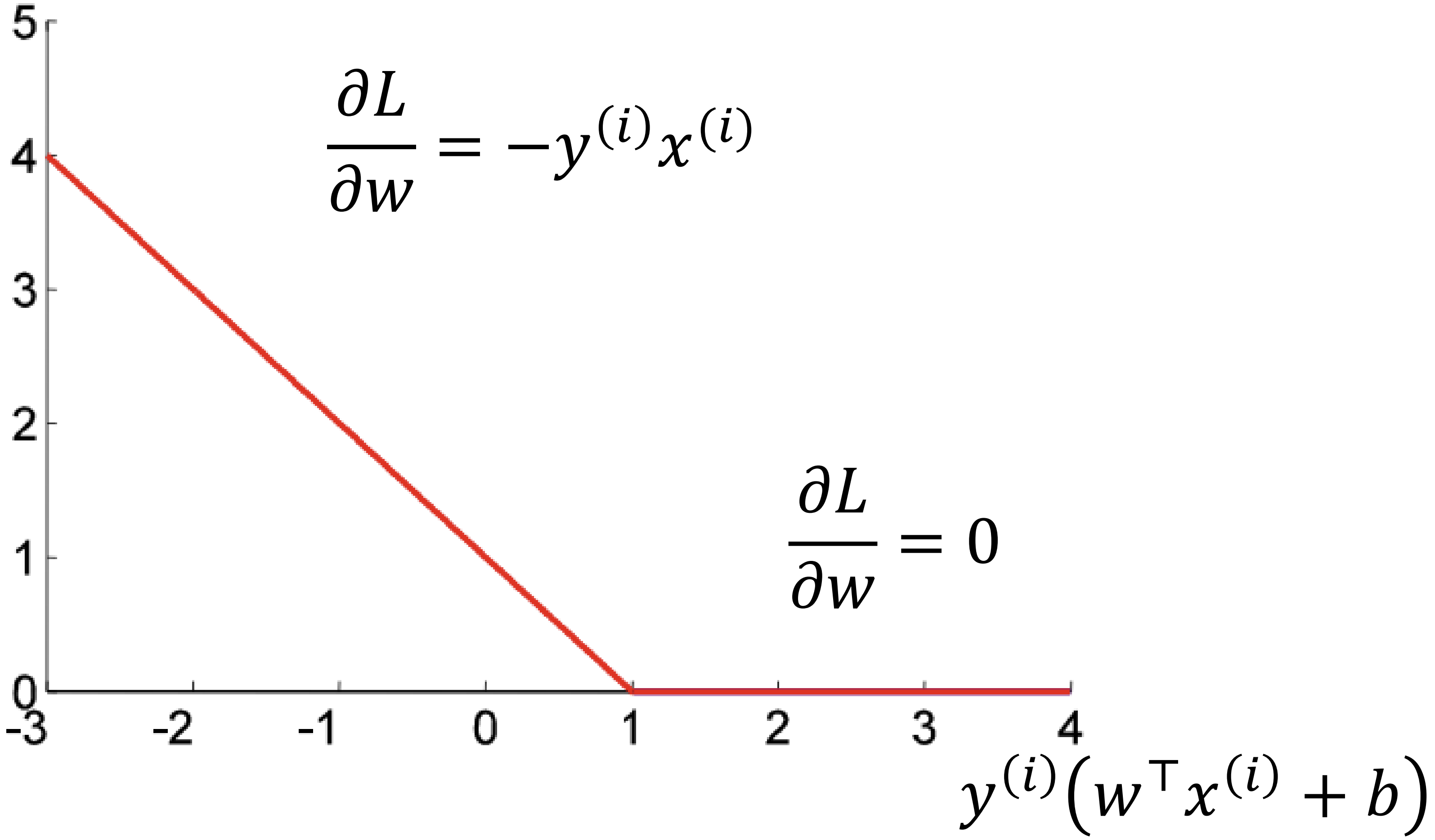
- Not differentiable everywhere

$$\min_{w,b} C \sum_{i=1}^{N} \max\left(1 - y^{(i)}\left(w^\top x^{(i)} + b\right), 0\right) + w^\top w$$

$$\min_{w,b} \sum_{i=1}^{N} \max\left(1 - y^{(i)}\left(w^\top x^{(i)} + b\right), 0\right) + \lambda w^\top w$$

Subgradient Descent

- Not differentiable everywhere



Subgradient Descent

- Stochastic gradient descent

$$w := w - \alpha\left(\lambda w - y^{(i)} x^{(i)}\right) \qquad if \quad y^{(i)}\left(w^\top x^{(i)} + b\right) < 1$$

$$w := w - \alpha\lambda w \qquad otherwise$$