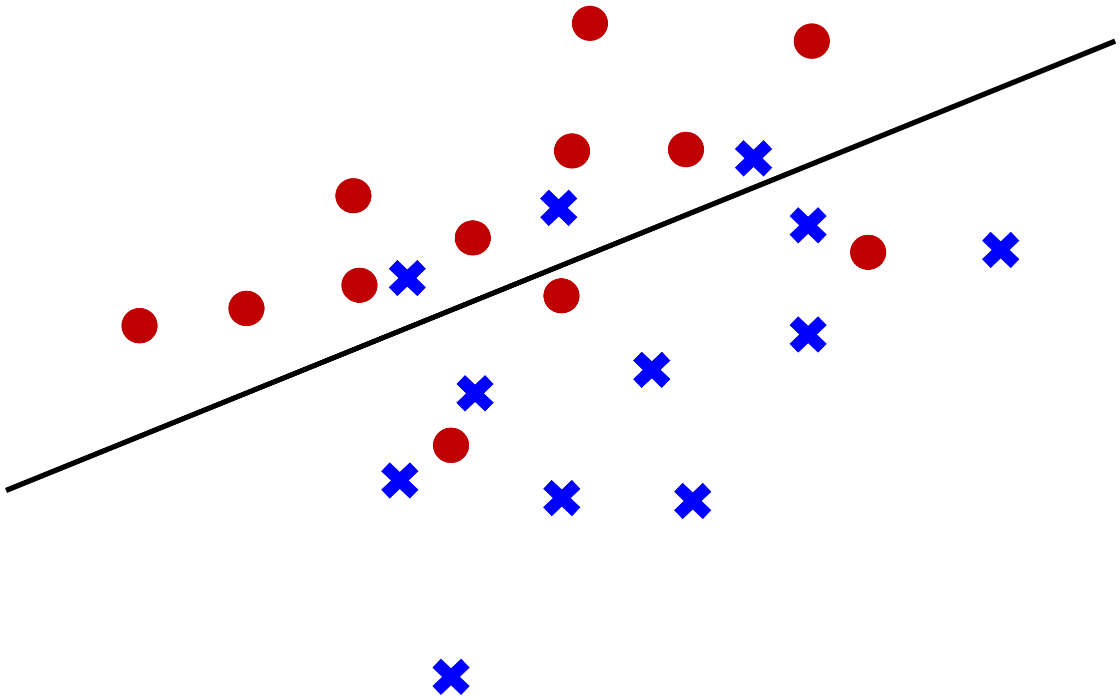
Support Vector Machines (SVM)

Data Intelligence and Learning (DIAL) Lab
Prof. Jongwuk Lee

Linear SVM with Soft Margin

Non-Linearly Separable Data

- It is **impossible** to find a linear boundary **without errors.**



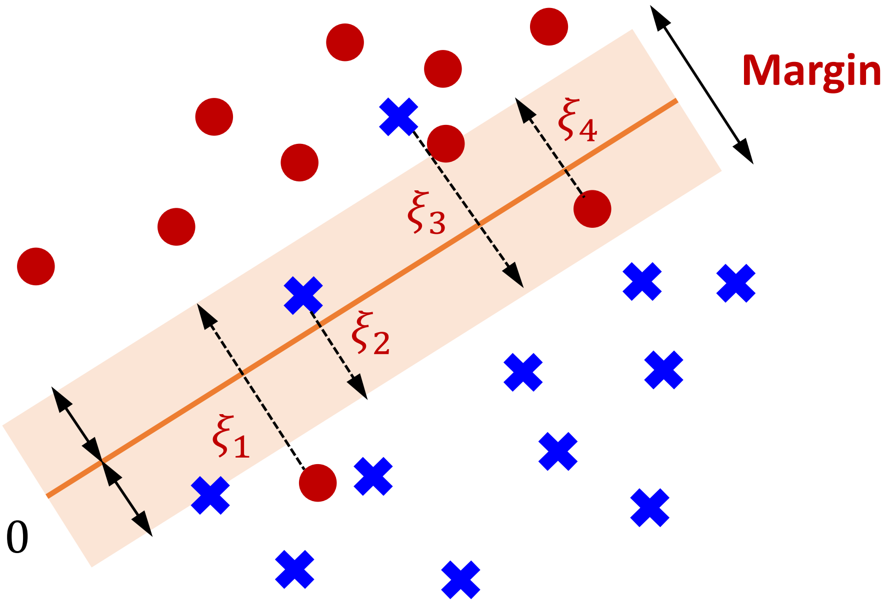
How to solve this problem?

How to Maximize the Margin?

- Allow some samples to be **in the margin** or **misclassified**.
  - **Correct**: $y^{(i)}\left(\mathbf{w}^T\mathbf{x}^{(i)} + b\right) \geq 1$
  - **Incorrect**: $0 < y^{(i)}\left(\mathbf{w}^T\mathbf{x}^{(i)} + b\right) < 1, y^{(i)}\left(\mathbf{w}^T\mathbf{x}^{(i)} + b\right) < 0$

- We introduce a slack variable ξ_i.

- Besides, penalize the sum of slack variables.



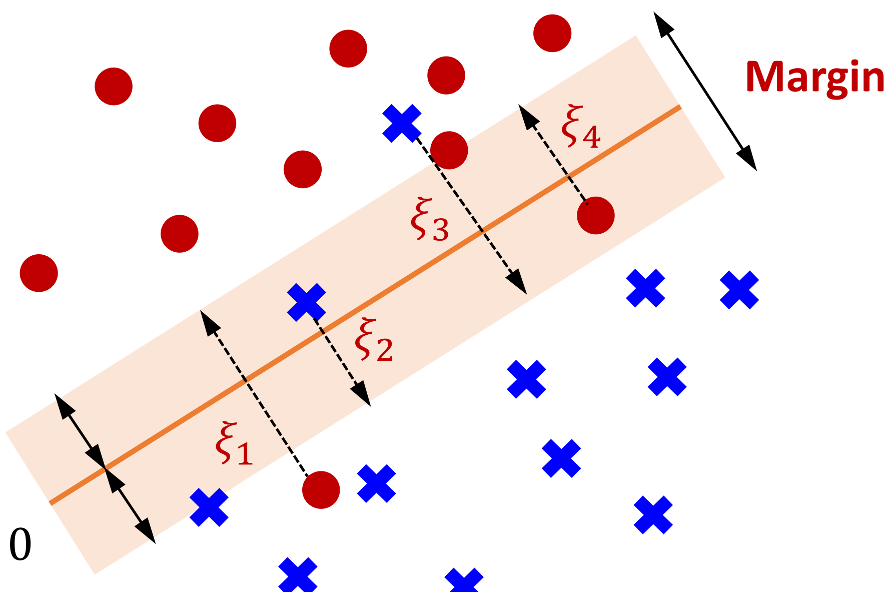
$$f(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\mathbf{x} + b = 0$$

Introducing Soft Margins

- Allow some samples to be **in the margin** or **misclassified**.
- We introduce a slack variable ξ_i.

$\xi_i \geq 0$

$$y^{(i)}\left(\mathbf{w}^T\mathbf{x}^{(i)} + b\right) \geq 1 \quad \Rightarrow \quad y^{(i)}\left(\mathbf{w}^T\mathbf{x}^{(i)} + b\right) \geq 1 - \xi_i$$



$f(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\mathbf{x} + b = 0$

Soft Margin SVM

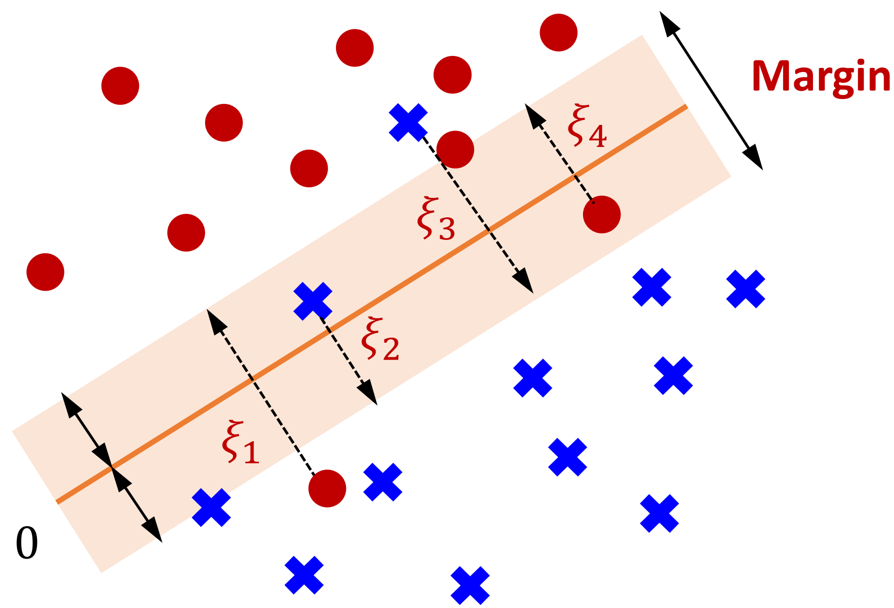
- Assume that an **error $\xi_i \geq 0$** for each sample $\mathbf{x}^{(i)}$.

$$y^{(i)}\left(\mathbf{w}^T\mathbf{x}^{(i)} + b\right) \geq 1 - \xi_i\,, \qquad \text{for } i = 1, 2, \cdots, n$$

Let ξ_i be a slack variable for $\mathbf{x}^{(i)}$.

- Penalize $\sum_i \xi_i$.

Finding a linear boundary that **maximizes the margin** and **minimizes the error**.



$$f(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\mathbf{x} + b = 0$$

Soft Margin SVM

- **Objective function**

Margin Error

$$\min_{\mathbf{w},b}\left(\frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}+C\sum_{i=1}^{n}\xi_i\right)$$

$$\text{subject to}\begin{cases} y^{(i)}(\mathbf{w}^{\mathrm{T}}\mathbf{x}^{(i)}+b)\geq 1-\xi_i \ \text{ for } i=1,\cdots,n \\ \xi_i\geq 0 \ \text{ for } \ i=1,\cdots,n \end{cases}$$

Slack variable

- **How to control C ?**

Effect of C

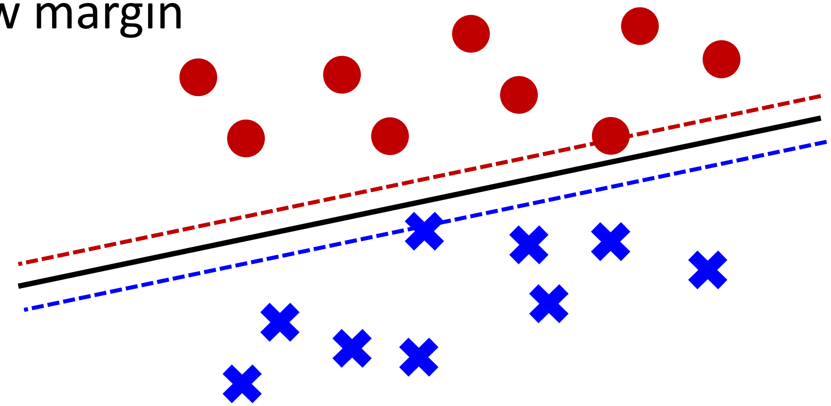
- **When C becomes ∞,**
  - **No allowance for errors** $\rightarrow$ Narrow margin
  - It is close to **hard margin SVM**.
  - Over-fitting

- **When $C = 0$,**
  - **Maximum allowance for errors** $\rightarrow$ Maximum margin
  - Over-generalization

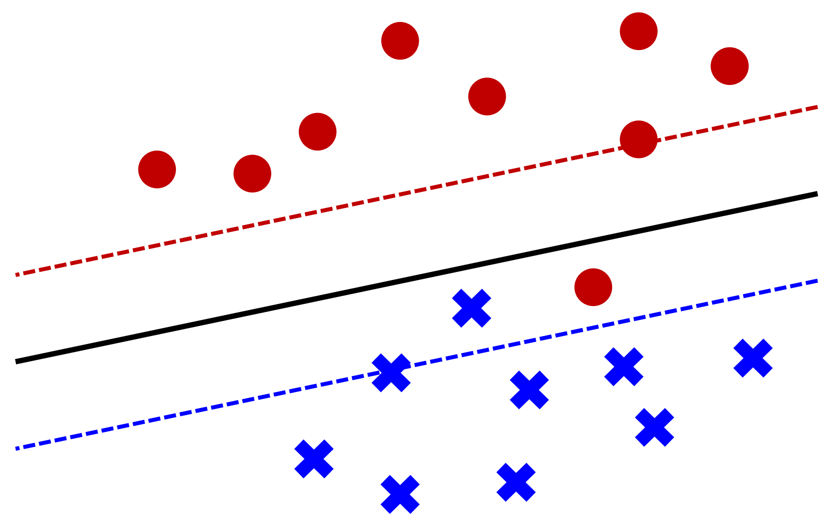
$$\min_{\mathbf{w},b}\left(\frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} + C\sum_{i=1}^{n}\xi_i\right)$$

Effect of C

- **When C becomes ∞,**
  - **No allowance for errors** $\rightarrow$ Narrow margin



- **When C is some value,**
  - Some allowance for errors



Understanding Sort Margin SVM

- **Simplifying the soft margin constraint by eliminating ξ_i**

$$
\begin{array}{ll}
y^{(i)}(\mathbf{w}^{\mathrm{T}}\mathbf{x}^{(i)} + b) \geq 1 - \xi_i & \text{for } i = 1, \cdots, n \\
\xi_i \geq 0 & \text{for } i = 1, \cdots, n
\end{array}
$$

$$
\Rightarrow \quad \boxed{\xi_i \geq 1 - y^{(i)}(\mathbf{w}^{\mathrm{T}}\mathbf{x}^{(i)} + b)}
$$

- **Case 1:** $1 - y^{(i)}\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}^{(i)} + b\right) \leq 0$
  - The smallest ξ_i that satisfies the constraint is $\xi_i = 0$.
- **Case 2:** $1 - y^{(i)}\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}^{(i)} + b\right) > 0$
  - The smallest ξ_i satisfies the constraint is $\xi_i = y^{(i)}\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}^{(i)} + b\right)$.

Understanding Sort Margin SVM

- **What is an optimal value as a function of w and b?**

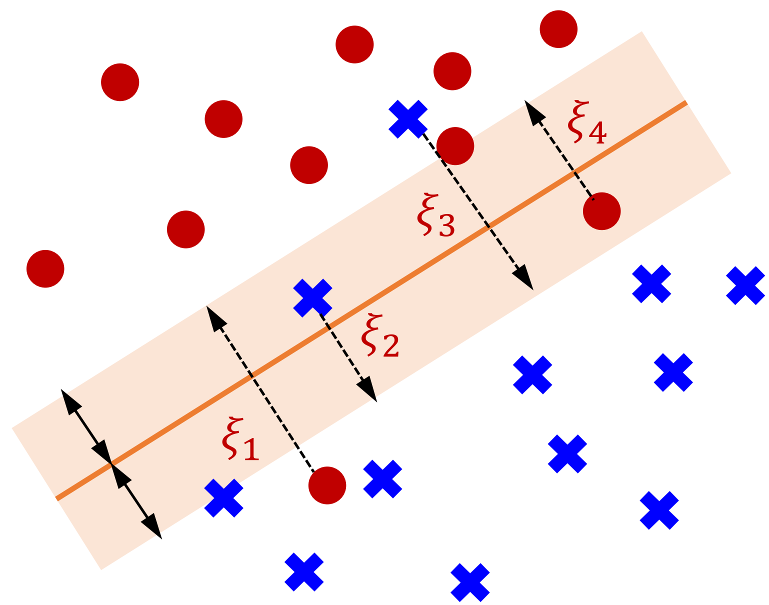
Case 1: If $y^{(i)}(\mathbf{w}^{\mathrm{T}}\mathbf{x}^{(i)} + b) \geq 1,$ then $\xi_i = 0$.

Case 2: If $y^{(i)}(\mathbf{w}^{\mathrm{T}}\mathbf{x}^{(i)} + b) < 1,$ then $\xi_i = 1 - y^{(i)}(\mathbf{w}^{\mathrm{T}}\mathbf{x}^{(i)} + b)$.

$$\Rightarrow \quad \boxed{\xi_i = \max\left(0, 1 - y^{(i)}(\mathbf{w}^{\mathrm{T}}\mathbf{x}^{(i)} + b)\right)}$$

The slack penalty

$$\boxed{\sum_{i=1}^{n} \xi_i = \sum_{i=1}^{n} \max\left(0, 1 - y^{(i)}(\mathbf{w}^{\mathrm{T}}\mathbf{x}^{(i)} + b)\right)}$$



Equivalent Hinge Loss Formulation

- **Substituting $\xi_i = \max\left(0, 1 - y^{(i)}\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}^{(i)} + b\right)\right)$ into the objective function, we can get**

$$\min_{\mathbf{w},b}\left(\frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} + C\sum_{i=1}^{n}\xi_i\right)$$

$$\text{subject to}\begin{cases} y^{(i)}(\mathbf{w}^{\mathrm{T}}\mathbf{x}^{(i)} + b) \geq 1 - \xi_i \ \text{ for } i = 1, \cdots, n \\ \xi_i \geq 0 \ \text{ for } \ i = 1, \cdots, n \end{cases}$$

$$\Downarrow$$

$$\min_{\mathbf{w},b}\left(\frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} + C\sum_{i=1}^{n}\max\left(0, 1 - y^{(i)}\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}^{(i)} + b\right)\right)\right)$$

The hinge loss is defined as $\mathcal{L}(y, \hat{y}) = \max\left(0,1 - y^{(i)}\hat{y}^{(i)}\right)$.

Equivalent to the Hinge Loss Function

- **Objective function of soft margin SVM**

$$\min_{\mathbf{w},b}\left(\frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}+C\sum_{i=1}^{n}\max\left(0,1-y^{(i)}\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}^{(i)}+b\right)\right)\right)$$

$$\Downarrow$$

$$\min_{\mathbf{w},b}\left(\sum_{i=1}^{n}\max\left(0,1-y^{(i)}\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}^{(i)}+b\right)\right)+\frac{1}{2C}\mathbf{w}^{\mathrm{T}}\mathbf{w}\right)$$

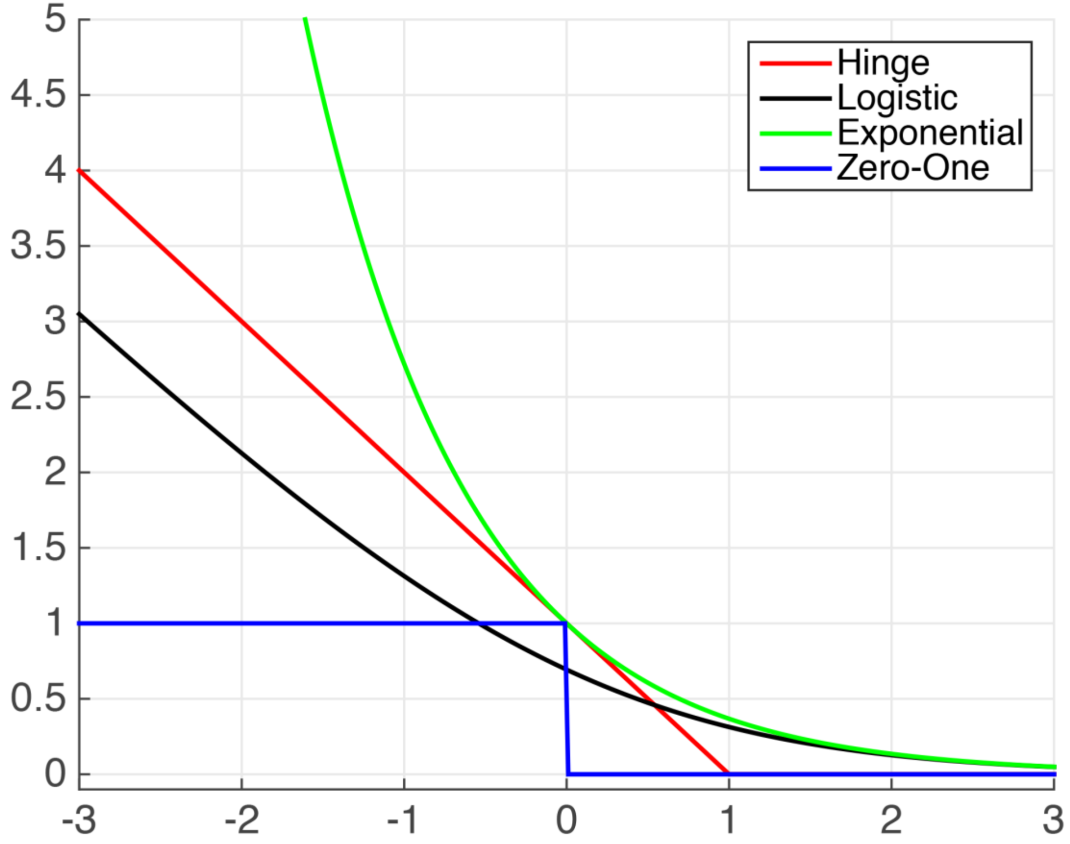
This first part is empirical risk minimization using a hinge loss.

This second term is the L2-regularization. It is used to prevent overfitting.

- **The soft margin SVM can be trained with a hinge loss function.**

Hinge Loss

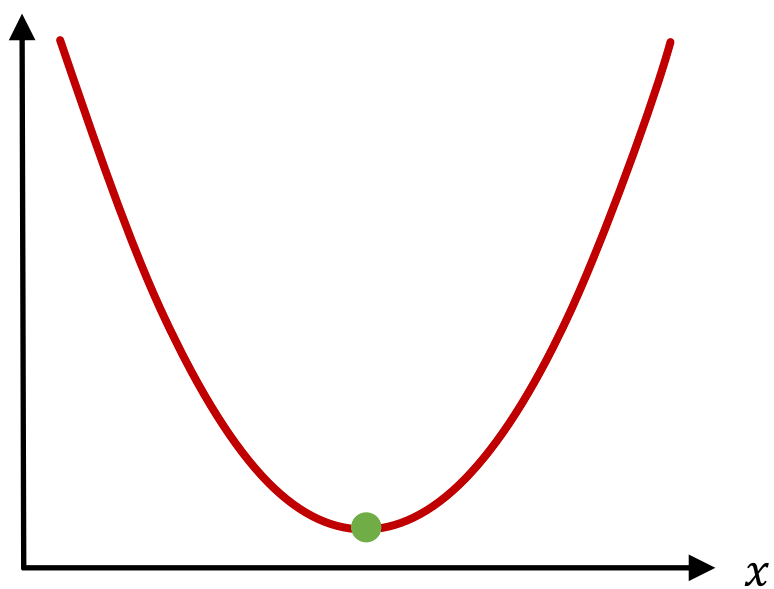
➢ **Hinge loss is upper bound of 0/1 loss!**

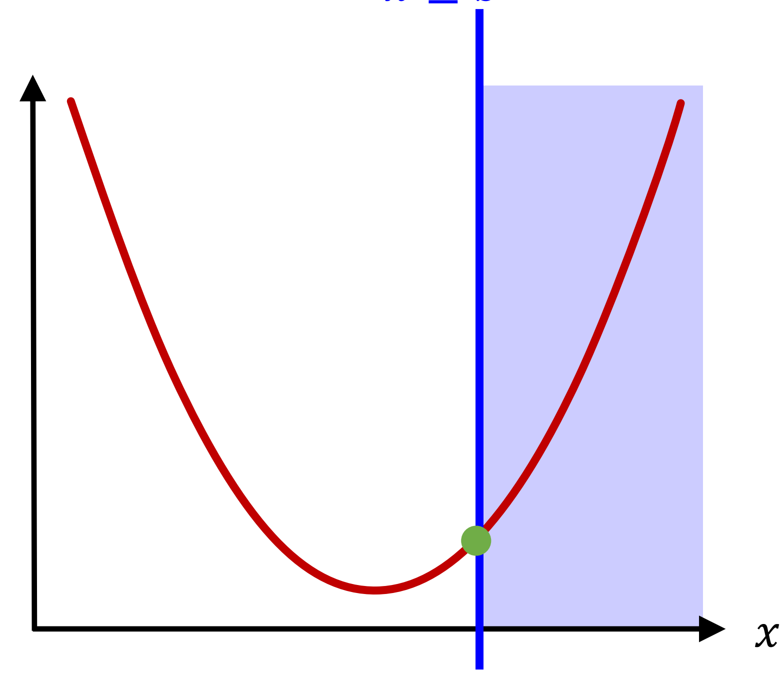


Dual Formulation of SVM

What is the Constrained Optimization?

$$\min_{x} x^2 \text{ such that } x \geq b$$

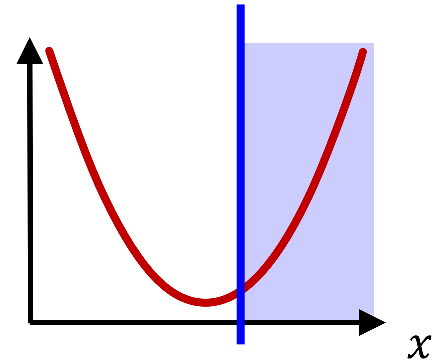




How to solve with constraints? → Lagrange Multiplier

Lagrange Multiplier: Dual Variables

$$\min_x x^2 \quad \text{subject to } x \geq b$$



Objective function:
Introduce a Lagrange multiplier.

$$L(x, \alpha) = x^2 - \alpha(x - b)$$

We will solve:

$$\min_x \max_\alpha L(x, \alpha) \quad \text{subject to } \alpha \geq 0$$

Add a new constraint.

➢ Why is it equivalent?

- $x < b \rightarrow (x - b) < 0 \rightarrow \max_\alpha -\alpha(x - b) = \infty$
  - Because min fights max, it does not happen.
- $x > b \rightarrow (x - b) > 0 \rightarrow \max_\alpha -\alpha(x - b) = 0, \;\; \alpha^* = 0$
  - Min is cooled with 0, and $\mathcal{L}(x, \alpha) = x^2$
- $x = b \rightarrow \alpha$ can be anything, and $\mathcal{L}(x, \alpha) = x^2$

$\therefore \;\; x^* = \max(b, 0)$

Dual Form of Hard-Margin SVM

- For simplicity, we mainly consider hard-margin SVM.

Original optimization problem

$$\min_{\mathbf{w},\mathbf{b}} \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} \text{ such that } \left(\mathbf{w}^{\mathbf{T}}\mathbf{x}^{(i)} + b\right)y^{(i)} \geq 1 \text{ for } i = 1, 2, \cdots, n$$

Rewrite constraints

One Lagrange multiplier per sample

Lagrangian form:

$$L(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} - \sum_{i=1}^{n} \alpha_i \left[\left(\mathbf{w}^{\mathbf{T}}\mathbf{x}^{(i)} + b\right)y^{(i)} - 1\right], \forall \alpha_i \geq 0$$

- Now, our goal is to solve $\min_{\mathbf{w},b} \max_{\boldsymbol{\alpha}} L(\mathbf{w}, b, \boldsymbol{\alpha})$ subject to $\forall \alpha_i \geq 0$

Dual Form of Hard-Margin SVM

- The dual form is more convenient to solve the objective function of SVM.

(Primal) $$\min_{\mathbf{w},b} \max_{\boldsymbol{\alpha}} L(\mathbf{w}, b, \boldsymbol{\alpha}) \text{ subject to } \alpha_i \geq 0 \text{ for } i = 1, \cdots, n$$

Swap min and max

(Dual) $$\max_{\boldsymbol{\alpha}} \min_{\mathbf{w},b} L(\mathbf{w}, b, \boldsymbol{\alpha}) \text{ subject to } \alpha_i \geq 0 \text{ for } i = 1, \cdots, n$$

First, compute the derivative of $\mathbf{w}$ and b, and it represents $L(\mathbf{w}, b, \boldsymbol{\alpha})$ as the function of $\boldsymbol{\alpha}$.

Dual SVM Derivation

- **Given the following Lagrangian function,**

$$L(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} - \sum_{i=1}^{n} \alpha_i \left[\left(\mathbf{w}^{\mathrm{T}} \mathbf{x}^{(i)} + b \right) y^{(i)} - 1 \right], \forall \alpha_i \geq 0$$

- **Compute the derivative of w and b and set them to zero.**

$$\frac{\partial L}{\partial \mathbf{w}} = \mathbf{w} - \sum_{i=1}^{n} \alpha_i \, \mathbf{x}^{(i)} y^{(i)} = 0 \quad \Rightarrow \quad \mathbf{w} = \sum_{i=1}^{n} \alpha_i \, \mathbf{x}^{(i)} y^{(i)}$$

$$\frac{\partial L}{\partial b} = -\sum_{i} \alpha_i y^{(i)} = 0 \quad \Rightarrow \quad \sum_{i} \alpha_i y^{(i)} = 0$$

Dual SVM Derivation

- **What is the meaning of $\alpha_i = 0$ and $\alpha_i > 0$?**
  - For $(\mathbf{x}^{(i)}, y^{(i)})$ corresponding to support vectors, $\alpha_i > 0$.
  - Otherwise, $\alpha_i = 0$.

$$L(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} - \sum_{i=1}^{n} \alpha_i\left[\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}^{(i)} + b\right)y^{(i)} - 1\right], \forall \alpha_i \geq 0$$

$$\frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y^{(i)} y^{(j)} \left(\mathbf{x}^{(i)}\right)^{\mathrm{T}}\left(\mathbf{x}^{(j)}\right)$$

$$\sum_{i=1}^{n} \alpha_i \left(\mathbf{w}^{\mathrm{T}}\mathbf{x}^{(i)}\right) y^{(i)} = \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y^{(i)} y^{(j)} \left(\mathbf{x}^{(i)}\right)^{\mathrm{T}}\left(\mathbf{x}^{(j)}\right)$$

Eliminating w and b using
$\mathbf{w} = \sum_{i=1}^{n} \alpha_i \mathbf{x}^{(i)} y^{(i)}$, $\sum_i \alpha_i y^{(i)} = 0$

$$L(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y^{(i)} y^{(j)} \left(\mathbf{x}^{(i)}\right)^{\mathrm{T}}\left(\mathbf{x}^{(j)}\right), \forall \alpha_i \geq 0$$

Dual SVM Derivation

➢ **Substituting these values, we can obtain the following form.**

$$\max_{\boldsymbol{\alpha}\geq 0}\min_{\mathbf{w},b}\frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}-\sum_{i=1}^{n}\alpha_i\left[\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}^{(i)}+b\right)y^{(i)}-1\right]$$

⬇

Scalars **Dot product**

$$\max_{\boldsymbol{\alpha}\geq\mathbf{0},\sum_i\alpha_i y^{(i)}=0}\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y^{(i)}y^{(j)}\left(\mathbf{x}^{(i)}\right)^{\mathrm{T}}\left(\mathbf{x}^{(j)}\right)$$

Sums over all the training samples.

Finding Parameters from α

- We determine w as follows.

$$\mathbf{w} = \sum_{i=1}^{n} \alpha_i \, \mathbf{x}^{(i)} y^{(i)}$$

- How do we determine b?
  - Given $\alpha_i\left[\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}^{(i)} + b\right)y^{(i)} - 1\right] = 0$,
    - **Support vectors**: $\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}^{(i)} + b\right)y^{(i)} - 1 = 0$ and $\alpha_i > 0$.
    - Otherwise, $\alpha_i = 0$.

$\alpha_i > 0$ implies the constraint is tight, i.e., $y^{(i)}\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}^{(i)} + b\right) = 1$.

$$b = y^{(i)} - \mathbf{w}^{\mathrm{T}}\mathbf{x}^{(i)} \quad \text{for any } \mathbf{x}^{(i)} \text{ such that } \alpha_i > 0$$

Solving Hard-Margin SVM

- **Given** $\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)}): 1 \le i \le n\}$**, where** $y^{(i)} \in \{-1, +1\}$**,**

$$\max_{\alpha_1, \cdots, \alpha_n} \left(\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y^{(i)} y^{(j)} \left(\mathbf{x}^{(i)} \cdot \mathbf{x}^{(j)} \right) \right)$$

$$\text{subject to } \begin{cases} \displaystyle\sum_{i=1}^{n} \alpha_i y^{(i)} = 0 \\ \alpha_i \ge 0 \ \text{ for } i = 1, \cdots, n \end{cases}$$

- **Solution**

$$\mathbf{w} = \sum_{i=1}^{n} \alpha_i y^{(i)} \mathbf{x}^{(i)}$$

$$b = y^{(i)} - \mathbf{w}^{\mathrm{T}} \mathbf{x}^{(i)} \quad \text{for any } \mathbf{x}^{(i)} \text{ such that } \alpha_i > 0$$

Prediction for Test Samples

- The solution of SVM is as follows.

$$\hat{y} = \text{sign}(\mathbf{w}^{\text{T}}\mathbf{x} + b)$$

$$\mathbf{w} = \sum_{i=1}^{n} \alpha_i y^{(i)} \mathbf{x}^{(i)}$$

$$b = y^{(i)} - \mathbf{w}^{\text{T}}\mathbf{x}^{(i)} \quad \text{for any } \mathbf{x}^{(i)} \text{ such that } \alpha_i > 0$$

- Given a new sample $\mathbf{x}_{new}$,

$$\hat{y} = \text{sign}(\mathbf{w}^{\text{T}}\mathbf{x}_{new} + b)$$

Example: Training Linear SVM

- Let $\mathcal{D} = \{(1, 1, -1), (2, 2, +1)\}$.

$$\max_{\alpha_1,\cdots,\alpha_n} \left(\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y^{(i)} y^{(j)} \left(\mathbf{x}^{(i)} \cdot \mathbf{x}^{(j)} \right) \right)$$

$$\text{subject to} \begin{cases} \displaystyle\sum_{i=1}^{n} \alpha_i y^{(i)} = 0 \\ \alpha_i \geq 0 \;\; \text{for } i = 1, \cdots, n \end{cases}$$

$$\Rightarrow \max_{\alpha_1,\alpha_2} \left((\alpha_1 + \alpha_2) - \frac{1}{2} \left(\begin{array}{c} \alpha_1 \alpha_1 y^{(1)} y^{(1)} \mathbf{x}^{(1)} \cdot \mathbf{x}^{(1)} + \alpha_1 \alpha_2 y^{(1)} y^{(2)} \mathbf{x}^{(1)} \cdot \mathbf{x}^{(2)} \\ + \alpha_2 \alpha_1 y^{(2)} y^{(1)} \mathbf{x}^{(2)} \cdot \mathbf{x}^{(1)} + \alpha_2 \alpha_2 y^{(2)} y^{(2)} \mathbf{x}^{(2)} \cdot \mathbf{x}^{(2)} \end{array} \right) \right)$$

$$\text{subject to} \begin{cases} \alpha_1 y^{(1)} + \alpha_2 y^{(2)} = 0 \\ \alpha_i \geq 0 \;\; \text{for } i = 1,2 \end{cases}$$

Example: Training Linear SVM

- Let $\mathcal{D} = \{(1, 1, -1), (2, 2, +1)\}$.

$$\max_{\alpha_1,\alpha_2} \Big((\alpha_1 + \alpha_2) - (\alpha_1^2 - 4\alpha_1\alpha_2 + 4\alpha_2^2)\Big) \text{ s.t. } \begin{cases} -\alpha_1 + \alpha_2 = 0 \\ \alpha_i \geq 0 \ \text{ for } i = 1,2 \end{cases}$$

$\Downarrow$ Since $\alpha_1 = \alpha_2$

$$\max_{\alpha_1} (\alpha_1^2 - 2\alpha_1) \text{ s.t. } \alpha_i \geq 0 \ \text{ for } i = 1,2$$

$\Rightarrow$ $\alpha_1 = \alpha_2 = 1$

- Using the solution, we can determine w and b.

Example: Training Linear SVM

- **Let** $\mathcal{D} = \{(1, 1, -1), (2, 2, +1)\}$ **and** $\alpha_1 = \alpha_2 = 1$

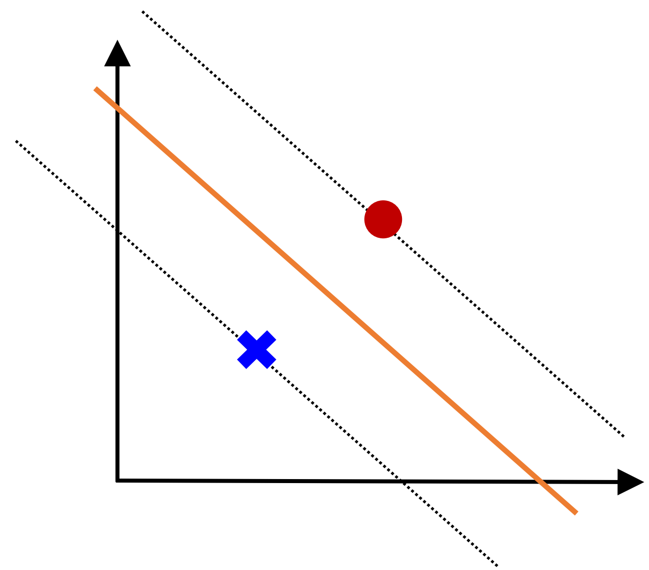
- **Solution**

$$\mathbf{w} = \sum_{i=1}^{n} \alpha_i y^{(i)} \mathbf{x}^{(i)}$$

$$b = y^{(i)} - \mathbf{w}^{\mathrm{T}} \mathbf{x}^{(i)} \quad \text{for any } \mathbf{x}^{(i)} \text{ such that } \alpha_i > 0$$

$$\mathbf{w} = (1)(-1)\begin{bmatrix}1\\1\end{bmatrix} + (1)(+1)\begin{bmatrix}2\\2\end{bmatrix} = \begin{bmatrix}1\\1\end{bmatrix}$$

$$b = (+1) - \begin{bmatrix}1 & 1\end{bmatrix}\begin{bmatrix}2\\2\end{bmatrix} = -3$$



- **Two samples are support vectors.**

$f(x) = x_1 + x_2 - 3$

Dual Form of Soft-Margin SVM

- Soft-margin SVM also considers slack variables.

Original optimization problem

$$\min_{\mathbf{w},\mathbf{b}} \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} + C\sum_{i=1}^{n}\xi_i \ \ \text{s.t.}\ (\mathbf{w}^{\mathbf{T}}\mathbf{x}^{(i)} + b)y^{(i)} \geq 1 - \xi_i, \forall\alpha_i \geq 0, \forall\xi_i \geq 0$$

Lagrangian form:

$$L(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}) = \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i\left[(\mathbf{w}^{\mathbf{T}}\mathbf{x}^{(i)} + b)y^{(i)} - 1 + \xi_i\right], \forall\alpha_i \geq 0$$

- Now, our goal is to solve

$$\min_{\mathbf{w},b,\boldsymbol{\xi}} \max_{\boldsymbol{\alpha}} L(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}) \ \text{ subject to } \forall\alpha \geq 0$$

Solving Soft-Margin SVM

- **Given** $\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)}) : 1 \leq i \leq n\}$, **where** $y^{(i)} \in \{-1, +1\}$,

$$\max_{\alpha_1,\cdots,\alpha_n} \left(\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y^{(i)} y^{(j)} \left(\mathbf{x}^{(i)} \cdot \mathbf{x}^{(j)} \right) \right)$$

$$\text{subject to} \begin{cases} \displaystyle\sum_{i=1}^{n} \alpha_i y^{(i)} = 0 \\ 0 \leq \alpha_i \leq C \quad i = 1, \cdots, n \end{cases}$$

It considers slack variables.

- **Solution**

$$\mathbf{w} = \sum_{i=1}^{n} \alpha_i y^{(i)} \mathbf{x}^{(i)}$$

$$b = y^{(i)} - \mathbf{w}^{\mathrm{T}} \mathbf{x}^{(i)} \quad \text{for any } \mathbf{x}^{(i)} \text{ such that } 0 < \alpha_i < C$$

Q&A