Support Vector Machine (SVM)

Machine Learning

Hamid R Rabiee – Zahra Dehghanian

Spring 2025

Sharif University
of Technology

Outline

- Margin concept
- Hard-Margin SVM
  - Dual Problem of Hard-Margin SVM
- Soft-Margin SVM
  - Dual Problem of Soft-Margin SVM

Hyperplanes

A *hyperplane* is a set of the form

$$\{x \mid a^T x = b\},$$

$a \in \mathbf{R}^n$, $a \neq 0$, and $b \in \mathbf{R}$.

a is the normal vector

Hyperplanes

A *hyperplane* is a set of the form

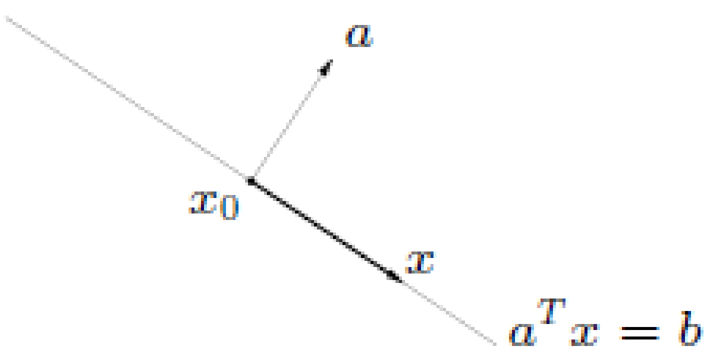
$$\{x \mid a^T x = b\},$$

$a \in \mathbf{R}^n$, $a \neq 0$, and $b \in \mathbf{R}$.

a is the normal vector

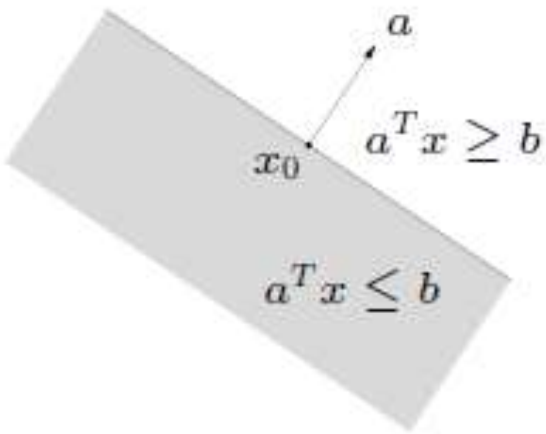
Geometrical interpretation

x_0 is any point in the hyperplane $\quad a^T x_0 = b \quad \{x \mid a^T(x - x_0) = 0\},$



halfspaces

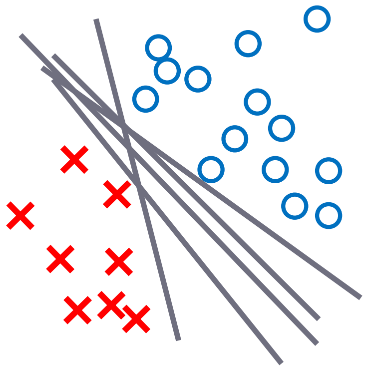
halfspace: set of the form $\{x \mid a^T x \leq b\}$ $(a \neq 0)$



Margin

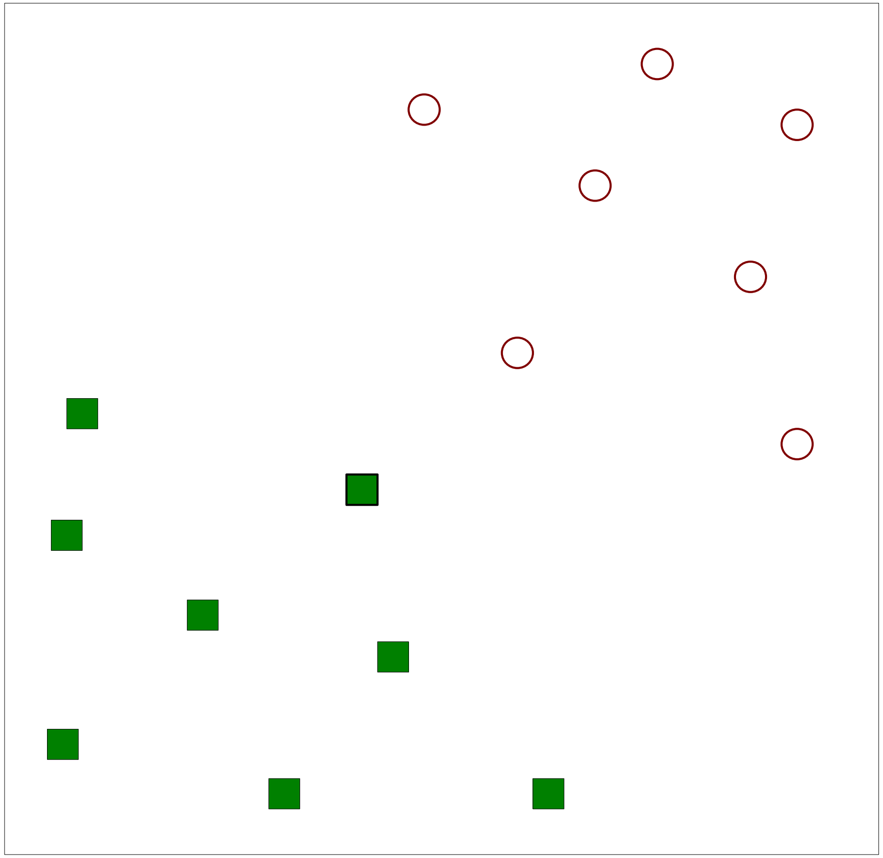
- Which line is better to select as the boundary to provide more generalization capability?

Larger margin provides better generalization to unseen data

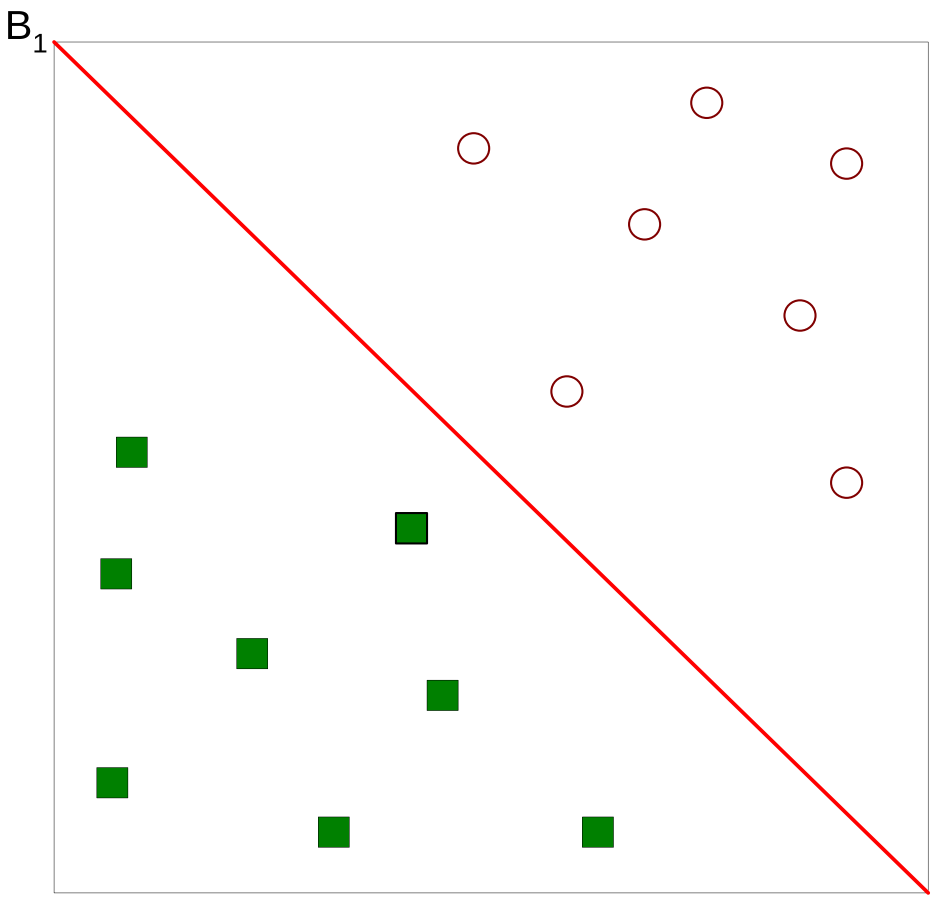


- **Margin** for a hyperplane that separates samples of two linearly separable classes is:
  - The smallest distance between the decision boundary and any of the training samples

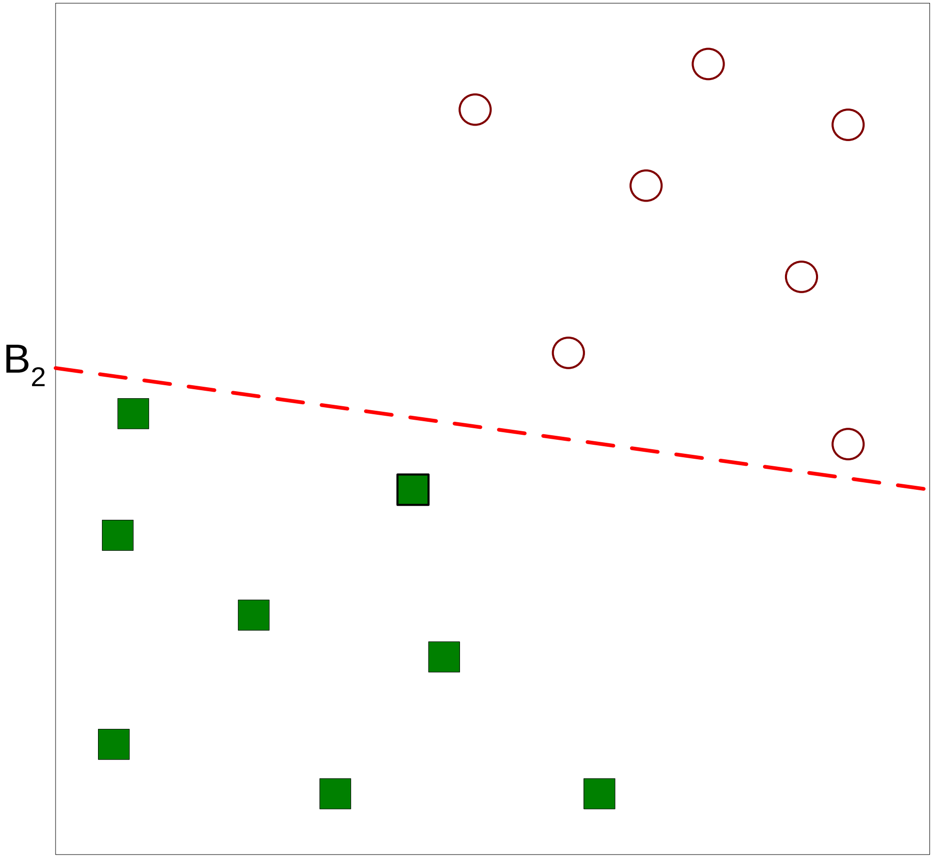
Maximum Margin Hyperplanes



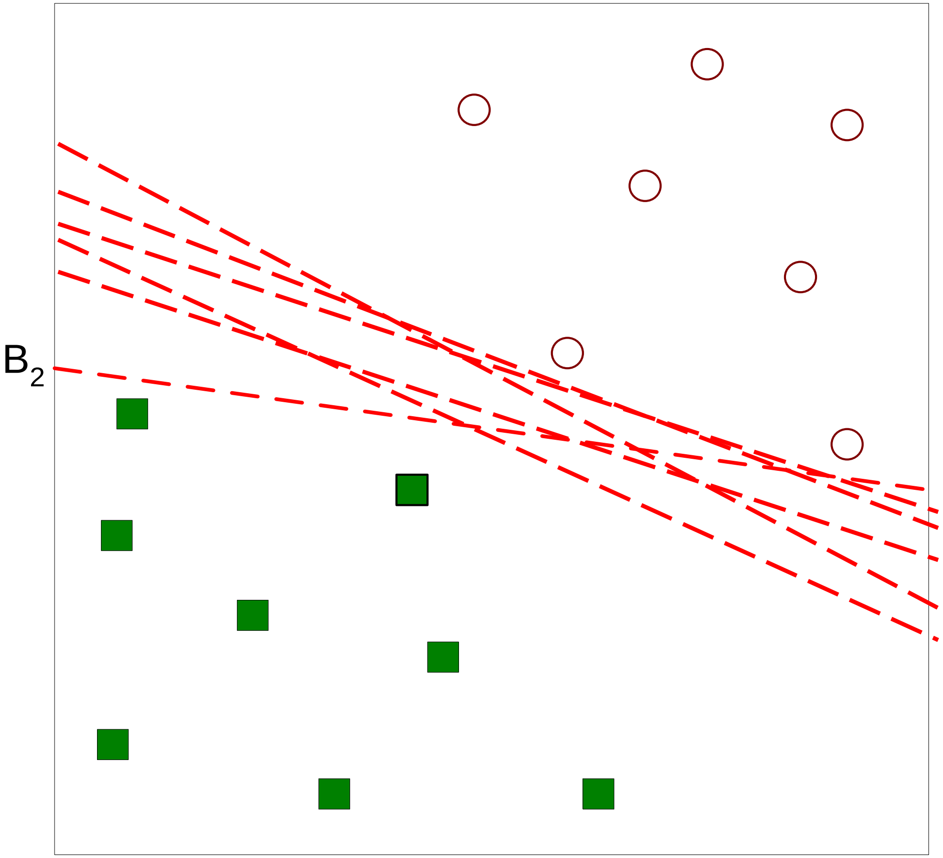
Maximum Margin Hyperplanes



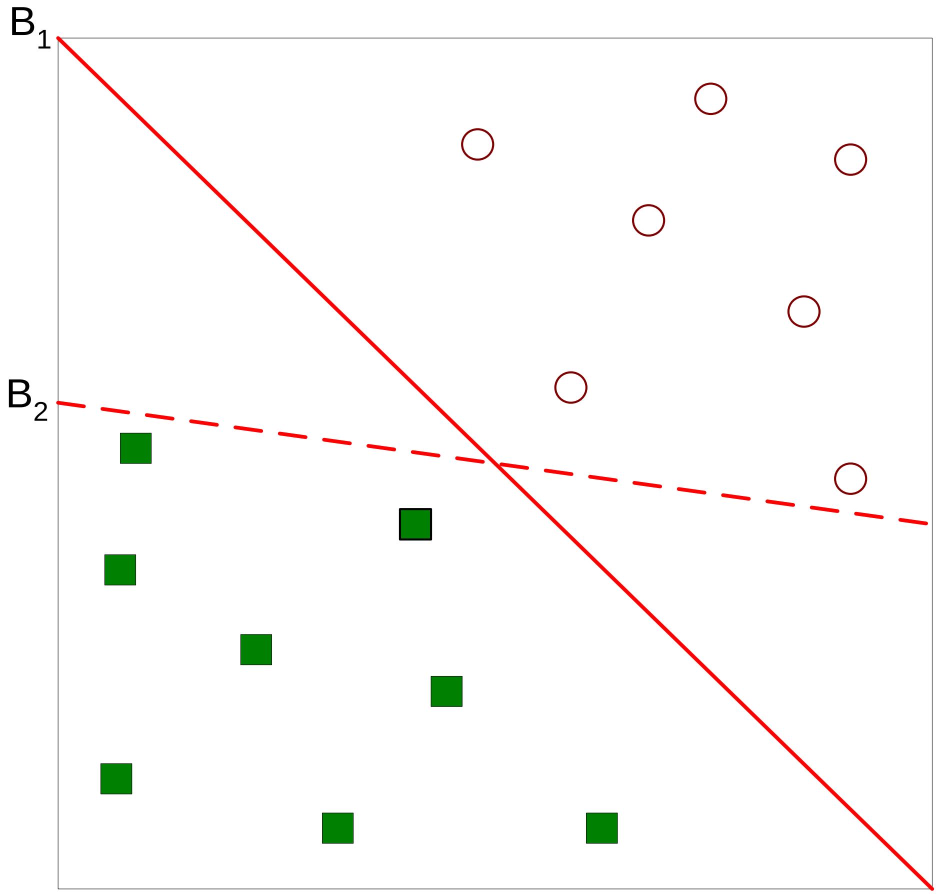
Maximum Margin Hyperplanes



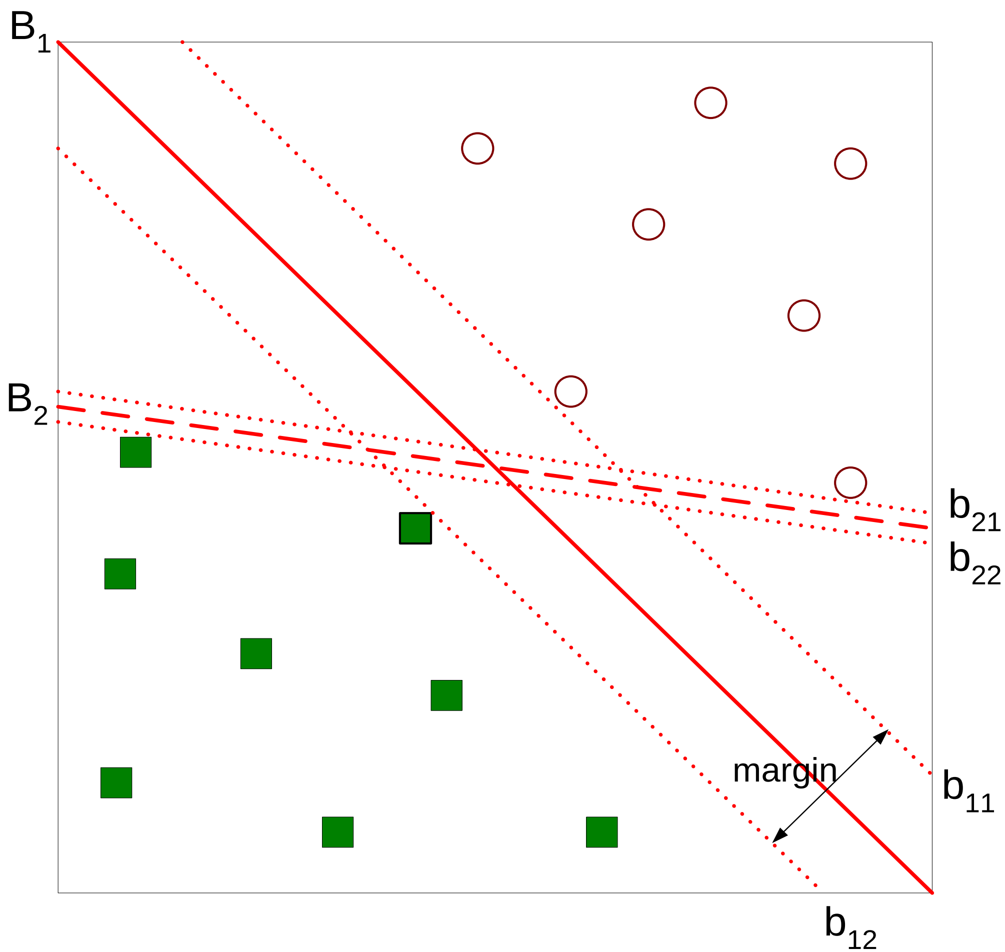
Maximum Margin Hyperplanes



Maximum Margin Hyperplanes



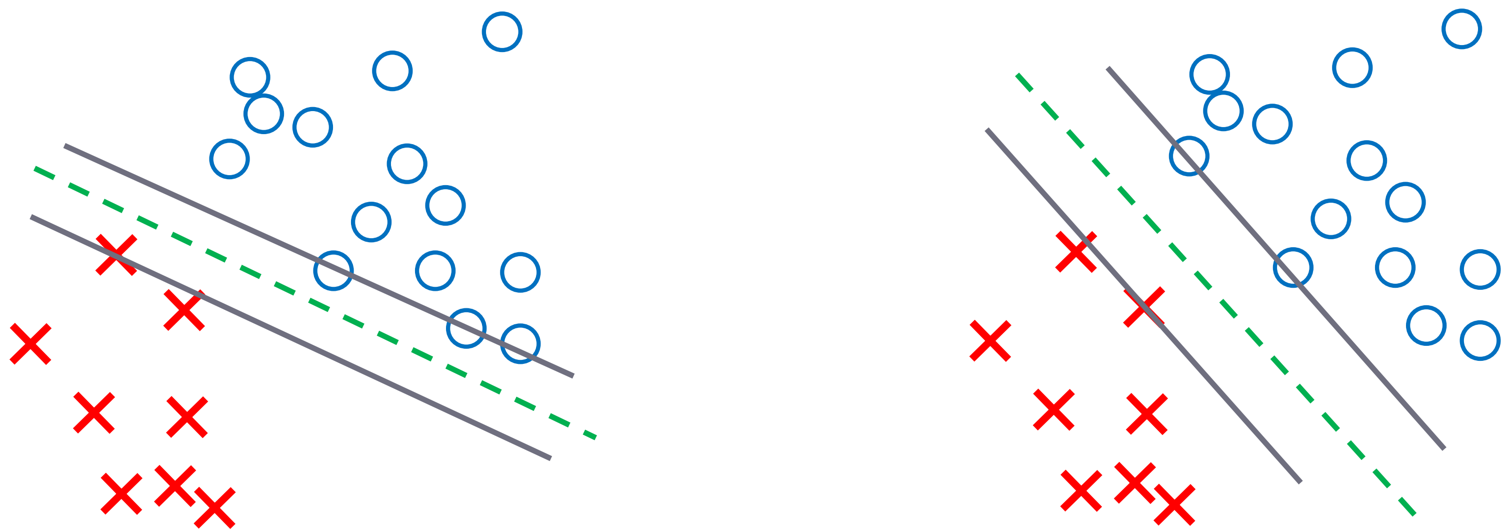
Maximum Margin Hyperplanes



maximizes

Maximum margin

- SVM finds the solution with maximum margin
  - Solution: a hyperplane that is farthest from all training samples



Larger margin

- The hyperplane with the largest margin has equal distances to the nearest sample of both classes

Linear SVM: Separable Case

A linear SVM is a classifier that searches for a hyperplane with the largest margin

$$(\mathbf{x_i}, y_i) \ (i = 1, 2, \ldots, N)$$

$$(x_{i1}, x_{i2}, \ldots, x_{id})^T \qquad y_i \in \{-1, 1\}$$

decision boundary of a linear classifier

$$\mathbf{w} \cdot \mathbf{x} + b = 0,$$

w and *b* are parameters of the model

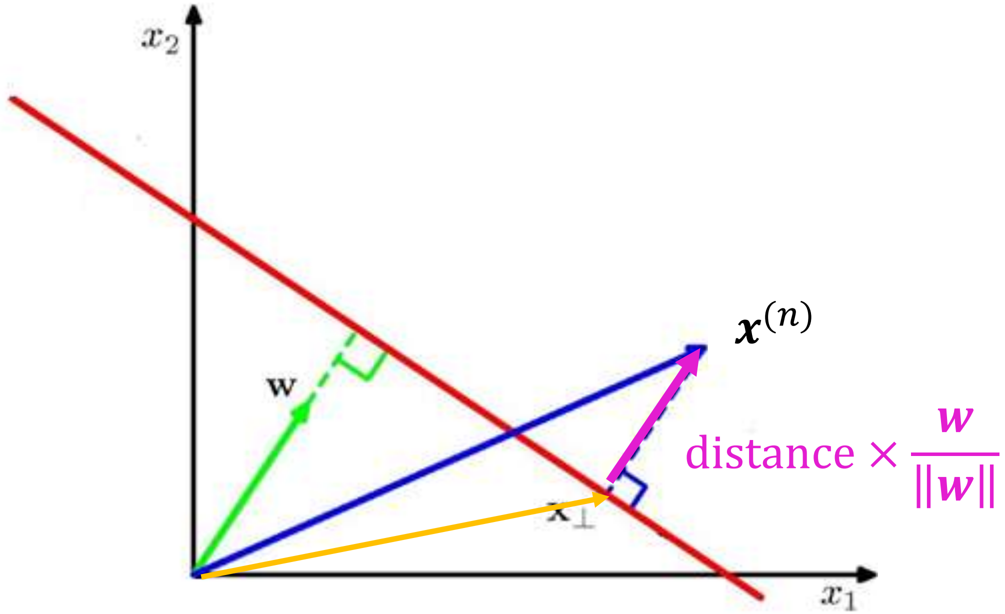
$$\mathbf{w} \cdot \mathbf{x}_s + b = k, \qquad k > 0.$$

$$\mathbf{w} \cdot \mathbf{x}_c + b = k', \qquad k' < 0.$$

$$y = \begin{cases} 1, & \text{if } \mathbf{w} \cdot \mathbf{z} + b > 0; \\ -1, & \text{if } \mathbf{w} \cdot \mathbf{z} + b < 0. \end{cases}$$

Distance between an $\boldsymbol{x}^{(n)}$ and the plane

$$\text{distance} = \frac{\left|\boldsymbol{w}^T \boldsymbol{x}^{(n)} + w_0\right|}{\|\boldsymbol{w}\|}$$

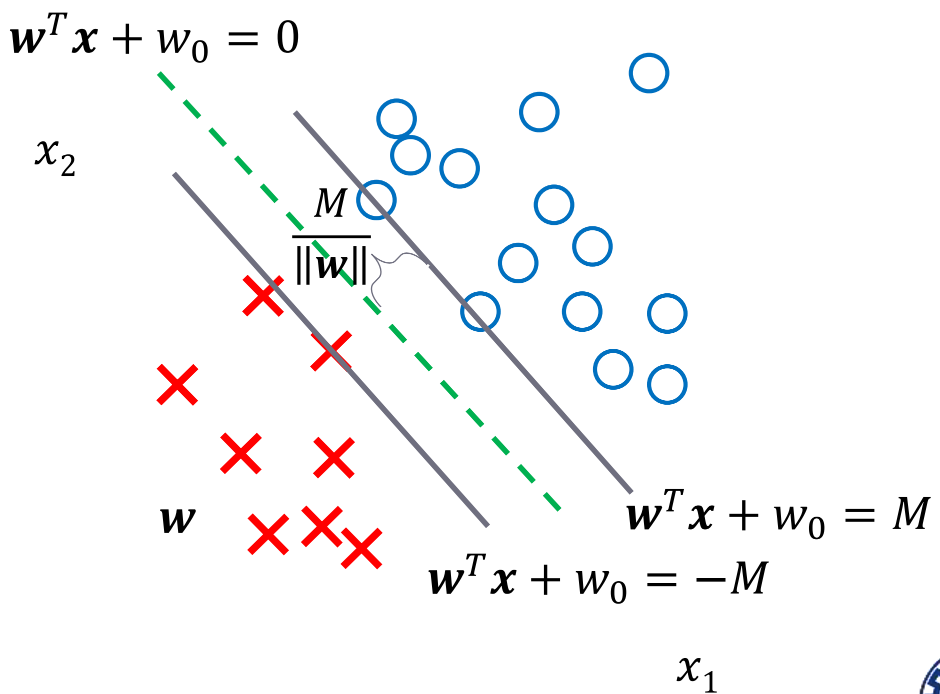


Hard-margin SVM: Optimization problem

$$\max_{M, \boldsymbol{w}, w_0} \frac{2M}{\|\boldsymbol{w}\|}$$

$$\text{s.t.} \left(\boldsymbol{w}^T \boldsymbol{x}^{(i)} + w_0\right) \geq M \quad \forall \boldsymbol{x}^{(i)} \in C_1 \longrightarrow y^{(i)} = 1$$

$$\left(\boldsymbol{w}^T \boldsymbol{x}^{(i)} + w_0\right) \leq -M \quad \forall \boldsymbol{x}^{(i)} \in C_2 \longrightarrow y^{(i)} = -1$$

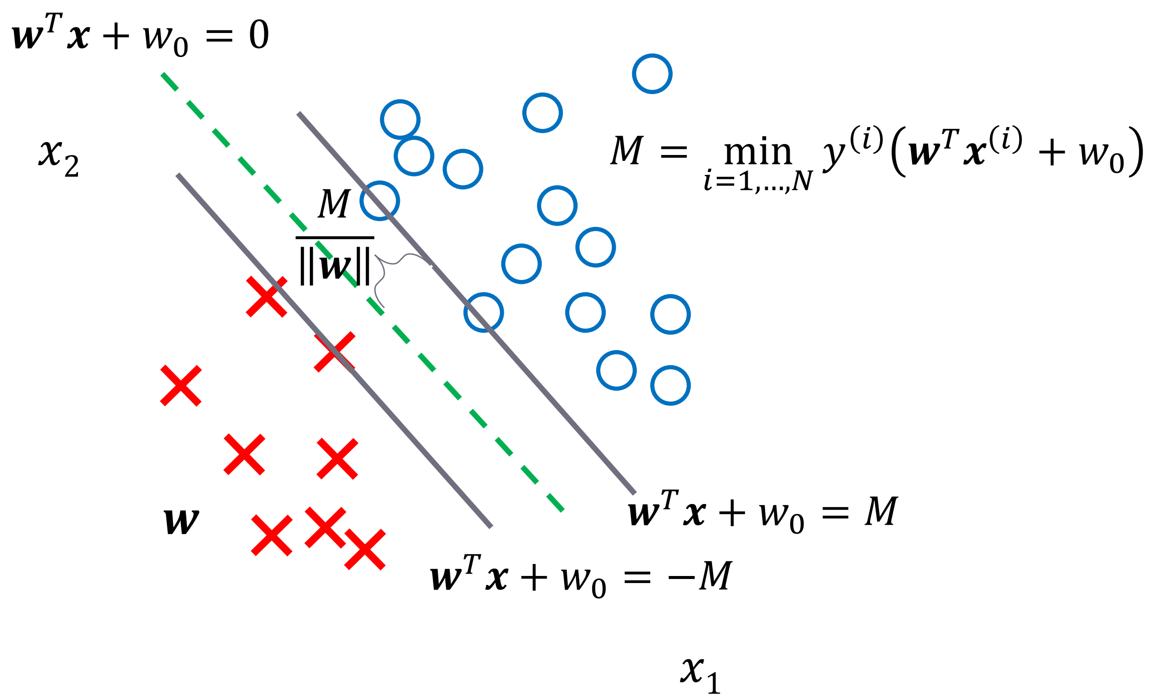


Margin: $2\frac{M}{\|\boldsymbol{w}\|}$

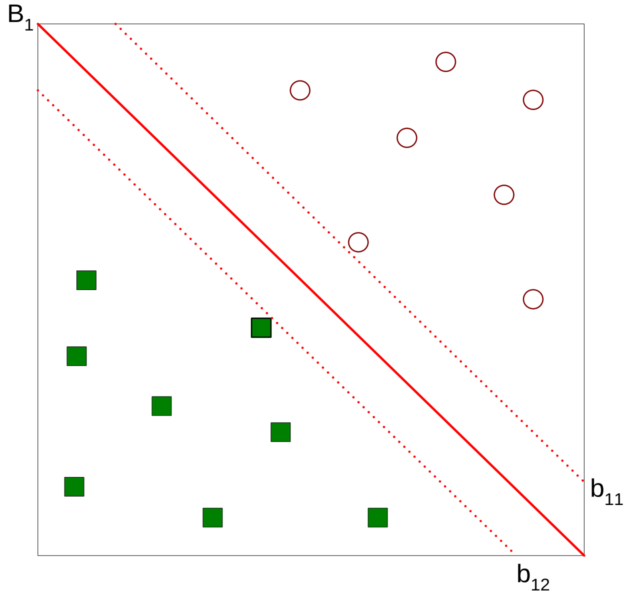
Hard-margin SVM: Optimization problem

$$\max_{M,\boldsymbol{w},w_0} \frac{2M}{\|\boldsymbol{w}\|}$$

$$\text{s.t. } y^{(i)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + w_0\right) \geq M \quad i = 1, \dots, N$$

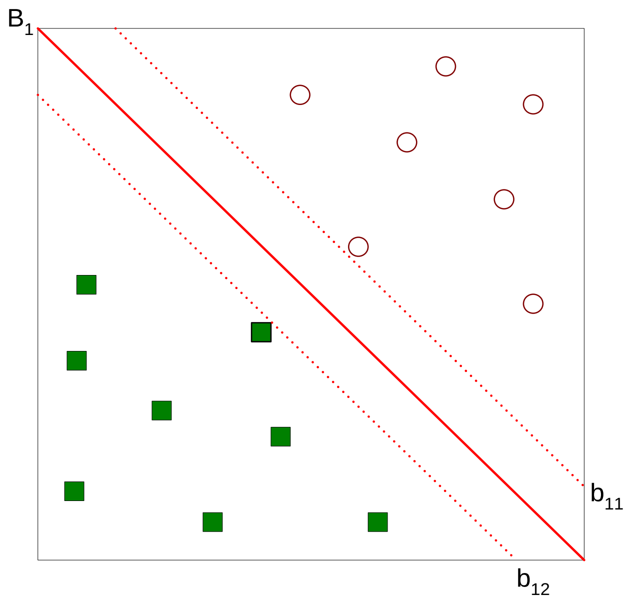


Linear SVM: Separable Case



$$b_{i1}: \ \mathbf{w} \cdot \mathbf{x} + b = 1,$$
$$b_{i2}: \ \mathbf{w} \cdot \mathbf{x} + b = -1.$$

Linear SVM: Separable Case



$$b_{i1}: \ \mathbf{w} \cdot \mathbf{x} + b = 1,$$
$$b_{i2}: \ \mathbf{w} \cdot \mathbf{x} + b = -1.$$

margin of the decision boundary is given by the distance between these two hyperplanes

$$d = \frac{2}{\|\mathbf{w}\|}$$

Hard-margin SVM: Optimization problem

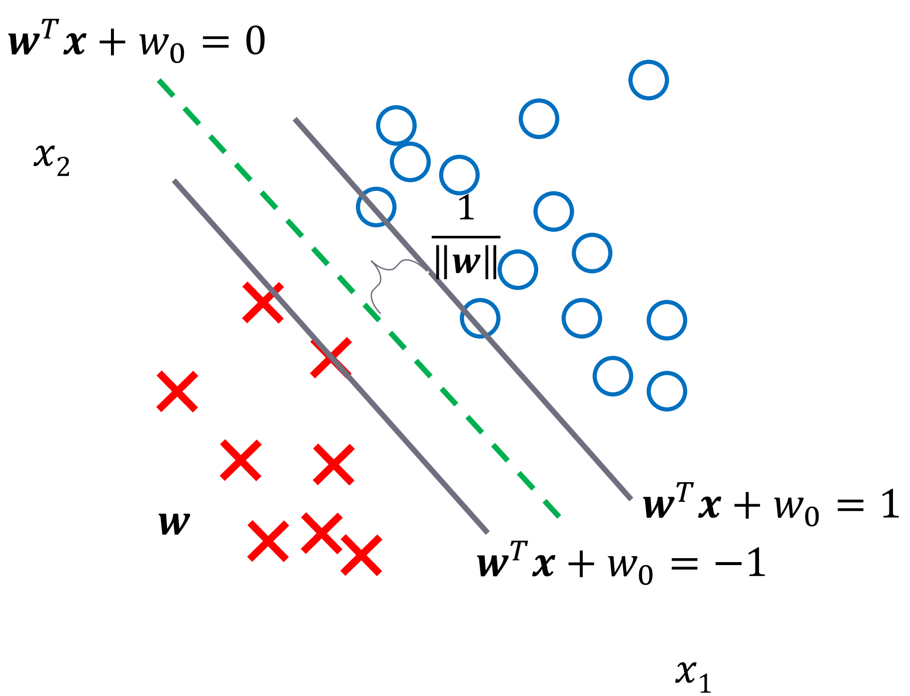
We can set $\boldsymbol{w}' = \frac{\boldsymbol{w}}{M}$, $w_0' = \frac{w_0}{M}$:

$$\max_{\boldsymbol{w},w_0} \frac{2}{\|\boldsymbol{w}\|}$$

$$\text{s.t.} \left(\boldsymbol{w}^T\boldsymbol{x}^{(n)} + w_0\right) \geq 1 \quad \forall y^{(n)} = 1$$

$$\left(\boldsymbol{w}^T\boldsymbol{x}^{(n)} + w_0\right) \leq -1 \quad \forall y^{(n)} = -1$$

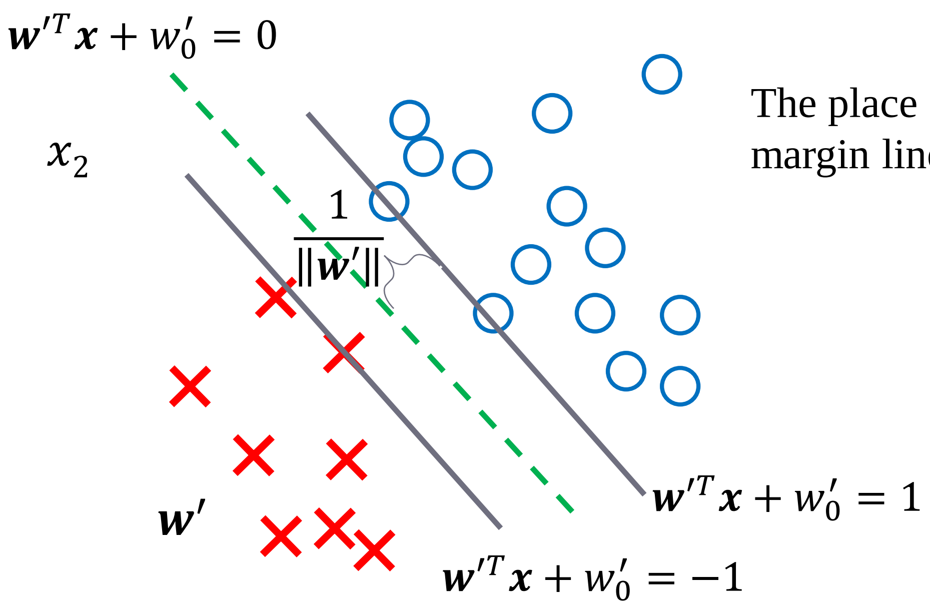
Margin: $\frac{2}{\|\boldsymbol{w}\|}$



Hard-margin SVM: Optimization problem

$$\max_{\boldsymbol{w}', w_0'} \frac{2}{\|\boldsymbol{w}'\|}$$

$$\text{s.t. } y^{(i)}\left(\boldsymbol{w}'^{T}\boldsymbol{x}^{(i)} + w_0'\right) \geq 1 \quad i = 1, \dots, N$$

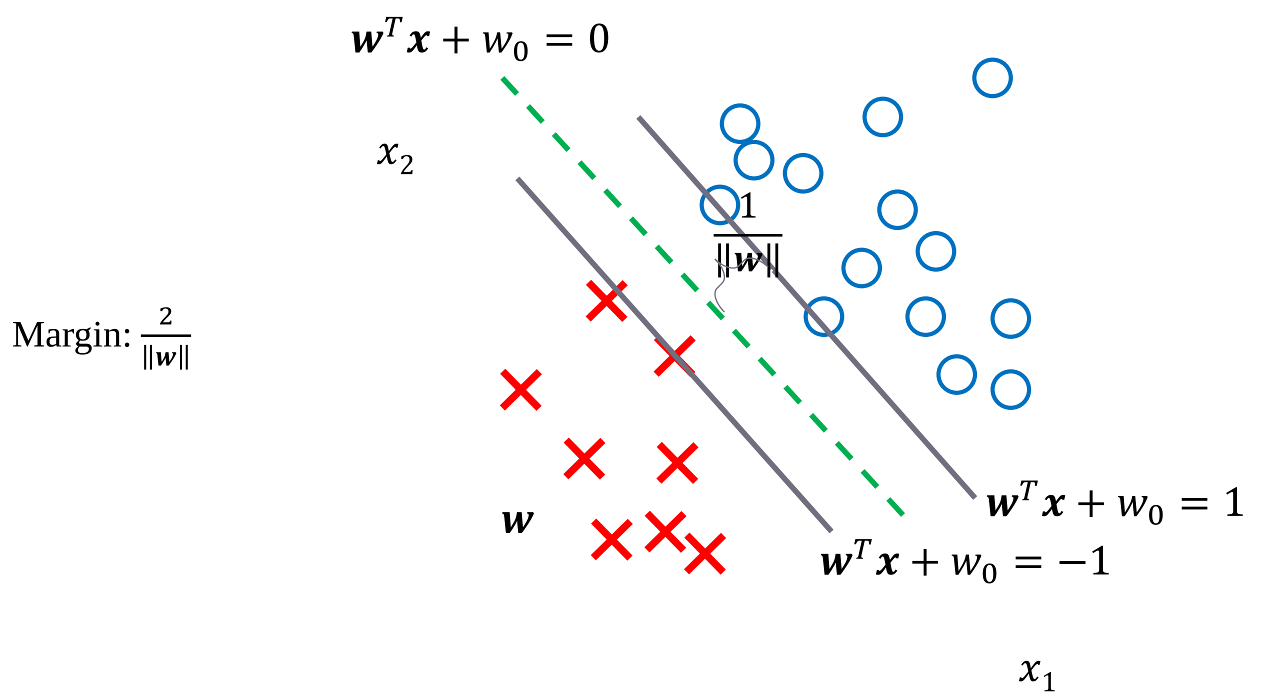


The place of boundary and margin lines do not change

Hard-margin SVM: Optimization problem

$$\max_{\boldsymbol{w},w_0} \frac{2}{\|\boldsymbol{w}\|}$$

$$\text{s.t. } y^{(i)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + w_0\right) \geq 1 \ , n = 1, \dots, N$$



SVM

$$\mathbf{w} \cdot \mathbf{x_i} + b \geq 1 \text{ if } y_i = 1,$$
$$\mathbf{w} \cdot \mathbf{x_i} + b \leq -1 \text{ if } y_i = -1$$

$$y_i(\mathbf{w} \cdot \mathbf{x_i} + b) \geq 1, \quad i = 1, 2, \ldots, N.$$

Definition 5.1 (Linear SVM: Separable Case). The learning task in SVM can be formalized as the following constrained optimization problem:

$$\min_{\mathbf{w}} \frac{\|\mathbf{w}\|^2}{2}$$
$$\text{subject to} \quad y_i(\mathbf{w} \cdot \mathbf{x_i} + b) \geq 1, \quad i = 1, 2, \ldots, N.$$

For Solving constrained optimization problem (like SVM Optimization) there exist Numerical approaches like **Quadratic Programming (QP) !**

Quadratic Programming (QP)

It is a convex Quadratic Programming (QP) problem

- There are computationally efficient packages to solve it.
- It has a <u>global</u> minimum (if any).

$$\min_{\boldsymbol{x}} \frac{1}{2}\boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{x} + \boldsymbol{c}^T\boldsymbol{x}$$

$$\text{s.t.} \quad \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}$$

$$\boldsymbol{E}\boldsymbol{x} = \boldsymbol{d}$$

Dual formulation of the SVM

- We are going to introduce the *dual* SVM problem which is equivalent to the original *primal* problem. The dual problem:
  - is often **easier**
  - It's **computationally more feasible** in high dimensional spaces where d is large
  - gives us further insights into the **optimal hyper-plane**
  - enable us to exploit the **kernel trick**

Optimization: Lagrangian multipliers

$$
\begin{aligned}
p^* = \min_{\boldsymbol{x}} f(\boldsymbol{x}) \\
\text{s.t. } g_i(\boldsymbol{x}) \leq 0 \quad i = 1, \dots, m \\
h_i(\boldsymbol{x}) = 0 \quad i = 1, \dots, p
\end{aligned}
$$

Optimization: Lagrangian multipliers

$$p^* = \min_{\boldsymbol{x}} f(\boldsymbol{x})$$
$$\text{s.t. } g_i(\boldsymbol{x}) \le 0 \quad i = 1, \dots, m$$
$$h_i(\boldsymbol{x}) = 0 \quad i = 1, \dots, p$$

Lagrangian multipliers

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\lambda}) = f(\boldsymbol{x}) + \sum_{i=1}^{m} \alpha_i g_i(\boldsymbol{x}) + \sum_{i=1}^{p} \lambda_i h_i(\boldsymbol{x})$$

Optimization: Lagrangian multipliers

$$p^* = \min_{\boldsymbol{x}} f(\boldsymbol{x})$$

$$\text{s.t. } g_i(\boldsymbol{x}) \le 0 \quad i = 1, \dots, m$$

$$h_i(\boldsymbol{x}) = 0 \quad i = 1, \dots, p$$

Lagrangian multipliers

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\lambda}) = f(\boldsymbol{x}) + \sum_{i=1}^{m} \alpha_i g_i(\boldsymbol{x}) + \sum_{i=1}^{p} \lambda_i h_i(\boldsymbol{x})$$

$$\max_{\{\alpha_i \ge 0\}, \{\lambda_i\}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\lambda}) = \begin{cases} \infty & \text{any } g_i(\boldsymbol{x}) > 0 \\ \infty & \text{any } h_i(\boldsymbol{x}) \ne 0 \\ f(\boldsymbol{x}) & \text{otherwise} \end{cases}$$

$$p^* = \min_{\boldsymbol{x}} \max_{\{\alpha_i \ge 0\}, \{\lambda_i\}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\lambda})$$

Optimization: Dual problem

- In general, we have:

$$\max_x \min_y h(x, y) \leq \min_y \max_x h(x, y)$$

- **Primal problem**: $p^* = \min_{\boldsymbol{x}} \max_{\{\alpha_i \geq 0\}, \{\lambda_i\}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\lambda})$

- **Dual problem**: $d^* = \max_{\{\alpha_i \geq 0\}, \{\lambda_i\}} \min_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\lambda})$
  - Obtained by swapping the order of min and max
  - $d^* \leq p^*$

- When the original problem is convex (f and g are convex functions and h is affine), we have strong duality $d^* = p^*$

Hard-margin SVM: Dual problem

$$\min_{\boldsymbol{w},w_0} \frac{1}{2}\|\boldsymbol{w}\|^2$$

$$\text{s.t.} \quad y^{(i)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + w_0\right) \geq 1 \quad i = 1, \dots, N$$

- By incorporating the constraints through Lagrangian multipliers, we will have:

$$\min_{\boldsymbol{w},w_0} \max_{\{\alpha_n \geq 0\}} \left\{\frac{1}{2}\|\boldsymbol{w}\|^2 + \sum_{n=1}^{N} \alpha_n \left(1 - y^{(n)}(\boldsymbol{w}^T\boldsymbol{x}^{(n)} + w_0)\right)\right\}$$

Hard-margin SVM: Dual problem

$$\min_{\boldsymbol{w},w_0} \frac{1}{2}\|\boldsymbol{w}\|^2$$
$$\text{s.t.} \quad y^{(i)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + w_0\right) \geq 1 \quad i = 1, \dots, N$$

- By incorporating the constraints through Lagrangian multipliers, we will have:

$$\min_{\boldsymbol{w},w_0} \max_{\{\alpha_n \geq 0\}} \left\{\frac{1}{2}\|\boldsymbol{w}\|^2 + \sum_{n=1}^{N} \alpha_n \left(1 - y^{(n)}(\boldsymbol{w}^T\boldsymbol{x}^{(n)} + w_0)\right)\right\}$$

- Dual problem (changing the order of min and max in the above problem):

$$\max_{\{\alpha_n \geq 0\}} \min_{\boldsymbol{w},w_0} \left\{\frac{1}{2}\|\boldsymbol{w}\|^2 + \sum_{n=1}^{N} \alpha_n \left(1 - y^{(n)}(\boldsymbol{w}^T\boldsymbol{x}^{(n)} + w_0)\right)\right\}$$

Hard-margin SVM: Dual problem

$$\max_{\{\alpha_n \geq 0\}} \min_{\boldsymbol{w}, w_0} \mathcal{L}(\boldsymbol{w}, w_0, \boldsymbol{\alpha})$$

$$\mathcal{L}(\boldsymbol{w}, w_0, \boldsymbol{\alpha}) = \frac{1}{2}\|\boldsymbol{w}\|^2 + \sum_{n=1}^{N} \left(1 - y^{(n)}(\boldsymbol{w}^T \boldsymbol{x}^{(n)} + w_0)\right)$$

Sharif University of Technology

Hard-margin SVM: Dual problem

$$\max_{\{\alpha_n \geq 0\}} \min_{\boldsymbol{w}, w_0} \mathcal{L}(\boldsymbol{w}, w_0, \boldsymbol{\alpha})$$

$$\mathcal{L}(\boldsymbol{w}, w_0, \boldsymbol{\alpha}) = \frac{1}{2}\|\boldsymbol{w}\|^2 + \sum_{n=1}^{N} \left(1 - y^{(n)}(\boldsymbol{w}^T \boldsymbol{x}^{(n)} + w_0)\right)$$

$$\nabla_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w}, w_0, \boldsymbol{\alpha}) = 0 \Rightarrow \boldsymbol{w} \quad \sum_{n=1}^{N} \alpha_n y^{(n)} \boldsymbol{x}^{(n)} = \boldsymbol{0}$$

$$\Rightarrow \boldsymbol{w} = \sum_{n=1}^{N} \alpha_n y^{(n)} \boldsymbol{x}^{(n)}$$

Hard-margin SVM: Dual problem

$$\max_{\{\alpha_n \geq 0\}} \min_{\boldsymbol{w}, w_0} \mathcal{L}(\boldsymbol{w}, w_0, \boldsymbol{\alpha})$$

$$\mathcal{L}(\boldsymbol{w}, w_0, \boldsymbol{\alpha}) = \frac{1}{2}\|\boldsymbol{w}\|^2 + \sum_{n=1}^{N} \left(1 - y^{(n)}(\boldsymbol{w}^T \boldsymbol{x}^{(n)} + w_0)\right)$$

$$\nabla_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w}, w_0, \boldsymbol{\alpha}) = 0 \Rightarrow \boldsymbol{w} \quad \sum_{n=1}^{N} \alpha_n y^{(n)} \boldsymbol{x}^{(n)} = \boldsymbol{0}$$

$$\Rightarrow \boldsymbol{w} = \sum_{n=1}^{N} \alpha_n y^{(n)} \boldsymbol{x}^{(n)}$$

$$\frac{\partial \mathcal{L}(\boldsymbol{w}, w_0, \boldsymbol{\alpha})}{\partial w_0} = 0 \Rightarrow \quad \underbrace{-\sum_{n=1}^{N} \alpha_n y^{(n)} = 0}$$

w_0 do not appear, instead, a “global” constraint on $\boldsymbol{\alpha}$ is created.

Substituting

$$\boldsymbol{w} = \sum_{n=1}^{N} \alpha_n y^{(n)} \boldsymbol{x}^{(n)} \qquad \sum_{n=1}^{N} \alpha_n y^{(n)} = 0$$

In the Largrangian

$$\mathcal{L}(\boldsymbol{w}, w_0, \boldsymbol{\alpha}) = \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \sum_{n=1}^{N} \alpha_n \left(1 - y^{(n)}(\boldsymbol{w}^T\boldsymbol{x}^{(n)} + w_0)\right)$$

Substituting

$$\boldsymbol{w} = \sum_{n=1}^{N} \alpha_n y^{(n)} \boldsymbol{x}^{(n)} \qquad \sum_{n=1}^{N} \alpha_n y^{(n)} = 0$$

In the Largrangian

$$\mathcal{L}(\boldsymbol{w}, w_0, \boldsymbol{\alpha}) = \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \sum\nolimits_{n=1}^{N} \alpha_n \left(1 - y^{(n)}(\boldsymbol{w}^T\boldsymbol{x}^{(n)} + w_0)\right)$$

We get

$$\mathcal{L}(\boldsymbol{\alpha}) = \sum\nolimits_{n=1}^{N} \alpha_n - \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} \alpha_n \alpha_m y^{(n)} y^{(m)} {\boldsymbol{x}^{(n)}}^T \boldsymbol{x}^{(m)}$$

Maximize w.r.t. $\boldsymbol{\alpha}$ subject to $\alpha_n \geq 0$ for $n = 1, \dots, N$ and $\sum_{n=1}^{N} \alpha_n y^{(n)} = 0$

Hard-margin SVM: Dual problem

$$\max_{\boldsymbol{\alpha}} \left\{ \sum_{n=1}^{N} \alpha_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} \alpha_n \alpha_m y^{(n)} y^{(m)} \boldsymbol{x}^{(n)^T} \boldsymbol{x}^{(m)} \right\}$$

Subject to $\sum_{n=1}^{N} \alpha_n y^{(n)} = 0$

$$\alpha_n \geq 0 \quad n = 1, \dots, N$$

- It is a convex QP

Solution

- Quadratic programming:

$$\min_{\boldsymbol{\alpha}} \frac{1}{2}\boldsymbol{\alpha}^T \begin{bmatrix} y^{(1)}y^{(1)}\boldsymbol{x}^{(1)^T}\boldsymbol{x}^{(1)} & \cdots & y^{(1)}y^{(N)}\boldsymbol{x}^{(1)^T}\boldsymbol{x}^{(N)} \\ \vdots & \ddots & \vdots \\ y^{(N)}y^{(1)}\boldsymbol{x}^{(N)^T}\boldsymbol{x}^{(1)} & \cdots & y^{(N)}y^{(N)}\boldsymbol{x}^{(N)^T}\boldsymbol{x}^{(N)} \end{bmatrix} \boldsymbol{\alpha} + (-\mathbf{1})^T\boldsymbol{\alpha}$$

$$\text{s.t.} -\boldsymbol{\alpha} \leq \mathbf{0}$$

$$\boldsymbol{y}^T\boldsymbol{\alpha} = \mathbf{0}$$

Finding the hyperplane

- After finding $\boldsymbol{\alpha}$ by QP, we find $\boldsymbol{w}$:

$$\boldsymbol{w} = \sum_{n=1}^{N} \alpha_n y^{(n)} \boldsymbol{x}^{(n)}$$

- How to find w_0?
  - we discuss it after introducing support vectors

Optimal Point

- Necessary conditions for the solution $[\boldsymbol{w}^*, w_0^*, \boldsymbol{\alpha}^*]$:
  - $\nabla_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w}, w_0, \boldsymbol{\alpha})|_{\boldsymbol{w}^*, w_0^*, \boldsymbol{\alpha}^*} = 0$
  - $\frac{\partial \mathcal{L}(\boldsymbol{w}, w_0, \boldsymbol{\alpha})}{\partial w_0}|_{\boldsymbol{w}^*, w_0^*, \boldsymbol{\alpha}^*} = 0$
  - $\alpha_n^* \geq 0 \quad n = 1, \dots, N$

- $y^{(n)}\left(\boldsymbol{w}^{*T}\boldsymbol{x}^{(n)} + w_0^*\right) \geq 1 \quad n = 1, \dots, N$
- $\alpha_i^* \left(1 - y^{(n)}\left(\boldsymbol{w}^{*T}\boldsymbol{x}^{(n)} + w_0^*\right)\right) = 0 \quad n = 1, \dots, N$

Optimization: Lagrangian multipliers

$$p^* = \min_{\boldsymbol{x}} f(\boldsymbol{x})$$

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

$$h_i(\boldsymbol{x}) = 0 \quad i = 1, \dots, p$$

Lagrangian multipliers

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\lambda}) = f(\boldsymbol{x}) + \sum_{i=1}^{m} \alpha_i g_i(\boldsymbol{x}) + \sum_{i=1}^{p} \lambda_i h_i(\boldsymbol{x})$$

$$\max_{\{\alpha_i \geq 0\}, \{\lambda_i\}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\lambda}) = \begin{cases} \infty & \text{any } g_i(\boldsymbol{x}) > 0 \\ \infty & \text{any } h_i(\boldsymbol{x}) \neq 0 \\ f(\boldsymbol{x}) & \text{otherwise} \end{cases}$$

$$p^* = \min_{\boldsymbol{x}} \max_{\{\alpha_i \geq 0\}, \{\lambda_i\}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\alpha}, \boldsymbol{\lambda})$$

Optimal Point

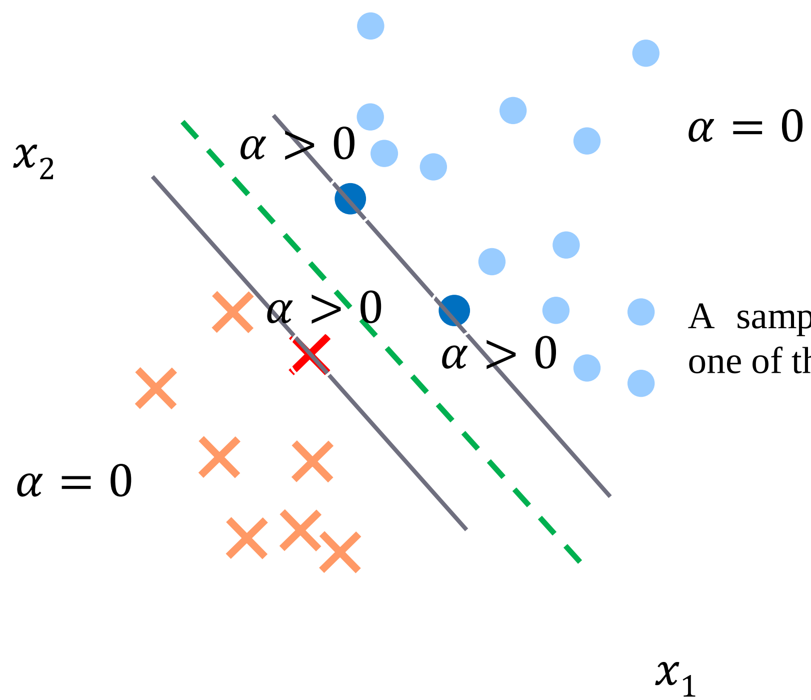
- Necessary conditions for the solution $[\boldsymbol{w}^*, w_0^*, \boldsymbol{\alpha}^*]$:
  - $\nabla_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w}, w_0, \boldsymbol{\alpha})|_{\boldsymbol{w}^*, w_0^*, \boldsymbol{\alpha}^*} = 0$
  - $\frac{\partial \mathcal{L}(\boldsymbol{w}, w_0, \boldsymbol{\alpha})}{\partial w_0}|_{\boldsymbol{w}^*, w_0^*, \boldsymbol{\alpha}^*} = 0$
  - $\alpha_n^* \geq 0 \quad n = 1, \dots, N$

- $y^{(n)}\left(\boldsymbol{w}^{*T} \boldsymbol{x}^{(n)} + w_0^*\right) \geq 1 \quad n = 1, \dots, N$
- $\alpha_i^* \left(1 - y^{(n)}\left(\boldsymbol{w}^{*T} \boldsymbol{x}^{(n)} + w_0^*\right)\right) = 0 \quad n = 1, \dots, N$

Karush-Kuhn-Tucker (KKT) conditions

Hard-margin SVM: Support vectors

- **Inactive** constraint: $y^{(n)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(n)} + w_0\right) > 1$
  - $\Rightarrow \alpha_n = 0$ and thus $\boldsymbol{x}^{(n)}$ is not a support vector.
- **Active** constraint: $y^{(n)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(n)} + w_0\right) = 1$
  - $\Rightarrow \alpha_n$ can be greater than 0 and thus $\boldsymbol{x}^{(i)}$ can be a support vector.



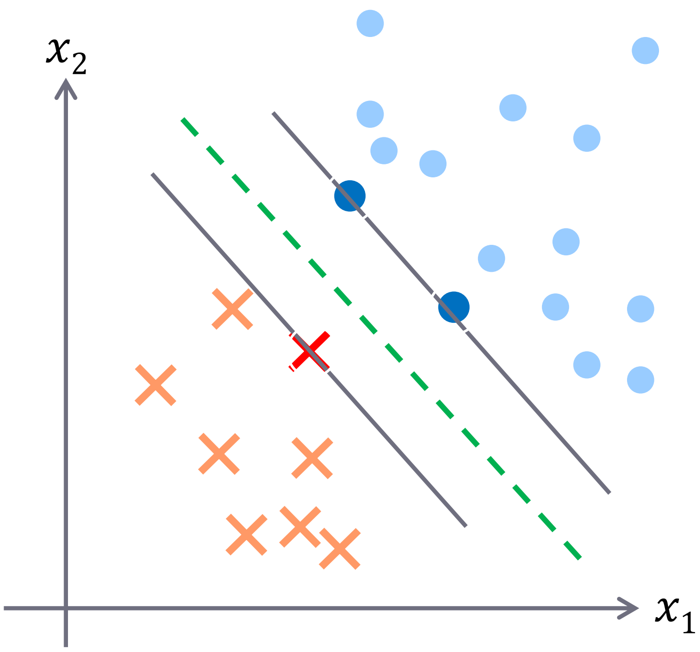
A sample with $\alpha_n > 0$ can also lie on one of the margin hyperplanes

Hard-margin SVM: Support vectors

- Support Vectors (SVs)= $\{\boldsymbol{x}^{(n)} | \alpha_n > 0\}$

- The **direction** of hyper-plane can be found only based on support vectors:

$$\boldsymbol{w} = \sum_{\alpha_n > 0} \alpha_n \, y^{(n)} \boldsymbol{x}^{(n)}$$



Finding the hyperplane

- After finding $\boldsymbol{\alpha}$ by QP, we find $\boldsymbol{w}$:

$$\boldsymbol{w} = \sum_{n=1}^{N} \alpha_n y^{(n)} \boldsymbol{x}^{(n)}$$

- How to find w_0?
  - Each of the samples that has α_s is on the margin, thus we solve for w_0 using any of SVs:

$$y^{(s)}\left(\boldsymbol{w}^T \boldsymbol{x}^{(s)} + w_0\right) = 1$$

$$\Rightarrow w_0 = y^{(s)} - \boldsymbol{w}^T \boldsymbol{x}^{(s)}$$

Hard-margin SVM: Dual problem Classifying new samples using only SVs

- Classification of a new sample $\boldsymbol{x}$:

$$\hat{y} = \text{sign}\left(w_0 + \boldsymbol{w}^T \boldsymbol{x}\right)$$

$$\hat{y} = \text{sign}\left(w_0 + \left(\sum_{\alpha_n > 0} \alpha_n y^{(n)} \boldsymbol{x}^{(n)}\right)^T \boldsymbol{x}\right)$$

$$\hat{y} = \text{sign}(\underbrace{y^{(s)} - \sum_{\alpha_n > 0} \alpha_n y^{(n)} \boldsymbol{x}^{(n)^T} \boldsymbol{x}^{(s)}}_{w_0} + \sum_{\alpha_n > 0} \alpha_n y^{(n)} \boldsymbol{x}^{(n)^T} \boldsymbol{x})$$

Support vectors are sufficient to predict labels of new samples

- The classifier is based on the expansion in terms of dot products of $\boldsymbol{x}$ with support vectors.

Hard-margin SVM dual problem: An important property

$$\max_{\boldsymbol{\alpha}} \left\{ \sum_{n=1}^{N} \alpha_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} \alpha_n \alpha_m y^{(n)} y^{(m)} \boldsymbol{x}^{(n)^T} \boldsymbol{x}^{(m)} \right\}$$

Subject to $\sum_{n=1}^{N} \alpha_n y^{(n)} = 0$

$$\alpha_n \geq 0 \quad n = 1, \dots, N$$

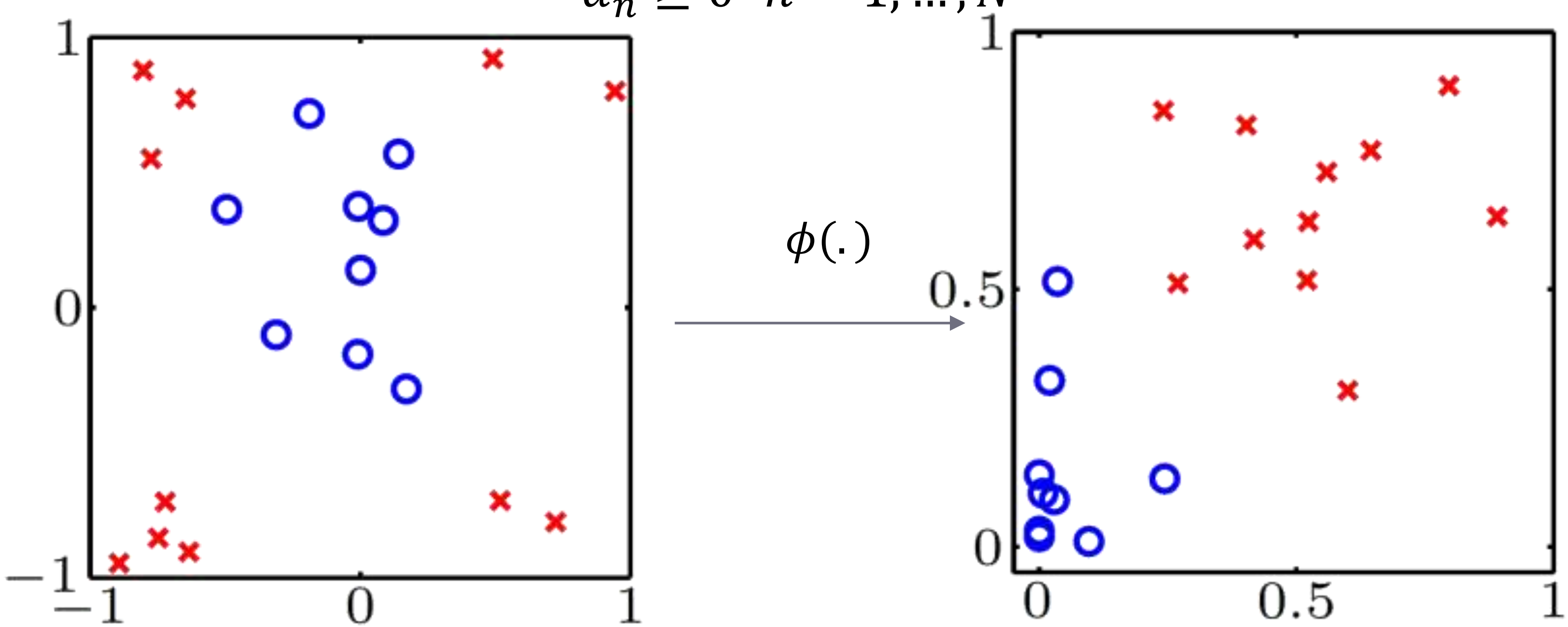
- Only the dot product of each pair of training data appears in the optimization problem
  - An important property that is helpful to extend to non-linear SVM
  - We will talk about it later (kernel-based methods)

In the transformed space

$$\max_{\boldsymbol{\alpha}} \left\{ \sum_{n=1}^{N} \alpha_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} \alpha_n \alpha_m y^{(n)} y^{(m)} \phi\left(\boldsymbol{x}^{(n)}\right)^T \phi\left(\boldsymbol{x}^{(m)}\right) \right\}$$

Subject to $\sum_{n=1}^{N} \alpha_n y^{(n)} = 0$

$\alpha_n \geq 0 \quad n = 1, \ldots, N$



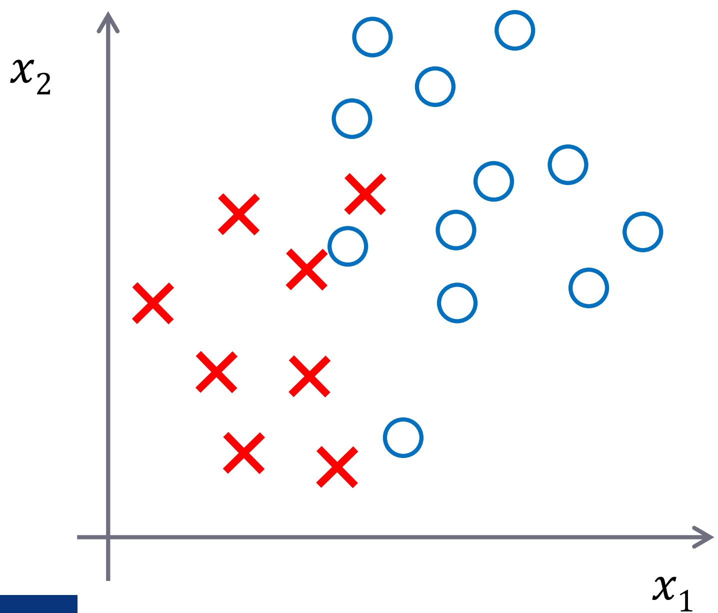
Beyond linear separability

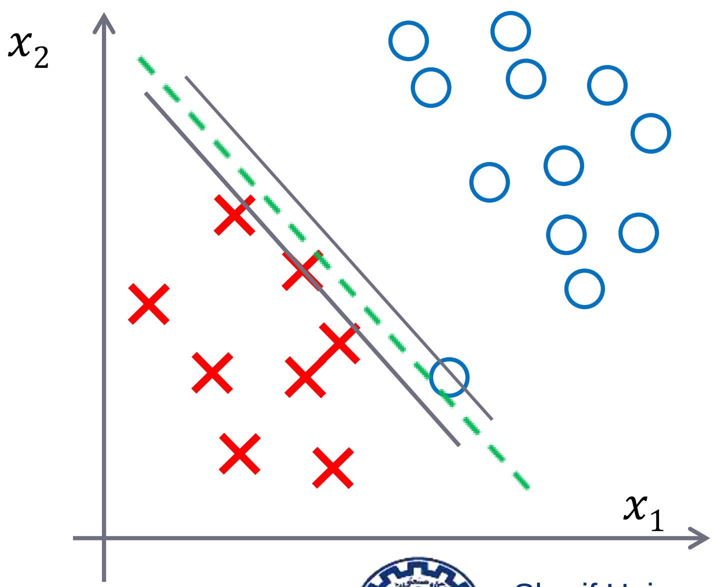
- When training samples are not linearly separable, it has no solution.

- How to extend it to find a solution even though the classes are not exactly linearly separable.

Near linear separability

- How to extend the hard-margin SVM to allow classification error
  - Overlapping classes that can be approximately separated by a linear boundary
  - Noise in the linearly separable classes





Near linear separability: Soft-margin SVM

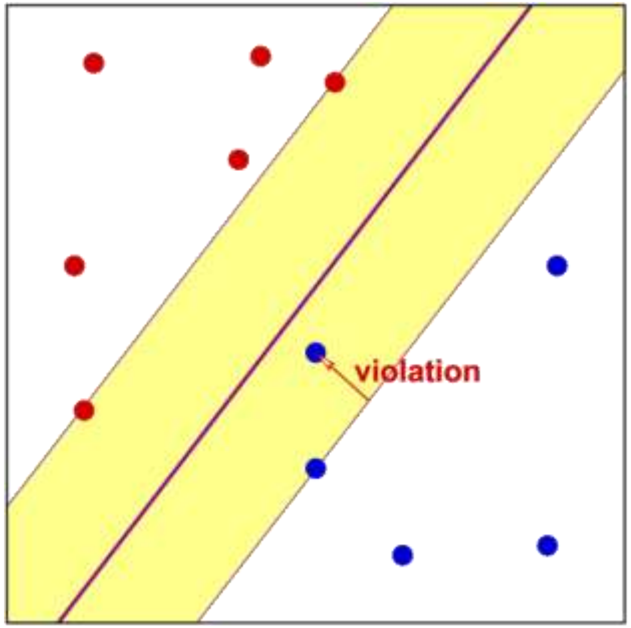
- Minimizing the number of misclassified points?!
  - NP-complete

- Soft margin:
  - Maximizing a margin while trying to minimize the *distance* between misclassified points and their correct margin plane

Error measure

- Margin violation amount ξ_n ($\xi_n \geq 0$):
  - $y^{(n)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(n)} + w_0\right) \geq 1 - \xi_n$

- Total violation: $\sum_{n=1}^{N} \xi_n$



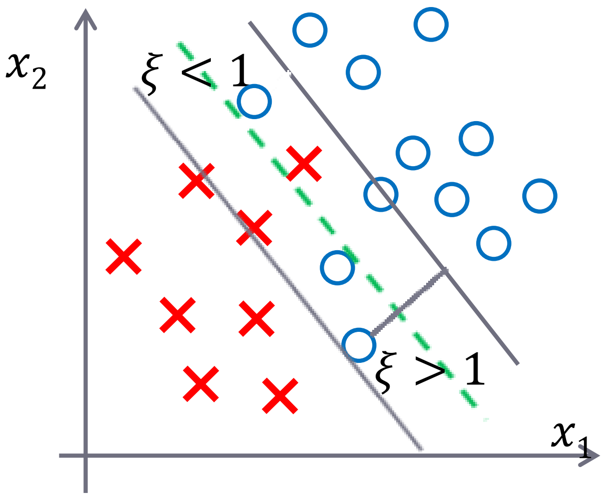
Soft-margin SVM: Optimization problem

- SVM with slack variables: allows samples to fall within the margin, but penalizes them

$$\min_{\boldsymbol{w},w_0,\{\xi_n\}_{n=1}^{N}} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{n=1}^{N}\xi_n$$

$$\text{s.t.} \quad y^{(n)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(n)} + w_0\right) \geq 1 - \xi_n \quad n = 1, \dots, N$$

$$\xi_n \geq 0$$



ξ_n: **slack** variables

$0 < \xi_n < 1$: if $\boldsymbol{x}^{(n)}$is correctly classified but inside margin

$\xi_n > 1$: if $\boldsymbol{x}^{(n)}$ is misclassifed

Soft-margin SVM

- linear penalty (hinge loss) for a sample if it is misclassified or lied in the margin
  - tries to maintain ξ_n small while maximizing the margin.
  - always finds a solution (as opposed to hard-margin SVM)
  - more robust to the outliers

- Soft margin problem is still a convex QP

Soft-margin SVM: Parameter C

- C is a tradeoff parameter:
  - small C allows margin constraints to be easily ignored
    - large margin
  - large C makes constraints hard to ignore
    - narrow margin

- $C \to \infty$ enforces all constraints: hard margin

- C can be determined using a technique like cross-validation

Soft-margin SVM: Cost function

$$\min_{\boldsymbol{w}, w_0, \{\xi_n\}_{n=1}^{N}} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{n=1}^{N}\xi_n$$

$$\text{s.t.} \quad y^{(n)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(n)} + w_0\right) \geq 1 - \xi_n \quad n = 1, \dots, N$$

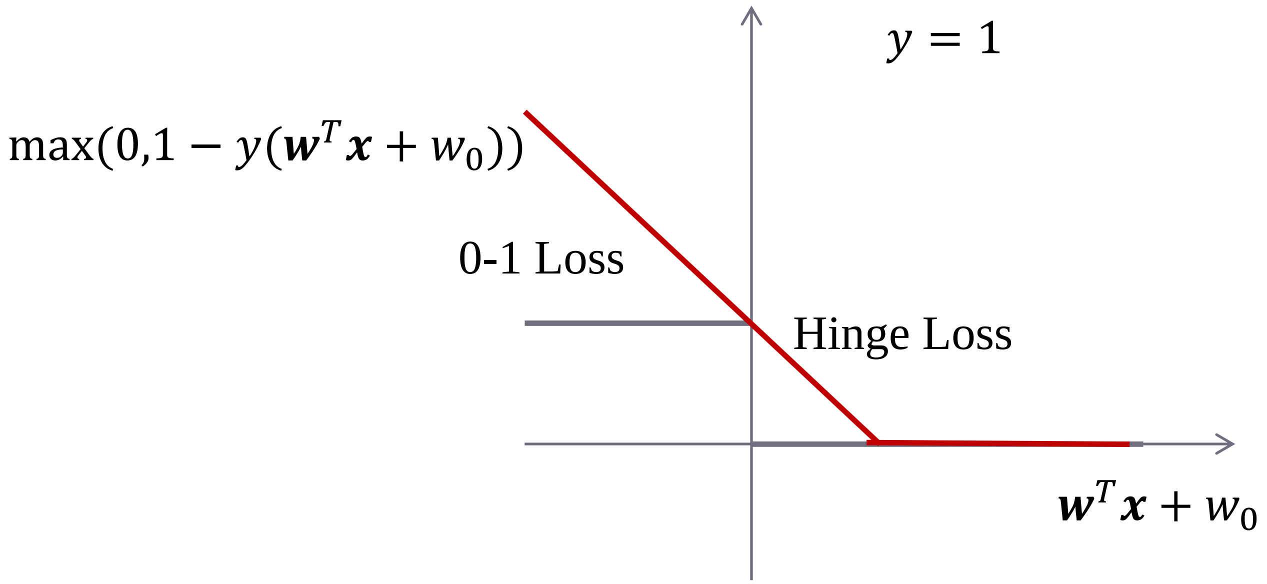
$$\xi_n \geq 0$$

- It is equivalent to the unconstrained optimization problem:

$$\min_{\boldsymbol{w}, w_0} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{n=1}^{N}\max(0, 1 - y^{(n)}(\boldsymbol{w}^T\boldsymbol{x}^{(n)} + w_0))$$

SVM loss function

- Hinge loss vs. 0-1 loss



Lagrange formulation

$$\mathcal{L}(\boldsymbol{w}, w_0, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta})$$
$$= \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{n=1}^{N}\xi_n + \sum_{n=1}^{N}\alpha_n\left(1 - \xi_n - y^{(n)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(n)} + w_0\right)\right)$$
$$- \sum_{n=1}^{N}\beta_n\xi_n$$

- Minimize w.r.t. $\boldsymbol{w}, w_0, \boldsymbol{\xi}$ and maximize w.r.t. $\alpha_n \geq 0$ and $\beta_n \geq 0$

$$\min_{\boldsymbol{w}, w_0, \{\xi_n\}_{n=1}^{N}} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{n=1}^{N}\xi_n$$
$$\text{s.t.} \quad y^{(n)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(n)} + w_0\right) \geq 1 - \xi_n \quad n = 1, \dots, N$$
$$\xi_n \geq 0$$

Lagrange formulation

$$\mathcal{L}(\boldsymbol{w}, w_0, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{n=1}^{N} \xi_n + \sum_{n=1}^{N} \alpha_n\big(1 - \xi_n$$

Sharif University of Technology

Soft-margin SVM: Dual problem

$$\max_{\boldsymbol{\alpha}} \left\{ \sum_{n=1}^{N} \alpha_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} \alpha_n \alpha_m y^{(n)} y^{(m)} \boldsymbol{x}^{(n)^T} \boldsymbol{x}^{(m)} \right\}$$

Subject to $\sum_{n=1}^{N} \alpha_n y^{(n)} = 0$

$$0 \leq \alpha_n \leq C \quad n = 1, \dots, N$$

- After solving the above quadratic problem, $\boldsymbol{w}$ is find as:

$$\boldsymbol{w} = \sum_{n=1}^{N} \alpha_n \, y^{(n)} \boldsymbol{x}^{(n)}$$

Soft-margin SVM: Support vectors

- Support Vectors: $\alpha_n > 0$
  - If $0 < \alpha_n < C$ (**margin** support vector)

SVs on the margin

$$y^{(n)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(n)} + w_0\right) = 1 \qquad (\xi_n = 0)$$

  - If $\alpha = C$ (**non-margin** support vector)

SVs on or over the margin

$$y^{(n)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(n)} + w_0\right) < 1 \qquad (\xi_n > 0)$$

$$C - \alpha_n - \beta_n = 0$$

SVM: Summary

- Hard margin: maximizing margin
- Soft margin: handling noisy data and overlapping classes
  - Slack variables in the problem
- Dual problems of hard-margin and soft-margin SVM
  - Classifier decision in terms of *support vectors*
- Dual problems lead us to non-linear SVM method easily by kernel substitution

Recourses

- C. Bishop, "Pattern Recognition and Machine Learning", Chapter 7.1.

- Yaser S. Abu-Mostafa, et al., "Learning from Data", Chapter 8.
- Course CE-717, Dr. M.Soleymani
- Course cs231n, Fei Fei Li, Stanford 2017.