

SVM and Kernels

Alexey Zaytsev
Slides by Evgeny Burnaev
Errors on slides by me

Skoltech, Moscow, Russia

- 1 Convex optimization and Duality: Basics
- 2 Support Vector Machine
- 3 SVMs with kernels
- 4 Support Vector Regression

1 Convex optimization and Duality: Basics

2 Support Vector Machine

3 SVMs with kernels

4 Support Vector Regression

- **Standard form problem** (not necessarily convex)

$$\begin{aligned} & \text{minimize}_{\mathbf{x}} \quad f_0(\mathbf{x}) \\ & \text{subject to} \quad f_i(\mathbf{x}) \leq 0, \quad i = 1, \dots, m \\ & \quad \quad \quad h_i(\mathbf{x}) = 0, \quad i = 1, \dots, p \end{aligned}$$

variable $\mathbf{x} \in X \subseteq \mathbb{R}^d$, optimal value f^*

- **Lagrangian:** $L : \mathbb{R}^d \times \mathbb{R}^m \times \mathbb{R}^p \rightarrow \mathbb{R}$, with $\text{dom}(L) = X \times \mathbb{R}^m \times \mathbb{R}^p$

$$L(\mathbf{x}, \lambda, \nu) = f_0(\mathbf{x}) + \sum_{i=1}^m \lambda_i f_i(\mathbf{x}) + \sum_{i=1}^p \nu_i h_i(\mathbf{x})$$

- weighted sum of objective and constraint functions
- λ_i is Lagrange multiplier associated with $f_i(\mathbf{x}) \leq 0$
- ν_i is Lagrange multiplier associate with $h_i(\mathbf{x}) = 0$

- **Standard form problem** (not necessarily convex)

$$\begin{aligned} & \text{minimize}_{\mathbf{x}} \quad f_0(\mathbf{x}) \\ & \text{subject to} \quad f_i(\mathbf{x}) \leq 0, \quad i = 1, \dots, m \\ & \quad \quad \quad h_i(\mathbf{x}) = 0, \quad i = 1, \dots, p \end{aligned}$$

variable $\mathbf{x} \in X \subseteq \mathbb{R}^d$, optimal value f^*

- **Lagrangian:** $L : \mathbb{R}^d \times \mathbb{R}^m \times \mathbb{R}^p \rightarrow \mathbb{R}$, with $\text{dom}(L) = X \times \mathbb{R}^m \times \mathbb{R}^p$

$$L(\mathbf{x}, \lambda, \nu) = f_0(\mathbf{x}) + \sum_{i=1}^m \lambda_i f_i(\mathbf{x}) + \sum_{i=1}^p \nu_i h_i(\mathbf{x})$$

- weighted sum of objective and constraint functions
- λ_i is Lagrange multiplier associated with $f_i(\mathbf{x}) \leq 0$
- ν_i is Lagrange multiplier associate with $h_i(\mathbf{x}) = 0$

- **Lagrange dual function** $g : \mathbb{R}^m \times \mathbb{R}^p \rightarrow \mathbb{R}$

$$\begin{aligned} g(\lambda, \nu) &= \inf_{\mathbf{x} \in X} L(\mathbf{x}, \lambda, \nu) \\ &= \inf_{\mathbf{x} \in X} \left(f_0(\mathbf{x}) + \sum_{i=1}^m \lambda_i f_i(\mathbf{x}) + \sum_{i=1}^p \nu_i h_i(\mathbf{x}) \right), \end{aligned}$$

g is concave, can be $-\infty$ for some λ, ν

- **Lower bound property:** if $\lambda \geq 0$, then $g(\lambda, \nu) \leq f^*$
- proof: if $\tilde{\mathbf{x}}$ is feasible and $\lambda \geq 0$, then

$$f_0(\tilde{\mathbf{x}}) \geq L(\tilde{\mathbf{x}}, \lambda, \nu) \geq \inf_{\mathbf{x} \in X} L(\mathbf{x}, \lambda, \nu) = g(\lambda, \nu)$$

minimizing over all feasible $\tilde{\mathbf{x}}$ gives $f^* \geq g(\lambda, \nu)$

- **Lagrange dual function** $g : \mathbb{R}^m \times \mathbb{R}^p \rightarrow \mathbb{R}$

$$\begin{aligned} g(\lambda, \nu) &= \inf_{\mathbf{x} \in X} L(\mathbf{x}, \lambda, \nu) \\ &= \inf_{\mathbf{x} \in X} \left(f_0(\mathbf{x}) + \sum_{i=1}^m \lambda_i f_i(\mathbf{x}) + \sum_{i=1}^p \nu_i h_i(\mathbf{x}) \right), \end{aligned}$$

g is concave, can be $-\infty$ for some λ, ν

- **Lower bound property:** if $\lambda \geq 0$, then $g(\lambda, \nu) \leq f^*$
- proof: if $\tilde{\mathbf{x}}$ is feasible and $\lambda \geq 0$, then

$$f_0(\tilde{\mathbf{x}}) \geq L(\tilde{\mathbf{x}}, \lambda, \nu) \geq \inf_{\mathbf{x} \in X} L(\mathbf{x}, \lambda, \nu) = g(\lambda, \nu)$$

minimizing over all feasible $\tilde{\mathbf{x}}$ gives $f^* \geq g(\lambda, \nu)$

- **Lagrange dual function** $g : \mathbb{R}^m \times \mathbb{R}^p \rightarrow \mathbb{R}$

$$\begin{aligned} g(\lambda, \nu) &= \inf_{\mathbf{x} \in X} L(\mathbf{x}, \lambda, \nu) \\ &= \inf_{\mathbf{x} \in X} \left(f_0(\mathbf{x}) + \sum_{i=1}^m \lambda_i f_i(\mathbf{x}) + \sum_{i=1}^p \nu_i h_i(\mathbf{x}) \right), \end{aligned}$$

g is concave, can be $-\infty$ for some λ, ν

- **Lower bound property:** if $\lambda \geq 0$, then $g(\lambda, \nu) \leq f^*$
- proof: if $\tilde{\mathbf{x}}$ is feasible and $\lambda \geq 0$, then

$$f_0(\tilde{\mathbf{x}}) \geq L(\tilde{\mathbf{x}}, \lambda, \nu) \geq \inf_{\mathbf{x} \in X} L(\mathbf{x}, \lambda, \nu) = g(\lambda, \nu)$$

minimizing over all feasible $\tilde{\mathbf{x}}$ gives $f^* \geq g(\lambda, \nu)$

Lower bound property: if $\lambda \geq 0$, then $g(\lambda, \nu) \leq f^*$

- finds best lower bound on f^* , obtained from Lagrange dual function
- a convex optimization problem; optimal value denoted g^*
- λ, ν are dual feasible if $\lambda \geq 0, (\lambda, \nu) \in \text{dom}(g)$

Lower bound property: if $\lambda \geq 0$, then $g(\lambda, \nu) \leq f^*$

Lagrange dual problem

$$\begin{aligned} & \text{maximize}_{\lambda, \nu} && g(\lambda, \nu) \\ & \text{subject to} && \lambda \geq 0 \end{aligned}$$

- finds best lower bound on f^* , obtained from Lagrange dual function
- a convex optimization problem; optimal value denoted g^*
- λ, ν are dual feasible if $\lambda \geq 0$, $(\lambda, \nu) \in \text{dom}(g)$

Lower bound property: if $\lambda \geq 0$, then $g(\lambda, \nu) \leq f^*$

Lagrange dual problem

$$\begin{aligned} & \text{maximize}_{\lambda, \nu} && g(\lambda, \nu) \\ & \text{subject to} && \lambda \geq 0 \end{aligned}$$

- finds best lower bound on f^* , obtained from Lagrange dual function
- a convex optimization problem; optimal value denoted g^*
- λ, ν are dual feasible if $\lambda \geq 0$, $(\lambda, \nu) \in \text{dom}(g)$

Lower bound property: if $\lambda \geq 0$, then $g(\lambda, \nu) \leq f^*$

Lagrange dual problem

$$\begin{aligned} & \text{maximize}_{\lambda, \nu} && g(\lambda, \nu) \\ & \text{subject to} && \lambda \geq 0 \end{aligned}$$

- finds best lower bound on f^* , obtained from Lagrange dual function
- a convex optimization problem; optimal value denoted g^*
- λ, ν are dual feasible if $\lambda \geq 0$, $(\lambda, \nu) \in \text{dom}(g)$

Lower bound property: if $\lambda \geq 0$, then $g(\lambda, \nu) \leq f^*$

Lagrange dual problem

$$\begin{aligned} & \text{maximize}_{\lambda, \nu} && g(\lambda, \nu) \\ & \text{subject to} && \lambda \geq 0 \end{aligned}$$

- finds best lower bound on f^* , obtained from Lagrange dual function
- a convex optimization problem; optimal value denoted g^*
- λ, ν are dual feasible if $\lambda \geq 0$, $(\lambda, \nu) \in \text{dom}(g)$

weak duality: $g^* \leq f^*$

- always holds (for convex and nonconvex problems)
- can be used to find nontrivial lower bounds for difficult problems

strong duality: $g^* = f^*$

- does not hold in general
- (usually) holds for convex problems
- conditions that guarantee strong duality in convex problems are called **constraint qualifications**

weak duality: $g^* \leq f^*$

- always holds (for convex and nonconvex problems)
- can be used to find nontrivial lower bounds for difficult problems

strong duality: $g^* = f^*$

- does not hold in general
- (usually) holds for convex problems
- conditions that guarantee strong duality in convex problems are called **constraint qualifications**

- Slater's condition (or Slater condition) is a sufficient condition for strong duality to hold for a convex optimization problem
- Strong duality holds for a convex problem

$$\begin{aligned} & \text{minimize}_{\mathbf{x}} && f_0(\mathbf{x}) \\ & \text{subject to} && f_i(\mathbf{x}) \leq 0, \quad i = 1, \dots, m \\ & && A\mathbf{x} = b \end{aligned}$$

if it is strictly feasible, i.e.,

$$\exists \mathbf{x} \in \text{int}(X) : f_i(\mathbf{x}) < 0, \quad i = 1, \dots, m, \quad A\mathbf{x} = b$$

- also guarantees that the dual optimum is attained (if $f^* > -\infty$)

- Slater's condition (or Slater condition) is a sufficient condition for strong duality to hold for a convex optimization problem
- Strong duality holds for a convex problem

$$\begin{aligned} & \text{minimize}_{\mathbf{x}} \quad f_0(\mathbf{x}) \\ & \text{subject to} \quad f_i(\mathbf{x}) \leq 0, \quad i = 1, \dots, m \\ & \quad \quad \quad A\mathbf{x} = b \end{aligned}$$

if it is strictly feasible, i.e.,

$$\exists \mathbf{x} \in \text{int}(X) : f_i(\mathbf{x}) < 0, \quad i = 1, \dots, m, \quad A\mathbf{x} = b$$

- also guarantees that the dual optimum is attained (if $f^* > -\infty$)

- Slater's condition (or Slater condition) is a sufficient condition for strong duality to hold for a convex optimization problem
- Strong duality holds for a convex problem

$$\begin{aligned} & \text{minimize}_{\mathbf{x}} \quad f_0(\mathbf{x}) \\ & \text{subject to} \quad f_i(\mathbf{x}) \leq 0, \quad i = 1, \dots, m \\ & \quad \quad \quad A\mathbf{x} = b \end{aligned}$$

if it is strictly feasible, i.e.,

$$\exists \mathbf{x} \in \text{int}(X) : f_i(\mathbf{x}) < 0, \quad i = 1, \dots, m, \quad A\mathbf{x} = b$$

- also guarantees that the dual optimum is attained (if $f^* > -\infty$)

- Slater's condition (or Slater condition) is a sufficient condition for strong duality to hold for a convex optimization problem
- Strong duality holds for a convex problem

$$\begin{aligned} & \text{minimize}_{\mathbf{x}} \quad f_0(\mathbf{x}) \\ & \text{subject to} \quad f_i(\mathbf{x}) \leq 0, \quad i = 1, \dots, m \\ & \quad \quad \quad A\mathbf{x} = b \end{aligned}$$

if it is strictly feasible, i.e.,

$$\exists \mathbf{x} \in \text{int}(X) : f_i(\mathbf{x}) < 0, \quad i = 1, \dots, m, \quad A\mathbf{x} = b$$

- also guarantees that the dual optimum is attained (if $f^* > -\infty$)

Assume strong duality holds, \mathbf{x}^* is primal optimal, (λ^*, ν^*) is dual optimal

$$\begin{aligned} f_0(\mathbf{x}^*) &= g(\lambda^*, \nu^*) = \inf_{\mathbf{x} \in X} L(\mathbf{x}, \lambda^*, \nu^*) = \\ \inf_{\mathbf{x}} &\left(f_0(\mathbf{x}) + \sum_{i=1}^m \lambda_i^* f_i(\mathbf{x}) + \sum_{i=1}^p \nu_i^* h_i(\mathbf{x}) \right) \\ &\leq f_0(\mathbf{x}^*) + \sum_{i=1}^m \lambda_i^* f_i(\mathbf{x}^*) + \sum_{i=1}^p \nu_i^* h_i(\mathbf{x}^*) \\ &\leq f_0(\mathbf{x}^*) \end{aligned}$$

hence, the two inequalities hold with equality

- \mathbf{x}^* minimizes $L(\mathbf{x}, \lambda^*, \nu^*)$
- $\lambda_i^* f_i(\mathbf{x}^*) = 0$ for $i = 1, 2, \dots, m$ (known as complementary slackness):

$$\lambda_i^* > 0 \Rightarrow f_i(\mathbf{x}^*) = 0, \quad f_i(\mathbf{x}^*) < 0 \Rightarrow \lambda_i^* = 0$$

Assume strong duality holds, \mathbf{x}^* is primal optimal, (λ^*, ν^*) is dual optimal

$$\begin{aligned} f_0(\mathbf{x}^*) &= g(\lambda^*, \nu^*) = \inf_{\mathbf{x} \in X} L(\mathbf{x}, \lambda^*, \nu^*) = \\ \inf_{\mathbf{x}} &\left(f_0(\mathbf{x}) + \sum_{i=1}^m \lambda_i^* f_i(\mathbf{x}) + \sum_{i=1}^p \nu_i^* h_i(\mathbf{x}) \right) \\ &\leq f_0(\mathbf{x}^*) + \sum_{i=1}^m \lambda_i^* f_i(\mathbf{x}^*) + \sum_{i=1}^p \nu_i^* h_i(\mathbf{x}^*) \\ &\leq f_0(\mathbf{x}^*) \end{aligned}$$

hence, the two inequalities hold with equality

- \mathbf{x}^* minimizes $L(\mathbf{x}, \lambda^*, \nu^*)$
- $\lambda_i^* f_i(\mathbf{x}^*) = 0$ for $i = 1, 2, \dots, m$ (known as complementary slackness):

$$\lambda_i^* > 0 \Rightarrow f_i(\mathbf{x}^*) = 0, \quad f_i(\mathbf{x}^*) < 0 \Rightarrow \lambda_i^* = 0$$

Assume strong duality holds, \mathbf{x}^* is primal optimal, (λ^*, ν^*) is dual optimal

$$\begin{aligned} f_0(\mathbf{x}^*) &= g(\lambda^*, \nu^*) = \inf_{\mathbf{x} \in X} L(\mathbf{x}, \lambda^*, \nu^*) = \\ \inf_{\mathbf{x}} &\left(f_0(\mathbf{x}) + \sum_{i=1}^m \lambda_i^* f_i(\mathbf{x}) + \sum_{i=1}^p \nu_i^* h_i(\mathbf{x}) \right) \\ &\leq f_0(\mathbf{x}^*) + \sum_{i=1}^m \lambda_i^* f_i(\mathbf{x}^*) + \sum_{i=1}^p \nu_i^* h_i(\mathbf{x}^*) \\ &\leq f_0(\mathbf{x}^*) \end{aligned}$$

hence, the two inequalities hold with equality

- \mathbf{x}^* minimizes $L(\mathbf{x}, \lambda^*, \nu^*)$
- $\lambda_i^* f_i(\mathbf{x}^*) = 0$ for $i = 1, 2, \dots, m$ (known as complementary slackness):

$$\lambda_i^* > 0 \Rightarrow f_i(\mathbf{x}^*) = 0, \quad f_i(\mathbf{x}^*) < 0 \Rightarrow \lambda_i^* = 0$$

Assume strong duality holds, \mathbf{x}^* is primal optimal, (λ^*, ν^*) is dual optimal

$$\begin{aligned} f_0(\mathbf{x}^*) &= g(\lambda^*, \nu^*) = \inf_{\mathbf{x} \in X} L(\mathbf{x}, \lambda^*, \nu^*) = \\ \inf_{\mathbf{x}} &\left(f_0(\mathbf{x}) + \sum_{i=1}^m \lambda_i^* f_i(\mathbf{x}) + \sum_{i=1}^p \nu_i^* h_i(\mathbf{x}) \right) \\ &\leq f_0(\mathbf{x}^*) + \sum_{i=1}^m \lambda_i^* f_i(\mathbf{x}^*) + \sum_{i=1}^p \nu_i^* h_i(\mathbf{x}^*) \\ &\leq f_0(\mathbf{x}^*) \end{aligned}$$

hence, the two inequalities hold with equality

- \mathbf{x}^* minimizes $L(\mathbf{x}, \lambda^*, \nu^*)$
- $\lambda_i^* f_i(\mathbf{x}^*) = 0$ for $i = 1, 2, \dots, m$ (known as complementary slackness):

$$\lambda_i^* > 0 \Rightarrow f_i(\mathbf{x}^*) = 0, \quad f_i(\mathbf{x}^*) < 0 \Rightarrow \lambda_i^* = 0$$

Assume strong duality holds, \mathbf{x}^* is primal optimal, (λ^*, ν^*) is dual optimal

$$\begin{aligned} f_0(\mathbf{x}^*) &= g(\lambda^*, \nu^*) = \inf_{\mathbf{x} \in X} L(\mathbf{x}, \lambda^*, \nu^*) = \\ \inf_{\mathbf{x}} &\left(f_0(\mathbf{x}) + \sum_{i=1}^m \lambda_i^* f_i(\mathbf{x}) + \sum_{i=1}^p \nu_i^* h_i(\mathbf{x}) \right) \\ &\leq f_0(\mathbf{x}^*) + \sum_{i=1}^m \lambda_i^* f_i(\mathbf{x}^*) + \sum_{i=1}^p \nu_i^* h_i(\mathbf{x}^*) \\ &\leq f_0(\mathbf{x}^*) \end{aligned}$$

hence, the two inequalities hold with equality

- \mathbf{x}^* minimizes $L(\mathbf{x}, \lambda^*, \nu^*)$
- $\lambda_i^* f_i(\mathbf{x}^*) = 0$ for $i = 1, 2, \dots, m$ (known as complementary slackness):

$$\lambda_i^* > 0 \Rightarrow f_i(\mathbf{x}^*) = 0, \quad f_i(\mathbf{x}^*) < 0 \Rightarrow \lambda_i^* = 0$$

Assume strong duality holds, \mathbf{x}^* is primal optimal, (λ^*, ν^*) is dual optimal

$$\begin{aligned} f_0(\mathbf{x}^*) &= g(\lambda^*, \nu^*) = \inf_{\mathbf{x} \in X} L(\mathbf{x}, \lambda^*, \nu^*) = \\ \inf_{\mathbf{x}} &\left(f_0(\mathbf{x}) + \sum_{i=1}^m \lambda_i^* f_i(\mathbf{x}) + \sum_{i=1}^p \nu_i^* h_i(\mathbf{x}) \right) \\ &\leq f_0(\mathbf{x}^*) + \sum_{i=1}^m \lambda_i^* f_i(\mathbf{x}^*) + \sum_{i=1}^p \nu_i^* h_i(\mathbf{x}^*) \\ &\leq f_0(\mathbf{x}^*) \end{aligned}$$

hence, the two inequalities hold with equality

- \mathbf{x}^* minimizes $L(\mathbf{x}, \lambda^*, \nu^*)$
- $\lambda_i^* f_i(\mathbf{x}^*) = 0$ for $i = 1, 2, \dots, m$ (known as complementary slackness):

$$\lambda_i^* > 0 \Rightarrow f_i(\mathbf{x}^*) = 0, \quad f_i(\mathbf{x}^*) < 0 \Rightarrow \lambda_i^* = 0$$

The following four conditions for \mathbf{x}, λ, ν are called KKT conditions (for a problem with differentiable f_i, h_i)

1. Primal constraints: $f_i(\mathbf{x}) \leq 0, i = 1, 2, \dots, m, h_i(\mathbf{x}) = 0, i = 1, 2, \dots, p$
2. Dual constraints: $\lambda \geq 0$
3. Complementary slackness: $\lambda_i f_i(\mathbf{x}) = 0, i = 1, 2, \dots, m$
4. Gradient of Lagrangian with respect to \mathbf{x} vanishes:

$$\nabla f_0(\mathbf{x}) + \sum_{i=1}^m \lambda_i \nabla f_i(\mathbf{x}) + \sum_{i=1}^p \nu_i \nabla h_i(\mathbf{x}) = 0$$

- If strong duality holds and \mathbf{x}, λ, ν are optimal, then they must satisfy the KKT conditions

The following four conditions for \mathbf{x}, λ, ν are called KKT conditions (for a problem with differentiable f_i, h_i)

1. Primal constraints: $f_i(\mathbf{x}) \leq 0, i = 1, 2, \dots, m, h_i(\mathbf{x}) = 0, i = 1, 2, \dots, p$
2. Dual constraints: $\lambda \geq 0$
3. Complementary slackness: $\lambda_i f_i(\mathbf{x}) = 0, i = 1, 2, \dots, m$
4. Gradient of Lagrangian with respect to \mathbf{x} vanishes:

$$\nabla f_0(\mathbf{x}) + \sum_{i=1}^m \lambda_i \nabla f_i(\mathbf{x}) + \sum_{i=1}^p \nu_i \nabla h_i(\mathbf{x}) = 0$$

- If strong duality holds and \mathbf{x}, λ, ν are optimal, then they must satisfy the KKT conditions

The following four conditions for \mathbf{x}, λ, ν are called KKT conditions (for a problem with differentiable f_i, h_i)

1. Primal constraints: $f_i(\mathbf{x}) \leq 0, i = 1, 2, \dots, m, h_i(\mathbf{x}) = 0, i = 1, 2, \dots, p$
2. Dual constraints: $\lambda \geq 0$
3. Complementary slackness: $\lambda_i f_i(\mathbf{x}) = 0, i = 1, 2, \dots, m$
4. Gradient of Lagrangian with respect to \mathbf{x} vanishes:

$$\nabla f_0(\mathbf{x}) + \sum_{i=1}^m \lambda_i \nabla f_i(\mathbf{x}) + \sum_{i=1}^p \nu_i \nabla h_i(\mathbf{x}) = 0$$

- If strong duality holds and \mathbf{x}, λ, ν are optimal, then they must satisfy the KKT conditions

The following four conditions for \mathbf{x}, λ, ν are called KKT conditions (for a problem with differentiable f_i, h_i)

1. Primal constraints: $f_i(\mathbf{x}) \leq 0, i = 1, 2, \dots, m, h_i(\mathbf{x}) = 0, i = 1, 2, \dots, p$
2. Dual constraints: $\lambda \geq 0$
3. Complementary slackness: $\lambda_i f_i(\mathbf{x}) = 0, i = 1, 2, \dots, m$
4. Gradient of Lagrangian with respect to \mathbf{x} vanishes:

$$\nabla f_0(\mathbf{x}) + \sum_{i=1}^m \lambda_i \nabla f_i(\mathbf{x}) + \sum_{i=1}^p \nu_i \nabla h_i(\mathbf{x}) = 0$$

- If strong duality holds and \mathbf{x}, λ, ν are optimal, then they must satisfy the KKT conditions

The following four conditions for \mathbf{x}, λ, ν are called KKT conditions (for a problem with differentiable f_i, h_i)

1. Primal constraints: $f_i(\mathbf{x}) \leq 0, i = 1, 2, \dots, m, h_i(\mathbf{x}) = 0, i = 1, 2, \dots, p$
2. Dual constraints: $\lambda \geq 0$
3. Complementary slackness: $\lambda_i f_i(\mathbf{x}) = 0, i = 1, 2, \dots, m$
4. Gradient of Lagrangian with respect to \mathbf{x} vanishes:

$$\nabla f_0(\mathbf{x}) + \sum_{i=1}^m \lambda_i \nabla f_i(\mathbf{x}) + \sum_{i=1}^p \nu_i \nabla h_i(\mathbf{x}) = 0$$

- If strong duality holds and \mathbf{x}, λ, ν are optimal, then they must satisfy the KKT conditions

If $\tilde{\mathbf{x}}, \tilde{\lambda}, \tilde{\nu}$ satisfy KKT for a convex problem, then they are optimal:

- from slackness: $f_0(\tilde{\mathbf{x}}) = L(\tilde{\mathbf{x}}, \tilde{\lambda}, \tilde{\nu})$
- from 4th condition (and convexity): $g(\tilde{\lambda}, \tilde{\nu}) = L(\tilde{\mathbf{x}}, \tilde{\lambda}, \tilde{\nu})$

hence, $f_0(\tilde{\mathbf{x}}) = g(\tilde{\lambda}, \tilde{\nu})$

If **Slater's condition** is satisfied:

- \mathbf{x} is optimal if and only if there exist λ, ν that satisfy KKT conditions
- recall that Slater implies strong duality, and dual optimum is attained
- generalizes optimality condition $\nabla f_0(\mathbf{x}) = 0$ for unconstrained problem

1 Convex optimization and Duality: Basics

2 Support Vector Machine

3 SVMs with kernels

4 Support Vector Regression

- **Training data:** sample drawn i.i.d. w.r.t. D on $X \subseteq \mathbb{R}^d$

$$S_m = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m)\} \in \{X \times \{-1, +1\}\}^m$$

- **Problem:** find hypothesis $h : X \rightarrow \{-1, +1\}$ in H (classifier) with small generalization error $R(h)$
- First we consider linear classification (hyperplanes) if dimension d is not too large

- **Training data:** sample drawn i.i.d. w.r.t. D on $X \subseteq \mathbb{R}^d$

$$S_m = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m)\} \in \{X \times \{-1, +1\}\}^m$$

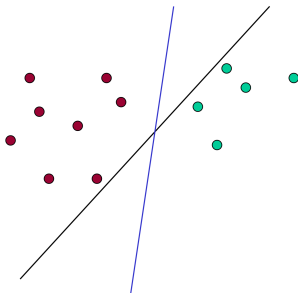
- **Problem:** find hypothesis $h : X \rightarrow \{-1, +1\}$ in H (classifier) with small generalization error $R(h)$
- First we consider linear classification (hyperplanes) if dimension d is not too large

- **Training data:** sample drawn i.i.d. w.r.t. D on $X \subseteq \mathbb{R}^d$

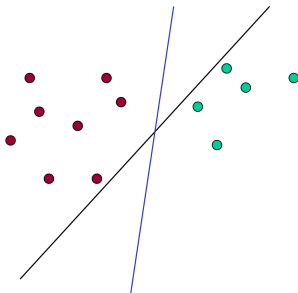
$$S_m = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m)\} \in \{X \times \{-1, +1\}\}^m$$

- **Problem:** find hypothesis $h : X \rightarrow \{-1, +1\}$ in H (classifier) with small generalization error $R(h)$
- First we consider linear classification (hyperplanes) if dimension d is not too large

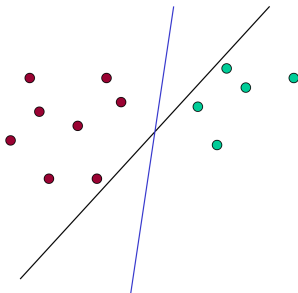
- Support vectors are the data points that lie closest to the decision surface (or hyperplane)
- Support vectors are the elements of the training set that would change the position of the dividing hyperplane if removed
- They are the data points most difficult to classify
- In general, lots of possible solutions for a hyperplane



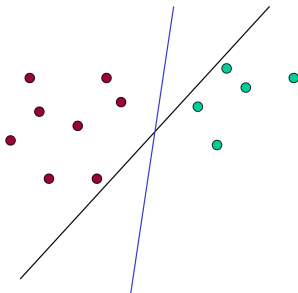
- Support vectors are the data points that lie closest to the decision surface (or hyperplane)
- Support vectors are the elements of the training set that would change the position of the dividing hyperplane if removed
- They are the data points most difficult to classify
- In general, lots of possible solutions for a hyperplane

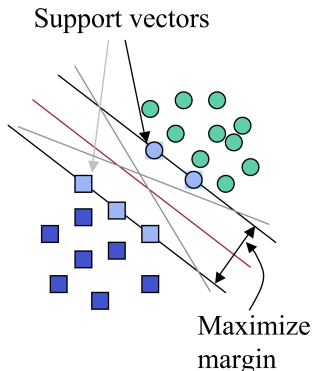


- Support vectors are the data points that lie closest to the decision surface (or hyperplane)
- Support vectors are the elements of the training set that would change the position of the dividing hyperplane if removed
- They are the data points most difficult to classify
- In general, lots of possible solutions for a hyperplane

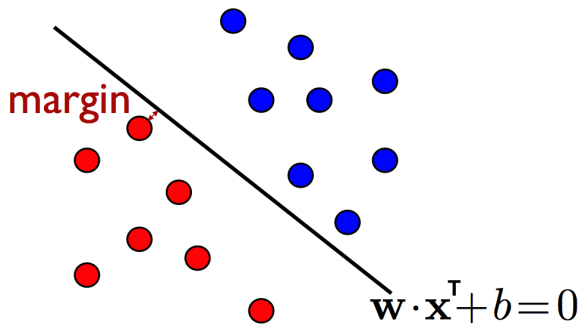


- Support vectors are the data points that lie closest to the decision surface (or hyperplane)
- Support vectors are the elements of the training set that would change the position of the dividing hyperplane if removed
- They are the data points most difficult to classify
- In general, lots of possible solutions for a hyperplane

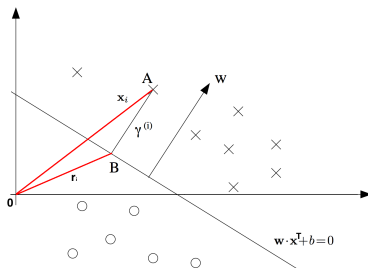




- Support Vector Machine (SVM) finds an optimal solution
- SVMs maximize the margin (the “street”) around the separating hyperplane
- The decision function is fully specified by a (usually very small) subset of training samples, the support vectors



- classifiers: $H = \{\mathbf{x} \rightarrow \text{sgn}(\mathbf{w} \cdot \mathbf{x}^T + b), \mathbf{w} \in \mathbb{R}^d, b \in \mathbb{R}\}$

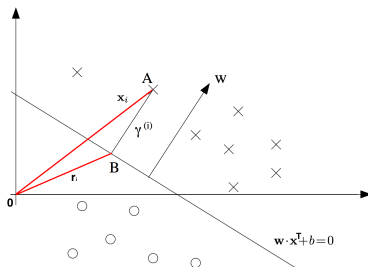


- $\gamma^{(i)}$ is a distance from \mathbf{x}_i to the hyperplane $\mathbf{w} \cdot \mathbf{x}^T + b = 0$
- $\mathbf{w}/\|\mathbf{w}\|$ is a unit perpendicular to the hyperplane
- Vector \mathbf{r}_i of a point B is equal to

$$\mathbf{r}_i = \mathbf{x}_i - \gamma^{(i)} \mathbf{w}/\|\mathbf{w}\|$$

- Since point B belongs to the hyperplane: $\mathbf{w} \cdot \mathbf{r}_i^T + b = 0$, i.e.

- Thus we get that $\gamma^{(i)} = \frac{\mathbf{w}}{\|\mathbf{w}\|} \mathbf{x}_i^T + \frac{b}{\|\mathbf{w}\|}$

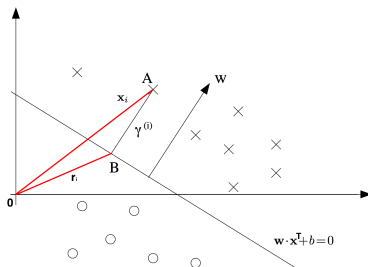


- $\gamma^{(i)}$ is a distance from \mathbf{x}_i to the hyperplane $\mathbf{w} \cdot \mathbf{x}^\top + b = 0$
- $\mathbf{w}/\|\mathbf{w}\|$ is a unit perpendicular to the hyperplane
- Vector \mathbf{r}_i of a point B is equal to

$$\mathbf{r}_i = \mathbf{x}_i - \gamma^{(i)} \mathbf{w}/\|\mathbf{w}\|$$

- Since point B belongs to the hyperplane: $\mathbf{w} \cdot \mathbf{r}_i^\top + b = 0$, i.e.

- Thus we get that $\gamma^{(i)} = \frac{\mathbf{w}}{\|\mathbf{w}\|} \mathbf{x}_i^\top + \frac{b}{\|\mathbf{w}\|}$

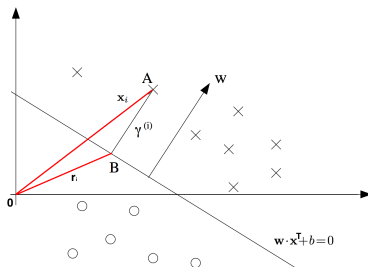


- $\gamma^{(i)}$ is a distance from \mathbf{x}_i to the hyperplane $\mathbf{w} \cdot \mathbf{x}^T + b = 0$
- $\mathbf{w}/\|\mathbf{w}\|$ is a unit perpendicular to the hyperplane
- Vector \mathbf{r}_i of a point B is equal to

$$\mathbf{r}_i = \mathbf{x}_i - \gamma^{(i)} \mathbf{w} / \|\mathbf{w}\|$$

- Since point B belongs to the hyperplane: $\mathbf{w} \cdot \mathbf{r}_i^T + b = 0$, i.e.

- Thus we get that $\gamma^{(i)} = \frac{\mathbf{w}}{\|\mathbf{w}\|} \mathbf{x}_i^T + \frac{b}{\|\mathbf{w}\|}$

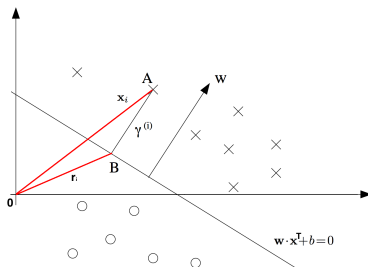


- $\gamma^{(i)}$ is a distance from \mathbf{x}_i to the hyperplane $\mathbf{w} \cdot \mathbf{x}^T + b = 0$
- $\mathbf{w}/\|\mathbf{w}\|$ is a unit perpendicular to the hyperplane
- Vector \mathbf{r}_i of a point B is equal to

$$\mathbf{r}_i = \mathbf{x}_i - \gamma^{(i)} \mathbf{w} / \|\mathbf{w}\|$$

- Since point B belongs to the hyperplane: $\mathbf{w} \cdot \mathbf{r}_i^T + b = 0$, i.e.

• Thus we get that $\gamma^{(i)} = \frac{\mathbf{w}}{\|\mathbf{w}\|} \mathbf{x}_i^T + \frac{b}{\|\mathbf{w}\|}$



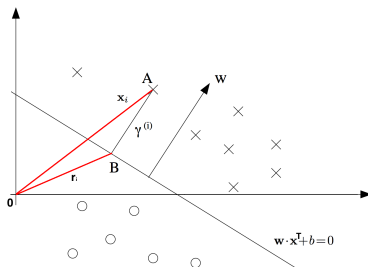
- $\gamma^{(i)}$ is a distance from \mathbf{x}_i to the hyperplane $\mathbf{w} \cdot \mathbf{x}^\top + b = 0$
- $\mathbf{w}/\|\mathbf{w}\|$ is a unit perpendicular to the hyperplane
- Vector \mathbf{r}_i of a point B is equal to

$$\mathbf{r}_i = \mathbf{x}_i - \gamma^{(i)} \mathbf{w} / \|\mathbf{w}\|$$

- Since point B belongs to the hyperplane: $\mathbf{w} \cdot \mathbf{r}_i^\top + b = 0$, i.e.

$$\mathbf{w} \left(\mathbf{x}_i^\top - \gamma^{(i)} \frac{\mathbf{w}^\top}{\|\mathbf{w}\|} \right) + b = 0$$

- Thus we get that $\gamma^{(i)} = \frac{\mathbf{w}}{\|\mathbf{w}\|} \mathbf{x}_i^\top + \frac{b}{\|\mathbf{w}\|}$



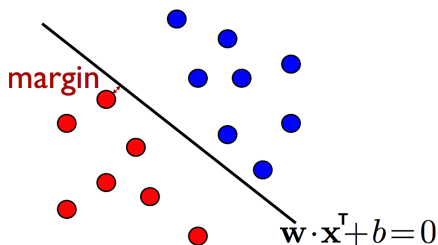
- $\gamma^{(i)}$ is a distance from \mathbf{x}_i to the hyperplane $\mathbf{w} \cdot \mathbf{x}^\top + b = 0$
- $\mathbf{w}/\|\mathbf{w}\|$ is a unit perpendicular to the hyperplane
- Vector \mathbf{r}_i of a point B is equal to

$$\mathbf{r}_i = \mathbf{x}_i - \gamma^{(i)} \mathbf{w} / \|\mathbf{w}\|$$

- Since point B belongs to the hyperplane: $\mathbf{w} \cdot \mathbf{r}_i^\top + b = 0$, i.e.

$$\mathbf{w} \left(\mathbf{x}_i^\top - \gamma^{(i)} \frac{\mathbf{w}^\top}{\|\mathbf{w}\|} \right) + b = 0$$

- Thus we get that $\gamma^{(i)} = \frac{\mathbf{w}}{\|\mathbf{w}\|} \mathbf{x}_i^\top + \frac{b}{\|\mathbf{w}\|}$

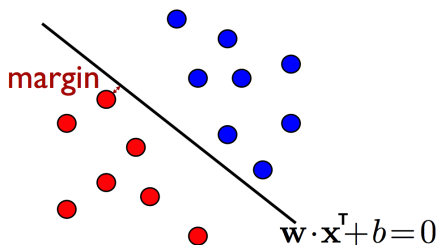


- In general case

$$\gamma^{(i)} = \left| \frac{\mathbf{w}}{\|\mathbf{w}\|} \mathbf{x}_i^T + \frac{b}{\|\mathbf{w}\|} \right| =$$

- The margin is

$$\rho = \max_{\mathbf{w}, b: y_i(\mathbf{w} \cdot \mathbf{x}_i^T + b) \geq 0} \left[\min_{i \in [1, m]} \gamma^{(i)} \right]$$

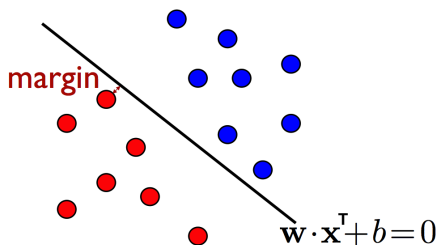


- In general case

$$\gamma^{(i)} = \left| \frac{\mathbf{w}}{\|\mathbf{w}\|} \mathbf{x}_i^\top + \frac{b}{\|\mathbf{w}\|} \right| = \frac{|\mathbf{w} \cdot \mathbf{x}_i^\top + b|}{\|\mathbf{w}\|}$$

- The margin is

$$\rho = \max_{\mathbf{w}, b: y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) \geq 0} \left[\min_{i \in [1, m]} \gamma^{(i)} \right]$$

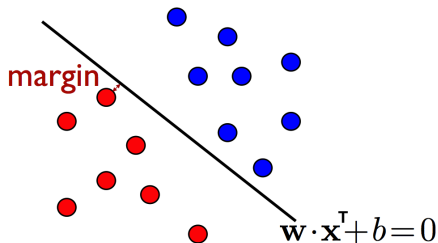


- In general case

$$\gamma^{(i)} = \left| \frac{\mathbf{w}}{\|\mathbf{w}\|} \mathbf{x}_i^\top + \frac{b}{\|\mathbf{w}\|} \right| = \frac{|\mathbf{w} \cdot \mathbf{x}_i^\top + b|}{\|\mathbf{w}\|} \rightarrow \min_{i \in [1, m]} \text{ (worst case!)}$$

- The margin is

$$\rho = \max_{\mathbf{w}, b: y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) \geq 0} \left[\min_{i \in [1, m]} \gamma^{(i)} \right]$$



- In general case

$$\gamma^{(i)} = \left| \frac{\mathbf{w}}{\|\mathbf{w}\|} \mathbf{x}_i^\top + \frac{b}{\|\mathbf{w}\|} \right| = \frac{|\mathbf{w} \cdot \mathbf{x}_i^\top + b|}{\|\mathbf{w}\|} \rightarrow \min_{i \in [1, m]} \text{ (worst case!)}$$

- The margin is

$$\rho = \max_{\mathbf{w}, b: y_i (\mathbf{w} \cdot \mathbf{x}_i^\top + b) \geq 0} \left[\min_{i \in [1, m]} \gamma^{(i)} \right]$$

- Optimization problem

$$\rho = \max_{\mathbf{w}, b: y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) \geq 0} \left[\min_{i \in [1, m]} \frac{|\mathbf{w} \cdot \mathbf{x}_i^\top + b|}{\|\mathbf{w}\|} \right]$$

- Target $\min_{i \in [1, m]} \frac{|\mathbf{w} \cdot \mathbf{x}_i^\top + b|}{\|\mathbf{w}\|}$ is scale-invariant, i.e.

- Inequality $y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) \geq 0 \Leftrightarrow y_i(\tilde{\mathbf{w}} \cdot \mathbf{x}_i^\top + \tilde{b}) \geq 0$
- We can always re-normalize \mathbf{w} and b such that
 - it holds: $\min_{i \in [1, m]} |\mathbf{w} \cdot \mathbf{x}_i^\top + b| = 1$
 - inequality $y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) \geq 0$ does not change
 - target function $\min_{i \in [1, m]} \frac{|\mathbf{w} \cdot \mathbf{x}_i^\top + b|}{\|\mathbf{w}\|}$ does not change

- Optimization problem

$$\rho = \max_{\mathbf{w}, b: y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) \geq 0} \left[\min_{i \in [1, m]} \frac{|\mathbf{w} \cdot \mathbf{x}_i^\top + b|}{\|\mathbf{w}\|} \right]$$

- Target $\min_{i \in [1, m]} \frac{|\mathbf{w} \cdot \mathbf{x}_i^\top + b|}{\|\mathbf{w}\|}$ is scale-invariant, i.e.

- Inequality $y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) \geq 0 \Leftrightarrow y_i(\tilde{\mathbf{w}} \cdot \mathbf{x}_i^\top + \tilde{b}) \geq 0$
- We can always re-normalize \mathbf{w} and b such that
 - it holds: $\min_{i \in [1, m]} |\mathbf{w} \cdot \mathbf{x}_i^\top + b| = 1$
 - inequality $y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) \geq 0$ does not change
 - target function $\min_{i \in [1, m]} \frac{|\mathbf{w} \cdot \mathbf{x}_i^\top + b|}{\|\mathbf{w}\|}$ does not change

- Optimization problem

$$\rho = \max_{\mathbf{w}, b: y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) \geq 0} \left[\min_{i \in [1, m]} \frac{|\mathbf{w} \cdot \mathbf{x}_i^\top + b|}{\|\mathbf{w}\|} \right]$$

- Target $\min_{i \in [1, m]} \frac{|\mathbf{w} \cdot \mathbf{x}_i^\top + b|}{\|\mathbf{w}\|}$ is scale-invariant, i.e.

$$\min_{i \in [1, m]} \frac{|\mathbf{w} \cdot \mathbf{x}_i^\top + b|}{\|\mathbf{w}\|} = \frac{\min_{i \in [1, m]} |\mathbf{w} \cdot \mathbf{x}_i^\top + b| \cdot \text{const}}{\|\mathbf{w}\| \cdot \text{const}}$$

- Inequality $y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) \geq 0 \Leftrightarrow y_i(\tilde{\mathbf{w}} \cdot \mathbf{x}_i^\top + \tilde{b}) \geq 0$
- We can always re-normalize \mathbf{w} and b such that
 - it holds: $\min_{i \in [1, m]} |\mathbf{w} \cdot \mathbf{x}_i^\top + b| = 1$
 - inequality $y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) \geq 0$ does not change
 - target function $\min_{i \in [1, m]} \frac{|\mathbf{w} \cdot \mathbf{x}_i^\top + b|}{\|\mathbf{w}\|}$ does not change

- Optimization problem

$$\rho = \max_{\mathbf{w}, b: y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) \geq 0} \left[\min_{i \in [1, m]} \frac{|\mathbf{w} \cdot \mathbf{x}_i^\top + b|}{\|\mathbf{w}\|} \right]$$

- Target $\min_{i \in [1, m]} \frac{|\mathbf{w} \cdot \mathbf{x}_i^\top + b|}{\|\mathbf{w}\|}$ is scale-invariant, i.e.

$$\begin{aligned} \min_{i \in [1, m]} \frac{|\mathbf{w} \cdot \mathbf{x}_i^\top + b|}{\|\mathbf{w}\|} &= \frac{\min_{i \in [1, m]} |\mathbf{w} \cdot \mathbf{x}_i^\top + b| \cdot \text{const}}{\|\mathbf{w}\| \cdot \text{const}} = \\ &= \frac{\min_{i \in [1, m]} |\tilde{\mathbf{w}} \cdot \mathbf{x}_i^\top + \tilde{b}|}{\|\tilde{\mathbf{w}}\|} \quad (\tilde{\mathbf{w}} = \mathbf{w} \cdot \text{const}, \tilde{b} = b \cdot \text{const}) \end{aligned}$$

- Inequality $y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) \geq 0 \Leftrightarrow y_i(\tilde{\mathbf{w}} \cdot \mathbf{x}_i^\top + \tilde{b}) \geq 0$
- We can always re-normalize \mathbf{w} and b such that

- it holds: $\min_{i \in [1, m]} |\mathbf{w} \cdot \mathbf{x}_i^\top + b| = 1$
- inequality $y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) \geq 0$ does not change
- target function $\min_{i \in [1, m]} \frac{|\mathbf{w} \cdot \mathbf{x}_i^\top + b|}{\|\mathbf{w}\|}$ does not change

- Optimization problem

$$\rho = \max_{\mathbf{w}, b: y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) \geq 0} \left[\min_{i \in [1, m]} \frac{|\mathbf{w} \cdot \mathbf{x}_i^\top + b|}{\|\mathbf{w}\|} \right]$$

- Target $\min_{i \in [1, m]} \frac{|\mathbf{w} \cdot \mathbf{x}_i^\top + b|}{\|\mathbf{w}\|}$ is scale-invariant, i.e.

$$\begin{aligned} \min_{i \in [1, m]} \frac{|\mathbf{w} \cdot \mathbf{x}_i^\top + b|}{\|\mathbf{w}\|} &= \frac{\min_{i \in [1, m]} |\mathbf{w} \cdot \mathbf{x}_i^\top + b| \cdot \text{const}}{\|\mathbf{w}\| \cdot \text{const}} = \\ &= \frac{\min_{i \in [1, m]} |\tilde{\mathbf{w}} \cdot \mathbf{x}_i^\top + \tilde{b}|}{\|\tilde{\mathbf{w}}\|} \quad (\tilde{\mathbf{w}} = \mathbf{w} \cdot \text{const}, \tilde{b} = b \cdot \text{const}) \end{aligned}$$

- Inequality $y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) \geq 0 \Leftrightarrow y_i(\tilde{\mathbf{w}} \cdot \mathbf{x}_i^\top + \tilde{b}) \geq 0$
- We can always re-normalize \mathbf{w} and b such that

- it holds: $\min_{i \in [1, m]} |\mathbf{w} \cdot \mathbf{x}_i^\top + b| = 1$
- inequality $y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) \geq 0$ does not change
- target function $\min_{i \in [1, m]} \frac{|\mathbf{w} \cdot \mathbf{x}_i^\top + b|}{\|\mathbf{w}\|}$ does not change

- Optimization problem

$$\rho = \max_{\mathbf{w}, b: y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) \geq 0} \left[\min_{i \in [1, m]} \frac{|\mathbf{w} \cdot \mathbf{x}_i^\top + b|}{\|\mathbf{w}\|} \right]$$

- Target $\min_{i \in [1, m]} \frac{|\mathbf{w} \cdot \mathbf{x}_i^\top + b|}{\|\mathbf{w}\|}$ is scale-invariant, i.e.

$$\begin{aligned} \min_{i \in [1, m]} \frac{|\mathbf{w} \cdot \mathbf{x}_i^\top + b|}{\|\mathbf{w}\|} &= \frac{\min_{i \in [1, m]} |\mathbf{w} \cdot \mathbf{x}_i^\top + b| \cdot \text{const}}{\|\mathbf{w}\| \cdot \text{const}} = \\ &= \frac{\min_{i \in [1, m]} |\tilde{\mathbf{w}} \cdot \mathbf{x}_i^\top + \tilde{b}|}{\|\tilde{\mathbf{w}}\|} \quad (\tilde{\mathbf{w}} = \mathbf{w} \cdot \text{const}, \tilde{b} = b \cdot \text{const}) \end{aligned}$$

- Inequality $y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) \geq 0 \Leftrightarrow y_i(\tilde{\mathbf{w}} \cdot \mathbf{x}_i^\top + \tilde{b}) \geq 0$
- We can always re-normalize \mathbf{w} and b such that
 - it holds: $\min_{i \in [1, m]} |\mathbf{w} \cdot \mathbf{x}_i^\top + b| = 1$
 - inequality $y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) \geq 0$ does not change
 - target function $\min_{i \in [1, m]} \frac{|\mathbf{w} \cdot \mathbf{x}_i^\top + b|}{\|\mathbf{w}\|}$ does not change

- Optimization problem

$$\rho = \max_{\mathbf{w}, b: y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) \geq 0} \left[\min_{i \in [1, m]} \frac{|\mathbf{w} \cdot \mathbf{x}_i^\top + b|}{\|\mathbf{w}\|} \right]$$

- Target $\min_{i \in [1, m]} \frac{|\mathbf{w} \cdot \mathbf{x}_i^\top + b|}{\|\mathbf{w}\|}$ is scale-invariant, i.e.

$$\begin{aligned} \min_{i \in [1, m]} \frac{|\mathbf{w} \cdot \mathbf{x}_i^\top + b|}{\|\mathbf{w}\|} &= \frac{\min_{i \in [1, m]} |\mathbf{w} \cdot \mathbf{x}_i^\top + b| \cdot \text{const}}{\|\mathbf{w}\| \cdot \text{const}} = \\ &= \frac{\min_{i \in [1, m]} |\tilde{\mathbf{w}} \cdot \mathbf{x}_i^\top + \tilde{b}|}{\|\tilde{\mathbf{w}}\|} \quad (\tilde{\mathbf{w}} = \mathbf{w} \cdot \text{const}, \tilde{b} = b \cdot \text{const}) \end{aligned}$$

- Inequality $y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) \geq 0 \Leftrightarrow y_i(\tilde{\mathbf{w}} \cdot \mathbf{x}_i^\top + \tilde{b}) \geq 0$
- We can always re-normalize \mathbf{w} and b such that
 - it holds: $\min_{i \in [1, m]} |\mathbf{w} \cdot \mathbf{x}_i^\top + b| = 1$
 - inequality $y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) \geq 0$ does not change
 - target function $\min_{i \in [1, m]} \frac{|\mathbf{w} \cdot \mathbf{x}_i^\top + b|}{\|\mathbf{w}\|}$ does not change

- Optimization problem

$$\rho = \max_{\mathbf{w}, b: y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) \geq 0} \left[\min_{i \in [1, m]} \frac{|\mathbf{w} \cdot \mathbf{x}_i^\top + b|}{\|\mathbf{w}\|} \right]$$

- Target $\min_{i \in [1, m]} \frac{|\mathbf{w} \cdot \mathbf{x}_i^\top + b|}{\|\mathbf{w}\|}$ is scale-invariant, i.e.

$$\begin{aligned} \min_{i \in [1, m]} \frac{|\mathbf{w} \cdot \mathbf{x}_i^\top + b|}{\|\mathbf{w}\|} &= \frac{\min_{i \in [1, m]} |\mathbf{w} \cdot \mathbf{x}_i^\top + b| \cdot \text{const}}{\|\mathbf{w}\| \cdot \text{const}} = \\ &= \frac{\min_{i \in [1, m]} |\tilde{\mathbf{w}} \cdot \mathbf{x}_i^\top + \tilde{b}|}{\|\tilde{\mathbf{w}}\|} \quad (\tilde{\mathbf{w}} = \mathbf{w} \cdot \text{const}, \tilde{b} = b \cdot \text{const}) \end{aligned}$$

- Inequality $y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) \geq 0 \Leftrightarrow y_i(\tilde{\mathbf{w}} \cdot \mathbf{x}_i^\top + \tilde{b}) \geq 0$
- We can always re-normalize \mathbf{w} and b such that
 - it holds: $\min_{i \in [1, m]} |\mathbf{w} \cdot \mathbf{x}_i^\top + b| = 1$
 - inequality $y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) \geq 0$ does not change
 - target function $\min_{i \in [1, m]} \frac{|\mathbf{w} \cdot \mathbf{x}_i^\top + b|}{\|\mathbf{w}\|}$ does not change

- Optimization problem

$$\rho = \max_{\mathbf{w}, b: y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) \geq 0} \left[\min_{i \in [1, m]} \frac{|\mathbf{w} \cdot \mathbf{x}_i^\top + b|}{\|\mathbf{w}\|} \right]$$

- Target $\min_{i \in [1, m]} \frac{|\mathbf{w} \cdot \mathbf{x}_i^\top + b|}{\|\mathbf{w}\|}$ is scale-invariant, i.e.

$$\begin{aligned} \min_{i \in [1, m]} \frac{|\mathbf{w} \cdot \mathbf{x}_i^\top + b|}{\|\mathbf{w}\|} &= \frac{\min_{i \in [1, m]} |\mathbf{w} \cdot \mathbf{x}_i^\top + b| \cdot \text{const}}{\|\mathbf{w}\| \cdot \text{const}} = \\ &= \frac{\min_{i \in [1, m]} |\tilde{\mathbf{w}} \cdot \mathbf{x}_i^\top + \tilde{b}|}{\|\tilde{\mathbf{w}}\|} \quad (\tilde{\mathbf{w}} = \mathbf{w} \cdot \text{const}, \tilde{b} = b \cdot \text{const}) \end{aligned}$$

- Inequality $y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) \geq 0 \Leftrightarrow y_i(\tilde{\mathbf{w}} \cdot \mathbf{x}_i^\top + \tilde{b}) \geq 0$
- We can always re-normalize \mathbf{w} and b such that
 - it holds: $\min_{i \in [1, m]} |\mathbf{w} \cdot \mathbf{x}_i^\top + b| = 1$
 - inequality $y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) \geq 0$ does not change
 - target function $\min_{i \in [1, m]} \frac{|\mathbf{w} \cdot \mathbf{x}_i^\top + b|}{\|\mathbf{w}\|}$ does not change

$$\rho = \max_{\mathbf{w}, b: y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) \geq 0} \left[\min_{i \in [1, m]} \frac{|\mathbf{w} \cdot \mathbf{x}_i^\top + b|}{\|\mathbf{w}\|} \right]$$

$$= \max_{\substack{\mathbf{w}, b: y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) \geq 0 \\ \min_{i \in [1, m]} |\mathbf{w} \cdot \mathbf{x}_i^\top + b| = 1}} \left[\min_{i \in [1, m]} \frac{|\mathbf{w} \cdot \mathbf{x}_i^\top + b|}{\|\mathbf{w}\|} \right] \quad (\text{scale-invar.})$$

$$= \max_{\substack{\mathbf{w}, b: y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) \geq 0 \\ \min_{i \in [1, m]} |\mathbf{w} \cdot \mathbf{x}_i^\top + b| = 1}} \left[\frac{1}{\|\mathbf{w}\|} \right]$$

$$= \max_{\mathbf{w}, b: y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) \geq 1} \left[\frac{1}{\|\mathbf{w}\|} \right]$$

$$\begin{aligned}\rho &= \max_{\mathbf{w}, b: y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) \geq 0} \left[\min_{i \in [1, m]} \frac{|\mathbf{w} \cdot \mathbf{x}_i^\top + b|}{\|\mathbf{w}\|} \right] \\&= \max_{\substack{\mathbf{w}, b: y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) \geq 0 \\ \min_{i \in [1, m]} |\mathbf{w} \cdot \mathbf{x}_i^\top + b| = 1}} \left[\min_{i \in [1, m]} \frac{|\mathbf{w} \cdot \mathbf{x}_i^\top + b|}{\|\mathbf{w}\|} \right] \quad (\text{scale-invar.}) \\&= \max_{\substack{\mathbf{w}, b: y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) \geq 0 \\ \min_{i \in [1, m]} |\mathbf{w} \cdot \mathbf{x}_i^\top + b| = 1}} \left[\frac{1}{\|\mathbf{w}\|} \right] \\&= \max_{\mathbf{w}, b: y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) \geq 1} \left[\frac{1}{\|\mathbf{w}\|} \right]\end{aligned}$$

$$\begin{aligned}\rho &= \max_{\mathbf{w}, b: y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) \geq 0} \left[\min_{i \in [1, m]} \frac{|\mathbf{w} \cdot \mathbf{x}_i^\top + b|}{\|\mathbf{w}\|} \right] \\&= \max_{\substack{\mathbf{w}, b: y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) \geq 0 \\ \min_{i \in [1, m]} |\mathbf{w} \cdot \mathbf{x}_i^\top + b| = 1}} \left[\min_{i \in [1, m]} \frac{|\mathbf{w} \cdot \mathbf{x}_i^\top + b|}{\|\mathbf{w}\|} \right] \quad (\text{scale-invar.}) \\&= \max_{\substack{\mathbf{w}, b: y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) \geq 0 \\ \min_{i \in [1, m]} |\mathbf{w} \cdot \mathbf{x}_i^\top + b| = 1}} \left[\frac{1}{\|\mathbf{w}\|} \right] \\&= \max_{\mathbf{w}, b: y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) \geq 1} \left[\frac{1}{\|\mathbf{w}\|} \right]\end{aligned}$$

$$\begin{aligned}\rho &= \max_{\mathbf{w}, b: y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) \geq 0} \left[\min_{i \in [1, m]} \frac{|\mathbf{w} \cdot \mathbf{x}_i^\top + b|}{\|\mathbf{w}\|} \right] \\&= \max_{\substack{\mathbf{w}, b: y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) \geq 0 \\ \min_{i \in [1, m]} |\mathbf{w} \cdot \mathbf{x}_i^\top + b| = 1}} \left[\min_{i \in [1, m]} \frac{|\mathbf{w} \cdot \mathbf{x}_i^\top + b|}{\|\mathbf{w}\|} \right] \quad (\text{scale-invar.}) \\&= \max_{\substack{\mathbf{w}, b: y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) \geq 0 \\ \min_{i \in [1, m]} |\mathbf{w} \cdot \mathbf{x}_i^\top + b| = 1}} \left[\frac{1}{\|\mathbf{w}\|} \right] \\&= \max_{\mathbf{w}, b: y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) \geq 1} \left[\frac{1}{\|\mathbf{w}\|} \right]\end{aligned}$$

- Optimization problem

$$\max_{\mathbf{w}, b: y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) \geq 1} \left[\frac{1}{\|\mathbf{w}\|} \right]$$

- Constrained Optimization:

$$\min_{\mathbf{w}, b} \frac{1}{2} \|\mathbf{w}\|^2$$

$$\text{s.t. } y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) \geq 1, i \in [1, m]$$

- Properties:

- Convex optimization
- Unique solution for linearly separable case

- Optimization problem

$$\max_{\mathbf{w}, b: y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) \geq 1} \left[\frac{1}{\|\mathbf{w}\|} \right]$$

- **Constrained Optimization:**

$$\min_{\mathbf{w}, b} \frac{1}{2} \|\mathbf{w}\|^2$$

$$\text{s.t. } y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) \geq 1, i \in [1, m]$$

- **Properties:**

- Convex optimization
- Unique solution for linearly separable case

- Optimization problem

$$\max_{\mathbf{w}, b: y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) \geq 1} \left[\frac{1}{\|\mathbf{w}\|} \right]$$

- **Constrained Optimization:**

$$\min_{\mathbf{w}, b} \frac{1}{2} \|\mathbf{w}\|^2$$

$$\text{s.t. } y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) \geq 1, i \in [1, m]$$

- **Properties:**

- Convex optimization
- Unique solution for linearly separable case

- **Lagrangian:** for all $\mathbf{w}, b, \alpha_i \geq 0$

$$L(\mathbf{w}, b, \alpha) = \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{i=1}^m \alpha_i [y_i (\mathbf{w} \cdot \mathbf{x}_i^\top + b) - 1]$$

- **KKT conditions:**

$$\nabla_{\mathbf{w}} L = \mathbf{w} - \sum_{i=1}^m \alpha_i y_i \mathbf{x}_i^\top = 0 \Leftrightarrow \mathbf{w} = \sum_{i=1}^m \alpha_i y_i \mathbf{x}_i$$

- **Lagrangian:** for all $\mathbf{w}, b, \alpha_i \geq 0$

$$L(\mathbf{w}, b, \alpha) = \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{i=1}^m \alpha_i [y_i (\mathbf{w} \cdot \mathbf{x}_i^\top + b) - 1]$$

- **KKT conditions:**

$$\nabla_{\mathbf{w}} L = \mathbf{w} - \sum_{i=1}^m \alpha_i y_i \mathbf{x}_i^\top = 0 \Leftrightarrow \mathbf{w} = \sum_{i=1}^m \alpha_i y_i \mathbf{x}_i$$

- **Lagrangian:** for all $\mathbf{w}, b, \alpha_i \geq 0$

$$L(\mathbf{w}, b, \alpha) = \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{i=1}^m \alpha_i [y_i (\mathbf{w} \cdot \mathbf{x}_i^\top + b) - 1]$$

- **KKT conditions:**

$$\nabla_{\mathbf{w}} L = \mathbf{w} - \sum_{i=1}^m \alpha_i y_i \mathbf{x}_i^\top = 0 \Leftrightarrow \mathbf{w} = \sum_{i=1}^m \alpha_i y_i \mathbf{x}_i$$

$$\nabla_b L = - \sum_{i=1}^m \alpha_i y_i = 0 \Leftrightarrow \sum_{i=1}^m \alpha_i y_i = 0$$

- **Lagrangian:** for all $\mathbf{w}, b, \alpha_i \geq 0$

$$L(\mathbf{w}, b, \alpha) = \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{i=1}^m \alpha_i [y_i (\mathbf{w} \cdot \mathbf{x}_i^\top + b) - 1]$$

- **KKT conditions:**

$$\nabla_{\mathbf{w}} L = \mathbf{w} - \sum_{i=1}^m \alpha_i y_i \mathbf{x}_i^\top = 0 \Leftrightarrow \mathbf{w} = \sum_{i=1}^m \alpha_i y_i \mathbf{x}_i$$

$$\nabla_b L = - \sum_{i=1}^m \alpha_i y_i = 0 \Leftrightarrow \sum_{i=1}^m \alpha_i y_i = 0$$

$$\forall i \in [1, m], \alpha_i [y_i (\mathbf{w} \cdot \mathbf{x}_i^\top + b) - 1] = 0$$

- **Complementary conditions:**

$$\alpha_i [y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) - 1] = 0 \Rightarrow \alpha_i = 0 \text{ or } y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) = 1$$

- **Support vectors:** vectors \mathbf{x}_i such that

$$\alpha_i \neq 0 \text{ and } y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) = 1$$

- Support Vectors: Input vectors that just touch the boundary of the margin (street)

- **Complementary conditions:**

$$\alpha_i [y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) - 1] = 0 \Rightarrow \alpha_i = 0 \text{ or } y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) = 1$$

- **Support vectors:** vectors \mathbf{x}_i such that

$$\alpha_i \neq 0 \text{ and } y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) = 1$$

- Support Vectors: Input vectors that just touch the boundary of the margin (street)

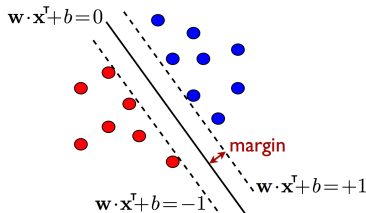
- **Complementary conditions:**

$$\alpha_i [y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) - 1] = 0 \Rightarrow \alpha_i = 0 \text{ or } y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) = 1$$

- **Support vectors:** vectors \mathbf{x}_i such that

$$\alpha_i \neq 0 \text{ and } y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) = 1$$

- **Support Vectors:** Input vectors that just touch the boundary of the margin (street)



- From KKT we get that optimal

$$\mathbf{w} = \sum_{i=1}^m \alpha_i y_i \mathbf{x}_i$$

- Plugging \mathbf{w} in $L = \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{i=1}^m \alpha_i [y_i (\mathbf{w} \cdot \mathbf{x}_i^\top + b) - 1]$ we get

$$\begin{aligned} L = & \underbrace{\frac{1}{2} \left\| \sum_{i=1}^m \alpha_i y_i \mathbf{x}_i \right\|^2 - \sum_{i,j=1}^m \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j^\top)}_{-\frac{1}{2} \sum_{i,j=1}^m \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j^\top)} \\ & - \underbrace{\sum_{i=1}^m \alpha_i y_i b}_{=0} + \sum_{i=1}^m \alpha_i \end{aligned}$$

- Thus

$$L = \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j^\top)$$

- From KKT we get that optimal

$$\mathbf{w} = \sum_{i=1}^m \alpha_i y_i \mathbf{x}_i$$

- Plugging \mathbf{w} in $L = \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{i=1}^m \alpha_i [y_i (\mathbf{w} \cdot \mathbf{x}_i^\top + b) - 1]$ we get

$$\begin{aligned} L &= \frac{1}{2} \left\| \sum_{i=1}^m \alpha_i y_i \mathbf{x}_i \right\|^2 - \sum_{i,j=1}^m \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j^\top) \\ &\quad \underbrace{- \frac{1}{2} \sum_{i,j=1}^m \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j^\top)}_{=0} \\ &\quad - \underbrace{\sum_{i=1}^m \alpha_i y_i b}_{=0} + \sum_{i=1}^m \alpha_i \end{aligned}$$

- Thus

$$L = \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j^\top)$$

- From KKT we get that optimal

$$\mathbf{w} = \sum_{i=1}^m \alpha_i y_i \mathbf{x}_i$$

- Plugging \mathbf{w} in $L = \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{i=1}^m \alpha_i [y_i (\mathbf{w} \cdot \mathbf{x}_i^\top + b) - 1]$ we get

$$\begin{aligned} L = & \underbrace{\frac{1}{2} \left\| \sum_{i=1}^m \alpha_i y_i \mathbf{x}_i \right\|^2 - \sum_{i,j=1}^m \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j^\top)}_{-\frac{1}{2} \sum_{i,j=1}^m \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j^\top)} \\ & - \underbrace{\sum_{i=1}^m \alpha_i y_i b}_{=0} + \sum_{i=1}^m \alpha_i \end{aligned}$$

- Thus

$$L = \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j^\top)$$

- From KKT we get that optimal

$$\mathbf{w} = \sum_{i=1}^m \alpha_i y_i \mathbf{x}_i$$

- Plugging \mathbf{w} in $L = \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{i=1}^m \alpha_i [y_i (\mathbf{w} \cdot \mathbf{x}_i^\top + b) - 1]$ we get

$$\begin{aligned} L &= \frac{1}{2} \left\| \sum_{i=1}^m \alpha_i y_i \mathbf{x}_i \right\|^2 - \underbrace{\sum_{i,j=1}^m \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j^\top)}_{-\frac{1}{2} \sum_{i,j=1}^m \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j^\top)} \\ &\quad - \underbrace{\sum_{i=1}^m \alpha_i y_i b}_{=0} + \sum_{i=1}^m \alpha_i \end{aligned}$$

- Thus

$$L = \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j^\top)$$

- **Constrained Optimization:**

$$\begin{aligned} \max_{\alpha} \quad & \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j}^m \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j^{\top}) \\ \text{s.t.} \quad & \alpha_i \geq 0 \text{ and } \sum_{i=1}^m \alpha_i y_i = 0, i \in [1, m] \end{aligned}$$

- Optimal

$$\mathbf{w} = \sum_{i=1}^m \alpha_i y_i \mathbf{x}_i$$

- **Solution:** classifier based on the separating hyperplane $\mathbf{w} \cdot \mathbf{x}^{\top} + b = 0$ has the form $h(\mathbf{x}) = \text{sign}(\mathbf{w} \cdot \mathbf{x}^{\top} + b)$, i.e.

- **Constrained Optimization:**

$$\begin{aligned} \max_{\alpha} \quad & \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j}^m \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j^{\top}) \\ \text{s.t.} \quad & \alpha_i \geq 0 \text{ and } \sum_{i=1}^m \alpha_i y_i = 0, i \in [1, m] \end{aligned}$$

- **Optimal**

$$\mathbf{w} = \sum_{i=1}^m \alpha_i y_i \mathbf{x}_i$$

- **Solution:** classifier based on the separating hyperplane $\mathbf{w} \cdot \mathbf{x}^{\top} + b = 0$ has the form $h(\mathbf{x}) = \text{sign}(\mathbf{w} \cdot \mathbf{x}^{\top} + b)$, i.e.

- **Constrained Optimization:**

$$\begin{aligned} \max_{\alpha} \quad & \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j}^m \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j^{\top}) \\ \text{s.t.} \quad & \alpha_i \geq 0 \text{ and } \sum_{i=1}^m \alpha_i y_i = 0, i \in [1, m] \end{aligned}$$

- Optimal

$$\mathbf{w} = \sum_{i=1}^m \alpha_i y_i \mathbf{x}_i$$

- **Solution:** classifier based on the separating hyperplane $\mathbf{w} \cdot \mathbf{x}^{\top} + b = 0$ has the form $h(\mathbf{x}) = \text{sign}(\mathbf{w} \cdot \mathbf{x}^{\top} + b)$, i.e.

- **Constrained Optimization:**

$$\begin{aligned} \max_{\alpha} \quad & \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j}^m \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j^{\top}) \\ \text{s.t.} \quad & \alpha_i \geq 0 \text{ and } \sum_{i=1}^m \alpha_i y_i = 0, i \in [1, m] \end{aligned}$$

- Optimal

$$\mathbf{w} = \sum_{i=1}^m \alpha_i y_i \mathbf{x}_i$$

- **Solution:** classifier based on the separating hyperplane $\mathbf{w} \cdot \mathbf{x}^{\top} + b = 0$ has the form $h(\mathbf{x}) = \text{sign}(\mathbf{w} \cdot \mathbf{x}^{\top} + b)$, i.e.

$$h(\mathbf{x}) = \text{sign} \left(\sum_{i=1}^m \alpha_i y_i (\mathbf{x}_i \cdot \mathbf{x}^{\top}) + b \right),$$

- **Constrained Optimization:**

$$\begin{aligned} \max_{\alpha} \quad & \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j}^m \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j^{\top}) \\ \text{s.t.} \quad & \alpha_i \geq 0 \text{ and } \sum_{i=1}^m \alpha_i y_i = 0, i \in [1, m] \end{aligned}$$

- Optimal

$$\mathbf{w} = \sum_{i=1}^m \alpha_i y_i \mathbf{x}_i$$

- **Solution:** classifier based on the separating hyperplane $\mathbf{w} \cdot \mathbf{x}^{\top} + b = 0$ has the form $h(\mathbf{x}) = \text{sign}(\mathbf{w} \cdot \mathbf{x}^{\top} + b)$, i.e.

$$h(\mathbf{x}) = \text{sign} \left(\sum_{i=1}^m \alpha_i y_i (\mathbf{x}_i \cdot \mathbf{x}^{\top}) + b \right),$$

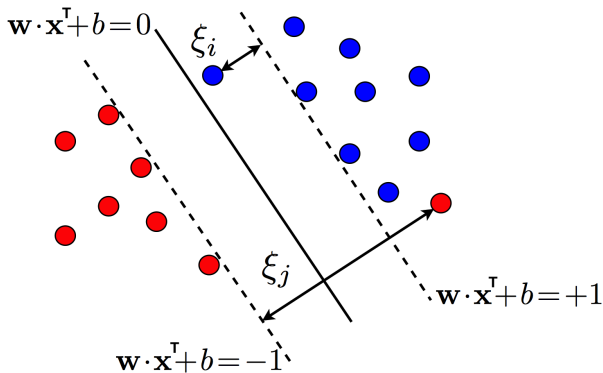
with $b = y_i - \sum_{j=1}^m \alpha_j y_j (\mathbf{x}_j \cdot \mathbf{x}_i^{\top})$ for any SV \mathbf{x}_i

- **Problem:** data often not linearly separable in practice. For any hyperplane there exists \mathbf{x}_i , such that

$$y_i[\mathbf{w} \cdot \mathbf{x}_i^\top + b] \not\geq 1$$

- **Approach:** relax constraints using slack variables $\xi_i \geq 0$

$$y_i[\mathbf{w} \cdot \mathbf{x}_i^\top + b] \geq 1 - \xi_i$$



- **Support vectors:** points along the margin or outliers
- **Soft margin:** $\rho = \frac{1}{\|\mathbf{w}\|}$

- **Constrained Optimization:**

$$\min_{\mathbf{w}, b, \xi} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^m \xi_i$$

$$\text{s.t. } y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) \geq 1 - \xi_i \text{ and } \xi_i \geq 0, i \in [1, m]$$

- **Properties:**

- Convex optimization
- Unique solution
- $C \geq 0$ is a trade-off parameter

- How to determine C ?
- The problem of determining a hyperplane minimizing the train error is NP-complete (as a function of dimension)
- Other convex functions of the slack variables can be used

- **Lagrangian:** for all $\mathbf{w}, b, \alpha_i \geq 0, \beta_i \geq 0$

$$L(\mathbf{w}, b, \xi, \alpha, \beta) = \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^m \xi_i - \sum_{i=1}^m \alpha_i [y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) - 1 + \xi_i] - \sum_{i=1}^m \beta_i \xi_i$$

- **KKT conditions:**

$$\nabla_{\mathbf{w}} L = \mathbf{w} - \sum_{i=1}^m \alpha_i y_i \mathbf{x}_i = 0 \Leftrightarrow \mathbf{w} = \sum_{i=1}^m \alpha_i y_i \mathbf{x}_i$$

$$\nabla_b L = - \sum_{i=1}^m \alpha_i y_i = 0 \Leftrightarrow \sum_{i=1}^m \alpha_i y_i = 0$$

$$\nabla_{\xi_i} L = C - \alpha_i - \beta_i = 0 \Leftrightarrow \alpha_i + \beta_i = C$$

$$\forall i \in [1, m], \alpha_i [y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) - 1 + \xi_i] = 0 \text{ and } \beta_i \xi_i = 0$$

- **Lagrangian:** for all $\mathbf{w}, b, \alpha_i \geq 0, \beta_i \geq 0$

$$L(\mathbf{w}, b, \xi, \alpha, \beta) = \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^m \xi_i - \sum_{i=1}^m \alpha_i [y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) - 1 + \xi_i] - \sum_{i=1}^m \beta_i \xi_i$$

- **KKT conditions:**

$$\nabla_{\mathbf{w}} L = \mathbf{w} - \sum_{i=1}^m \alpha_i y_i \mathbf{x}_i = 0 \Leftrightarrow \mathbf{w} = \sum_{i=1}^m \alpha_i y_i \mathbf{x}_i$$

$$\nabla_b L = - \sum_{i=1}^m \alpha_i y_i = 0 \Leftrightarrow \sum_{i=1}^m \alpha_i y_i = 0$$

$$\nabla_{\xi_i} L = C - \alpha_i - \beta_i = 0 \Leftrightarrow \alpha_i + \beta_i = C$$

$$\forall i \in [1, m], \alpha_i [y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) - 1 + \xi_i] = 0 \text{ and } \beta_i \xi_i = 0$$

- **Complementary conditions:**

$$\alpha_i[y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) - 1 + \xi_i] = 0 \Rightarrow \alpha_i = 0 \text{ or } y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) = 1 - \xi_i$$

- **Support vectors:** vectors \mathbf{x}_i such that

$$\alpha_i \neq 0 \text{ and } y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) = 1 - \xi_i$$

- **Complementary conditions:**

$$\alpha_i[y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) - 1 + \xi_i] = 0 \Rightarrow \alpha_i = 0 \text{ or } y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) = 1 - \xi_i$$

- **Support vectors:** vectors \mathbf{x}_i such that

$$\alpha_i \neq 0 \text{ and } y_i(\mathbf{w} \cdot \mathbf{x}_i^\top + b) = 1 - \xi_i$$

- Plugging optimal $\mathbf{w} = \sum_{i=1}^m \alpha_i y_i \mathbf{x}_i$ in L we get

$$L = \underbrace{\frac{1}{2} \left\| \sum_{i=1}^m \alpha_i y_i \mathbf{x}_i \right\|^2 - \sum_{i,j=1}^m \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j^\top)}_{-\frac{1}{2} \sum_{i,j=1}^m \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j^\top)} - \underbrace{\sum_{i=1}^m \alpha_i y_i b}_{=0} + \sum_{i=1}^m \alpha_i$$

- Thus

$$L = \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j^\top)$$

- Since $\beta_i = C - \alpha_i$, the condition $\beta_i \geq 0$ is equivalent to $\alpha_i \leq C$

- Plugging optimal $\mathbf{w} = \sum_{i=1}^m \alpha_i y_i \mathbf{x}_i$ in L we get

$$\begin{aligned} L = & \underbrace{\frac{1}{2} \left\| \sum_{i=1}^m \alpha_i y_i \mathbf{x}_i \right\|^2 - \sum_{i,j=1}^m \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j^\top)}_{-\frac{1}{2} \sum_{i,j=1}^m \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j^\top)} \\ & - \underbrace{\sum_{i=1}^m \alpha_i y_i b}_{=0} + \sum_{i=1}^m \alpha_i \end{aligned}$$

- Thus

$$L = \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j^\top)$$

- Since $\beta_i = C - \alpha_i$, the condition $\beta_i \geq 0$ is equivalent to $\alpha_i \leq C$

- Plugging optimal $\mathbf{w} = \sum_{i=1}^m \alpha_i y_i \mathbf{x}_i$ in L we get

$$\begin{aligned} L = & \underbrace{\frac{1}{2} \left\| \sum_{i=1}^m \alpha_i y_i \mathbf{x}_i \right\|^2 - \sum_{i,j=1}^m \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j^\top)}_{-\frac{1}{2} \sum_{i,j=1}^m \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j^\top)} \\ & - \underbrace{\sum_{i=1}^m \alpha_i y_i b}_{=0} + \sum_{i=1}^m \alpha_i \end{aligned}$$

- Thus

$$L = \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j^\top)$$

- Since $\beta_i = C - \alpha_i$, the condition $\beta_i \geq 0$ is equivalent to $\alpha_i \leq C$

- **Constrained Optimization:**

$$\max_{\alpha} \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j}^m \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j^{\top})$$

$$\text{s.t. } 0 \leq \alpha_i \leq C \text{ and } \sum_{i=1}^m \alpha_i y_i = 0, i \in [1, m]$$

- **Solution:** separating hyperplane $\mathbf{w} \cdot \mathbf{x}^{\top} + b = 0$

$$h(\mathbf{x}) = \text{sgn} \left(\sum_{i=1}^m \alpha_i y_i (\mathbf{x}_i \cdot \mathbf{x}^{\top}) + b \right),$$

with $b = y_i - \sum_{j=1}^m \alpha_j y_j (\mathbf{x}_j \cdot \mathbf{x}_i^{\top})$ for any SV \mathbf{x}_i with $0 < \alpha_i < C$

- **Constrained Optimization:**

$$\max_{\alpha} \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j}^m \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j^{\top})$$

$$\text{s.t. } 0 \leq \alpha_i \leq C \text{ and } \sum_{i=1}^m \alpha_i y_i = 0, i \in [1, m]$$

- **Solution:** separating hyperplane $\mathbf{w} \cdot \mathbf{x}^{\top} + b = 0$

$$h(\mathbf{x}) = \text{sgn} \left(\sum_{i=1}^m \alpha_i y_i (\mathbf{x}_i \cdot \mathbf{x}^{\top}) + b \right),$$

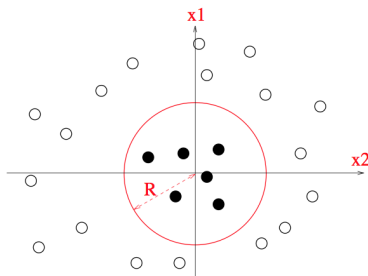
with $b = y_i - \sum_{j=1}^m \alpha_j y_j (\mathbf{x}_j \cdot \mathbf{x}_i^{\top})$ for any SV \mathbf{x}_i with $0 < \alpha_i < C$

1 Convex optimization and Duality: Basics

2 Support Vector Machine

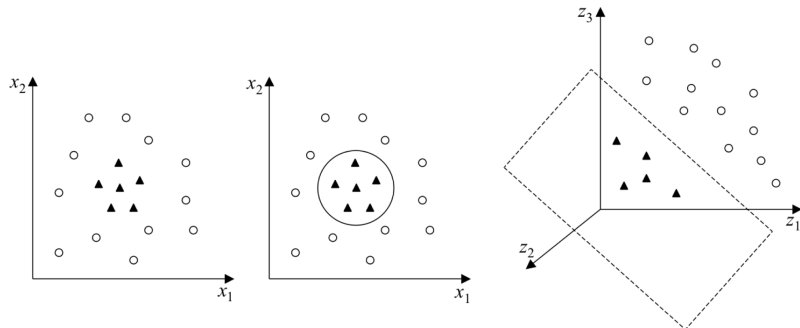
3 SVMs with kernels

4 Support Vector Regression



- Linear separation impossible in most problems
- Non-linear mapping $\Phi : X \rightarrow \mathbb{H}$ from input space to high-dimensional feature space
- Generalization ability: independent of $\dim(\mathbb{H})$, depends only on d and m

Example: polynomial kernel



For $\mathbf{x} = (x_1, x_2) \in \mathbb{R}^2$, let $\Phi(\mathbf{x}) = (x_1^2, \sqrt{2}x_1x_2, x_2^2) \in \mathbb{R}^3$. Then

$$\begin{aligned} K(\mathbf{x}', \mathbf{x}) &= \Phi(\mathbf{x}') \cdot \Phi(\mathbf{x})^\top \quad [\text{dot product of features}] \\ &= x_1^2(x_1')^2 + 2x_1x_2x_1'x_2' + x_2^2(x_2')^2 \\ &= (x_1x_1' + x_2x_2')^2 = (\mathbf{x}' \cdot \mathbf{x})^2 \end{aligned}$$

- **Idea:**

- Define $K : X \times X \rightarrow \mathbb{R}$ called kernel, such that

$$\Phi(\mathbf{x}) \cdot \Phi(\mathbf{x}')^\top = K(\mathbf{x}, \mathbf{x}')$$

- K is often interpreted as a similarity measure

- **Benefits:**

- Efficiency: K is often more efficient to compute than Φ and the dot product
- Flexibility: K can be chosen arbitrarily so long as the existence of Φ is guaranteed (PDS condition or Mercer's condition)

- **Idea:**

- Define $K : X \times X \rightarrow \mathbb{R}$ called kernel, such that

$$\Phi(\mathbf{x}) \cdot \Phi(\mathbf{x}')^\top = K(\mathbf{x}, \mathbf{x}')$$

- K is often interpreted as a similarity measure

- **Benefits:**

- Efficiency: K is often more efficient to compute than Φ and the dot product
- Flexibility: K can be chosen arbitrarily so long as the existence of Φ is guaranteed (PDS condition or Mercer's condition)

- **Gaussian kernel:**

$$K(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\sigma^2}\right), \sigma \neq 0$$

- **Constrained Dual Optimization** problem:

$$\max_{\alpha} \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j^{\top})$$

$$\text{s.t. } 0 \leq \alpha_i \leq C \text{ and } \sum_{i=1}^m \alpha_i y_i = 0, i \in [1, m]$$

- **Decision function** $h(\mathbf{x}) = \text{sgn}(\mathbf{w} \cdot \mathbf{x}^{\top} + b)$ has the form

$$h(\mathbf{x}) = \text{sgn} \left(\sum_{i=1}^m \alpha_i y_i (\mathbf{x}_i \cdot \mathbf{x}^{\top}) + b \right),$$

with

$$\mathbf{w} = \sum_{i=1}^m \alpha_i y_i \mathbf{x}_i, \quad b = y_i - \sum_{j=1}^m \alpha_j y_j (\mathbf{x}_i \cdot \mathbf{x}_j^{\top})$$

for any SV \mathbf{x}_i with $0 < \alpha_i < C$

- **Constrained Dual Optimization** problem:

$$\max_{\alpha} \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j}^m \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j^{\top})$$

$$\text{s.t. } 0 \leq \alpha_i \leq C \text{ and } \sum_{i=1}^m \alpha_i y_i = 0, i \in [1, m]$$

- **Decision function** $h(\mathbf{x}) = \text{sgn}(\mathbf{w} \cdot \mathbf{x}^{\top} + b)$ has the form

$$h(\mathbf{x}) = \text{sgn} \left(\sum_{i=1}^m \alpha_i y_i (\mathbf{x}_i \cdot \mathbf{x}^{\top}) + b \right),$$

with

$$\mathbf{w} = \sum_{i=1}^m \alpha_i y_i \mathbf{x}_i, \quad b = y_i - \sum_{j=1}^m \alpha_j y_j (\mathbf{x}_i \cdot \mathbf{x}_j^{\top})$$

for any SV \mathbf{x}_i with $0 < \alpha_i < C$

- **Constrained Dual Optimization** problem:

$$\max_{\alpha} \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \underbrace{\Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)}_{K(\mathbf{x}_i, \mathbf{x}_j)}^{\top}$$

$$\text{s.t. } 0 \leq \alpha_i \leq C \text{ and } \sum_{i=1}^m \alpha_i y_i = 0, i \in [1, m]$$

- **Decision function** $h(\mathbf{x}) = \text{sgn}(\mathbf{w} \cdot \Phi(\mathbf{x})^{\top} + b)$ has the form

$$h(\mathbf{x}) = \text{sgn} \left(\sum_{i=1}^m \alpha_i y_i \underbrace{\Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x})^{\top}}_{K(\mathbf{x}_i, \mathbf{x})} + b \right),$$

with

$$\mathbf{w} = \sum_{i=1}^m \alpha_i y_i \Phi(\mathbf{x}_i), \quad b = y_i - \sum_{j=1}^m \alpha_j y_j \underbrace{\Phi(\mathbf{x}_j) \cdot \Phi(\mathbf{x}_i)^{\top}}_{K(\mathbf{x}_j, \mathbf{x}_i)}$$

for any SV \mathbf{x}_i with $0 < \alpha_i < C$

- **Constrained Dual Optimization** problem:

$$\max_{\alpha} \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \underbrace{\Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)}_{K(\mathbf{x}_i, \mathbf{x}_j)}^{\top}$$

$$\text{s.t. } 0 \leq \alpha_i \leq C \text{ and } \sum_{i=1}^m \alpha_i y_i = 0, i \in [1, m]$$

- **Decision function** $h(\mathbf{x}) = \text{sgn}(\mathbf{w} \cdot \Phi(\mathbf{x})^{\top} + b)$ has the form

$$h(\mathbf{x}) = \text{sgn} \left(\sum_{i=1}^m \alpha_i y_i \underbrace{\Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x})^{\top}}_{K(\mathbf{x}_i, \mathbf{x})} + b \right),$$

with

$$\mathbf{w} = \sum_{i=1}^m \alpha_i y_i \Phi(\mathbf{x}_i), \quad b = y_i - \sum_{j=1}^m \alpha_j y_j \underbrace{\Phi(\mathbf{x}_j) \cdot \Phi(\mathbf{x}_i)^{\top}}_{K(\mathbf{x}_j, \mathbf{x}_i)}$$

for any SV \mathbf{x}_i with $0 < \alpha_i < C$

- **Constrained Dual Optimization** problem:

$$\max_{\alpha} \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j)$$

$$\text{s.t. } 0 \leq \alpha_i \leq C \text{ and } \sum_{i=1}^m \alpha_i y_i = 0, i \in [1, m]$$

- **Decision function** $h(\mathbf{x}) = \text{sgn}(\mathbf{w} \cdot \Phi(\mathbf{x})^\top + b)$ has the form

$$h(\mathbf{x}) = \text{sgn} \left(\sum_{i=1}^m \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b \right),$$

with

$$\mathbf{w} = \sum_{i=1}^m \alpha_i y_i \Phi(\mathbf{x}_i), \quad b = y_i - \sum_{j=1}^m \alpha_j y_j K(\mathbf{x}_j, \mathbf{x}_i)$$

for any SV \mathbf{x}_i with $0 < \alpha_i < C$

- **Constrained Dual Optimization** problem:

$$\max_{\alpha} \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j}^m \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j)$$

$$\text{s.t. } 0 \leq \alpha_i \leq C \text{ and } \sum_{i=1}^m \alpha_i y_i = 0, i \in [1, m]$$

- **Decision function** $h(\mathbf{x}) = \text{sgn}(\mathbf{w} \cdot \Phi(\mathbf{x})^\top + b)$ has the form

$$h(\mathbf{x}) = \text{sgn} \left(\sum_{i=1}^m \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b \right),$$

with

$$\mathbf{w} = \sum_{i=1}^m \alpha_i y_i \Phi(\mathbf{x}_i), \quad b = y_i - \sum_{j=1}^m \alpha_j y_j K(\mathbf{x}_j, \mathbf{x}_i)$$

for any SV \mathbf{x}_i with $0 < \alpha_i < C$

1 Convex optimization and Duality: Basics

2 Support Vector Machine

3 SVMs with kernels

4 Support Vector Regression

- Hypothesis set

$$\{x \rightarrow \mathbf{w} \cdot \Phi(\mathbf{x})^\top + b : \mathbf{w} \in \mathbb{R}^p, b \in \mathbb{R}\}$$

- Loss function: ϵ -insensitive loss

$$L(y, y') = |y - y'|_\epsilon = \max(0, |y' - y| - \epsilon)$$

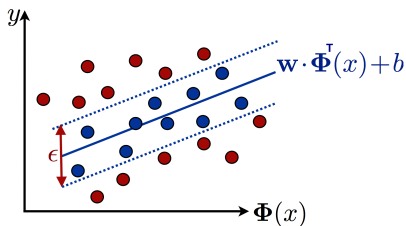


Figure – Fit “tube” with width ϵ to data

- **Optimization problem:** similar to that of SVM

$$\frac{1}{2}\|\mathbf{w}\|^2 + C \sum_{i=1}^m |y_i - (\mathbf{w} \cdot \Phi(\mathbf{x}_i)^\top + b)|_\epsilon \rightarrow \min_{\mathbf{w}, b}$$

- Equivalent formulation

$$\min_{\mathbf{w}, b, \xi, \xi'} \frac{1}{2}\|\mathbf{w}\|^2 + C \sum_{i=1}^m (\xi_i + \xi'_i)$$

$$\text{subject to } (\mathbf{w} \cdot \Phi(\mathbf{x}_i)^\top + b) - y_i \leq \epsilon + \xi_i$$

$$y_i - (\mathbf{w} \cdot \Phi(\mathbf{x}_i)^\top + b) \leq \epsilon + \xi'_i$$

$$\xi_i \geq 0, \xi'_i \geq 0$$

- **Optimization problem:** similar to that of SVM

$$\frac{1}{2}\|\mathbf{w}\|^2 + C \sum_{i=1}^m |y_i - (\mathbf{w} \cdot \Phi(\mathbf{x}_i)^\top + b)|_\epsilon \rightarrow \min_{\mathbf{w}, b}$$

- Equivalent formulation

$$\min_{\mathbf{w}, b, \xi, \xi'} \frac{1}{2}\|\mathbf{w}\|^2 + C \sum_{i=1}^m (\xi_i + \xi'_i)$$

$$\text{subject to } (\mathbf{w} \cdot \Phi(\mathbf{x}_i)^\top + b) - y_i \leq \epsilon + \xi_i$$

$$y_i - (\mathbf{w} \cdot \Phi(\mathbf{x}_i)^\top + b) \leq \epsilon + \xi'_i$$

$$\xi_i \geq 0, \xi'_i \geq 0$$

- **Optimization problem:**

$$\begin{aligned} \max_{\alpha, \alpha'} & -\epsilon(\alpha' + \alpha)^\top \mathbf{1} + (\alpha' - \alpha)^\top \mathbf{Y} \\ & - \frac{1}{2}(\alpha' - \alpha)^\top \mathbf{K}(\alpha' - \alpha) \\ \text{s.t. } & (\mathbf{0} \leq \alpha \leq \mathbf{C}) \text{ or } (\mathbf{0} \leq \alpha' \leq \mathbf{C}) \text{ or } ((\alpha' - \alpha)^\top \mathbf{1} = 0) \end{aligned}$$

Here $\mathbf{K} = \{\Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)^\top\}_{i,j=1}^m = \{K(\mathbf{x}_i, \mathbf{x}_j)\}_{i,j=1}^m \in \mathbb{R}^{m \times m}$

- **Solution**

$$h(\mathbf{x}) = \sum_{i=1}^m (\alpha'_i - \alpha_i) \underbrace{K(\mathbf{x}_i, \mathbf{x})}_{\Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x})^\top} + b$$

with

$$b = \begin{cases} -\sum_{i=1}^m (\alpha'_i - \alpha_i) K(\mathbf{x}_i, \mathbf{x}_i) + y_i + \epsilon, & \text{when } 0 < \alpha_i < C \\ -\sum_{i=1}^m (\alpha'_i - \alpha_i) K(\mathbf{x}_i, \mathbf{x}_i) + y_i - \epsilon, & \text{when } 0 < \alpha'_i < C \end{cases}$$

- Support vectors: points strictly outside the tube

- **Optimization problem:**

$$\begin{aligned} \max_{\alpha, \alpha'} & -\epsilon(\alpha' + \alpha)^\top \mathbf{1} + (\alpha' - \alpha)^\top \mathbf{Y} \\ & - \frac{1}{2}(\alpha' - \alpha)^\top \mathbf{K}(\alpha' - \alpha) \\ \text{s.t. } & (\mathbf{0} \leq \alpha \leq \mathbf{C}) \text{ or } (\mathbf{0} \leq \alpha' \leq \mathbf{C}) \text{ or } ((\alpha' - \alpha)^\top \mathbf{1} = 0) \end{aligned}$$

Here $\mathbf{K} = \{\Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)^\top\}_{i,j=1}^m = \{K(\mathbf{x}_i, \mathbf{x}_j)\}_{i,j=1}^m \in \mathbb{R}^{m \times m}$

- **Solution**

$$h(\mathbf{x}) = \sum_{i=1}^m (\alpha'_i - \alpha_i) \underbrace{K(\mathbf{x}_i, \mathbf{x})}_{\Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x})^\top} + b$$

with

$$b = \begin{cases} -\sum_{i=1}^m (\alpha'_i - \alpha_i) K(\mathbf{x}_i, \mathbf{x}_i) + y_i + \epsilon, & \text{when } 0 < \alpha_i < C \\ -\sum_{i=1}^m (\alpha'_i - \alpha_i) K(\mathbf{x}_i, \mathbf{x}_i) + y_i - \epsilon, & \text{when } 0 < \alpha'_i < C \end{cases}$$

- Support vectors: points strictly outside the tube

- **Optimization problem:**

$$\begin{aligned} \max_{\alpha, \alpha'} & -\epsilon(\alpha' + \alpha)^\top \mathbf{1} + (\alpha' - \alpha)^\top \mathbf{Y} \\ & - \frac{1}{2}(\alpha' - \alpha)^\top \mathbf{K}(\alpha' - \alpha) \\ \text{s.t. } & (\mathbf{0} \leq \alpha \leq \mathbf{C}) \text{ or } (\mathbf{0} \leq \alpha' \leq \mathbf{C}) \text{ or } ((\alpha' - \alpha)^\top \mathbf{1} = 0) \end{aligned}$$

Here $\mathbf{K} = \{\Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)^\top\}_{i,j=1}^m = \{K(\mathbf{x}_i, \mathbf{x}_j)\}_{i,j=1}^m \in \mathbb{R}^{m \times m}$

- **Solution**

$$h(\mathbf{x}) = \sum_{i=1}^m (\alpha'_i - \alpha_i) \underbrace{K(\mathbf{x}_i, \mathbf{x})}_{\Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x})^\top} + b$$

with

$$b = \begin{cases} -\sum_{i=1}^m (\alpha'_i - \alpha_i) K(\mathbf{x}_i, \mathbf{x}_i) + y_i + \epsilon, & \text{when } 0 < \alpha_i < C \\ -\sum_{i=1}^m (\alpha'_i - \alpha_i) K(\mathbf{x}_i, \mathbf{x}_i) + y_i - \epsilon, & \text{when } 0 < \alpha'_i < C \end{cases}$$

- Support vectors: points strictly outside the tube

- Advantages

- strong theoretical guarantees (for that loss)
- sparser solution
- use of kernels

- Disadvantages

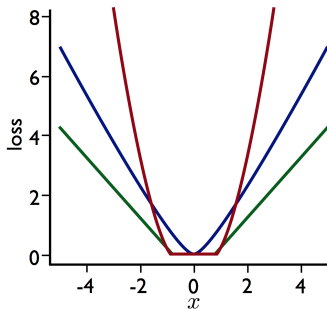
- selection of two parameters: C and ϵ . Heuristics for that:
 - * search C near maximum y , ϵ near average difference of y -s, measure of no. of SVs
- large matrices: low-rank approximations of kernel matrix

- Advantages

- strong theoretical guarantees (for that loss)
- sparser solution
- use of kernels

- Disadvantages

- selection of two parameters: C and ϵ . Heuristics for that:
 - * search C near maximum y , ϵ near average difference of y -s, measure of no. of SVs
- large matrices: low-rank approximations of kernel matrix



- quadratic ϵ -insensitive

$$x \rightarrow \max(0, |x| - \epsilon)^2$$

- Huber

$$x \rightarrow \begin{cases} x^2, & \text{if } |x| \leq c \\ 2c|x| - c^2, & \text{otherwise} \end{cases}$$

- ϵ -insensitive

$$x \rightarrow \max(0, |x| - \epsilon)$$

- SVR in case of quadratic ϵ -insensitive for $\epsilon = 0$ coincides with Kernel Ridge Regression (see lecture 2)

$$h(\mathbf{x}) = \sum_{i=1}^m \alpha_i K(\mathbf{x}_i, \mathbf{x}), \quad (*)$$

where

$$\boldsymbol{\alpha} = (\Phi(\mathbf{X}) \cdot \Phi(\mathbf{X})^\top + \lambda \mathbf{I})^{-1} \mathbf{Y} = (\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{Y},$$

where

$$\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_m\} \in \mathbb{R}^{m \times d}, \quad \mathbf{Y} = (y_1, \dots, y_m) \in \mathbb{R}^{m \times 1}$$

- In case of $\epsilon > 0$ SVR allows to reduce a number of terms in (*) above thanks to the support vector concept: explicit solution vs. sparsity!