Outline

1 Constrained optimization, Lagrangian duality and KKT

2 Support vector machines
Outline

1. Constrained optimization, Lagrangian duality and KKT

2. Support vector machines
Constrained optimization, Lagrangian duality and KKT
Optimization problem in canonical form

\[
\min_{x \in \mathcal{X}} \quad f(x)
\]
\[
\text{s.t.} \quad h_i(x) = 0, \quad i \in [1, n]
\]
\[
g_j(x) \leq 0, \quad j \in [1, m]
\]

with

- \( \mathcal{X} \subset \mathbb{R}^p \).
- \( f, g_j \) functions,
- \( h_i \) affine functions.
Review: Constrained optimization

Optimization problem in canonical form

\[
\min_{x \in \mathcal{X}} \quad f(x)
\]

s.t.
\[
\begin{align*}
  h_i(x) &= 0, & i & \in [1, n] \\
  g_j(x) &\leq 0, & j & \in [1, m]
\end{align*}
\]

with
\[
\begin{align*}
  &\mathcal{X} \subset \mathbb{R}^p. \\
  &f, g_j \text{ functions,} \\
  &h_i \text{ affine functions.}
\end{align*}
\]

The problem is convex if \( f, g_j \) and \( \mathcal{X} \) are convex (w.l.o.g. \( \mathcal{X} \neq \emptyset \)).
Review: Constrained optimization

Optimization problem in canonical form

\[
\min_{x \in \mathcal{X}} \quad f(x)
\]

s.t. \( h_i(x) = 0, \quad i \in [1, n] \)

\( g_j(x) \leq 0, \quad j \in [1, m] \)

with

- \( \mathcal{X} \subset \mathbb{R}^p \).
- \( f, g_j \) functions,
- \( h_i \) affine functions.

The problem is convex if \( f, g_j \) and \( \mathcal{X} \) are convex (w.l.o.g. \( \mathcal{X} \neq \emptyset \)).

Lagrangian

\[
\mathcal{L}(x, \lambda, \mu) = f(x) + \sum_{i=1}^{n} \lambda_i h_i(x) + \sum_{j=1}^{m} \mu_j g_j(x)
\]
Lagrangian duality

Lagrangian

\[ \mathcal{L}(x, \lambda, \mu) = f(x) + \sum_{i=1}^{n} \lambda_i h_i(x) + \sum_{j=1}^{m} \mu_j g_j(x) \]
Lagrangian duality

Lagrangian

\[ \mathcal{L}(x, \lambda, \mu) = f(x) + \sum_{i=1}^{n} \lambda_i h_i(x) + \sum_{j=1}^{m} \mu_j g_j(x) \]

Primal vs Dual problem

\[ p^* = \min_{x \in \mathcal{X}} \max_{\lambda \in \mathbb{R}^n, \mu \in \mathbb{R}_+^m} \mathcal{L}(x, \lambda, \mu) \quad (P) \]

\[ d^* = \max_{\lambda \in \mathbb{R}^n, \mu \in \mathbb{R}_+^m} \min_{x \in \mathcal{X}} \mathcal{L}(x, \lambda, \mu) \quad (D) \]
Maxmin inequalities

\[
\max_y \min_x f(x, y) \leq \min_x \max_y f(x, y)
\]

Weak duality

In general, we have \( d^* \leq p^* \). This is called weak duality.

Strong duality

In some cases, we have strong duality: \( d^* = p^* \). Solutions to (P) and (D) are the same.
Maxmin inequalities

\[ f(x, y) \leq \max_y f(x, y) \]
Maxmin inequalities

\[ \min_x f(x, y) \leq \min_x \max_y f(x, y) \]
Maxmin inequalities

\[ \max_y \min_x f(x, y) \leq \min_x \max_y f(x, y) \]
Maxmin inequalities

\[ \max_y \min_x f(x, y) \leq \min_x \max_y f(x, y) \]

**Weak duality**

In general, we have \( d^* \leq p^* \). This is called weak duality.
Maxmin inequalities

\[
\max_y \min_x f(x, y) \leq \min_x \max_y f(x, y)
\]

**Weak duality**
In general, we have \( d^* \leq p^* \). This is called weak duality.

**Strong duality**
In some cases, we have strong duality:
- \( d^* = p^* \)
- Solutions to (P) and (D) are the same
Slater’s qualification condition

Slater’s qualification condition is a condition on the constraints that guarantees that strong duality holds.
Consider an optimization problem in canonical form.
Slater’s qualification condition

Slater’s qualification condition is a condition on the constraints that guarantees that strong duality holds.
Consider an optimization problem in canonical form.

Definition: Slater’s condition (strong form)
There exists $\mathbf{x} \in \mathcal{X}$ such that $h(\mathbf{x}) = 0$ and $g(\mathbf{x}) < 0$ entrywise.

Definition: Slater’s condition (weak form)
There exists $\mathbf{x} \in \mathcal{X}$ such that $h(\mathbf{x}) = 0$ and $g(\mathbf{x}) \leq 0$ entrywise, but with $g_i(\mathbf{x}) < 0$ if $g_i$ is not affine.

Slater’s conditions requires that there exists a feasible point which is strictly feasible for all non-affine constraints.
Slater’s qualification condition

Slater’s qualification condition is a condition on the constraints that guarantees that strong duality holds.

Consider an optimization problem in canonical form.

Definition: Slater’s condition (strong form)
There exists \( \mathbf{x} \in \mathcal{X} \) such that \( h(\mathbf{x}) = 0 \) and \( g(\mathbf{x}) < 0 \) entrywise.

Definition: Slater’s condition (weak form)
There exists \( \mathbf{x} \in \mathcal{X} \) such that \( h(\mathbf{x}) = 0 \) and \( g(\mathbf{x}) \leq 0 \) entrywise, but with \( g_i(\mathbf{x}) < 0 \) if \( g_i \) is not affine.
Slater’s qualification condition

Slater’s qualification condition is a condition on the constraints that guarantees that strong duality holds. Consider an optimization problem in canonical form.

**Definition: Slater’s condition (strong form)**

There exists \( x \in \mathcal{X} \) such that \( h(x) = 0 \) and \( g(x) < 0 \) entrywise.

**Definition: Slater’s condition (weak form)**

There exists \( x \in \mathcal{X} \) such that \( h(x) = 0 \) and \( g(x) \leq 0 \) entrywise, but with \( g_i(x) < 0 \) if \( g_i \) is not affine.

Slater’s conditions requires that there exists a feasible point which is strictly feasible for all non-affine constraints.
Karush-Kuhn-Tucker conditions

**Theorem**

*For a convex problem defined by differentiable functions* \( f, h_i, g_j, \) \( x \) *is an optimal solution if and only if there exists* \((\lambda, \mu)\) *such that the KKT conditions are satisfied.*

**KKT conditions**

\[
\nabla f(x) + \sum_{i=1}^{n} \lambda_i \nabla h_i(x) + \sum_{j=1}^{m} \mu_j \nabla g_j(x) = 0 \quad \text{(Lagrangian stationarity)}
\]

\[
h(x) = 0, \quad g(x) \leq 0 \quad \text{(primal feasibility)}
\]

\[
\mu_j \geq 0 \quad \text{(dual feasibility)}
\]

\[
\forall j \in [1, m], \quad \mu_j g_j(x) = 0 \quad \text{(complementary slackness)}
\]
Outline

1 Constrained optimization, Lagrangian duality and KKT

2 Support vector machines
Support vector machines
Hard margin SVM

- Binary classification problem with $y_i \in \{-1, 1\}$. 

\[
\text{Constraints: for } y_i = 1 \text{ require } w^\top x_i + b \geq 1 \\
\text{for } y_i = -1 \text{ require } w^\top x_i + b \leq -1
\]

This leads to
\[
\min \frac{1}{2} \|w\|^2 \quad \text{s.t.} \quad \forall i, y_i(w^\top x_i + b) \geq 1
\]

quadratic program (not a so useful property nowadays)

unfeasible if the data is not separable
Hard margin SVM

- Binary classification problem with $y_i \in \{-1, 1\}$.
- Margin $\frac{1}{\|w\|}$
Hard margin SVM

- Binary classification problem with \( y_i \in \{-1, 1\} \).
- Margin \( \frac{1}{\|w\|} \)
- Constraints:
  - for \( y_i = 1 \) require \( w^\top x_i + b \geq 1 \)
  - for \( y_i = -1 \) require \( w^\top x_i + b \leq -1 \)

This leads to
\[
\min \frac{1}{2} \|w\|^2 \quad \text{s.t.} \quad \forall i, y_i (w^\top x_i + b) \geq 1
\]
quadratic program (not a so useful property nowadays)
unfeasible if the data is not separable

SVM, kernel methods and multiclass
Hard margin SVM

- Binary classification problem with $y_i \in \{-1, 1\}$.
- Margin $\frac{1}{\|w\|}$
- Constraints:
  - for $y_i = 1$ require $w^T x_i + b \geq 1$
  - for $y_i = -1$ require $w^T x_i + b \leq -1$

This leads to

$$\min \frac{1}{2}\|w\|^2 \quad \text{s.t.} \quad \forall i, \quad y_i(w^T x_i + b) \geq 1$$
Hard margin SVM

- Binary classification problem with $y_i \in \{-1, 1\}$.
- Margin $\frac{1}{\|w\|}$
- Constraints:
  - for $y_i = 1$ require $w^T x_i + b \geq 1$
  - for $y_i = -1$ require $w^T x_i + b \leq -1$

This leads to

$$\min \frac{1}{2}\|w\|^2 \quad \text{s.t.} \quad \forall i, \quad y_i (w^T x_i + b) \geq 1$$

- quadratic program (not a so useful property nowadays)
- unfeasible if the data is not separable
Hard-margin SVM
Soft margin SVM

- Authorize some points to be on the wrong side of the margin
- Penalize by a cost proportional to the distance to the margin
- Introduce some slack variables $\xi_i$ measuring the violation for each datapoint.

$$
\begin{align*}
\min & \quad \|w\|^2 + C \sum_{i=1}^{n} \xi_i \\
\text{s.t.} & \quad \forall i, \quad \left\{ y_i (w^\top x_i + b) \geq 1 - \xi_i \right\} \quad \xi_i \geq 0
\end{align*}
$$
Soft margin SVM

- Authorize some points to be on the wrong side of the margin
- Penalize by a cost proportional to the distance to the margin
- Introduce some slack variables $\xi_i$ measuring the violation for each datapoint.

\[
\min_{\mathbf{w}, \xi} \quad \frac{1}{2} \| \mathbf{w} \|^2 + C \sum_{i=1}^{n} \xi_i \\
\text{s.t.} \quad \forall i, \begin{cases} 
  y_i (\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i \\
  \xi_i \geq 0
\end{cases}
\]
Lagrangian of the SVM

\[ \mathcal{L}(w, \xi, \alpha, \nu) \]

\[ = \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} \xi_i + \sum_{i=1}^{n} \alpha_i (1 - \xi_i - y_i (w^T x_i + b)) - \sum_{i=1}^{n} \nu_i \xi_i \]
Lagrangian of the SVM

\[ \mathcal{L}(w, \xi, \alpha, \nu) \]

\[ = \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} \xi_i + \sum_{i=1}^{n} \alpha_i (1 - \xi_i - y_i (w^\top x_i + b)) - \sum_{i=1}^{n} \nu_i \xi_i \]

\[ = \frac{1}{2} \|w\|^2 - w^\top \left( \sum_{i=1}^{n} \alpha_i y_i x_i \right) + \sum_{i=1}^{n} \xi_i (C - \alpha_i - \nu_i) - \sum_{i=1}^{n} \alpha_i y_i b + \sum_{i=1}^{n} \alpha_i \]

Stationarity of the Lagrangian

\[ \nabla_w \mathcal{L} = w - \sum_{i=1}^{n} \alpha_i y_i x_i \], \[ \frac{\partial \mathcal{L}}{\partial \xi_i} = C - \alpha_i - \nu_i \]

So that \[ \nabla \mathcal{L} = 0 \] leads to \[ w = \sum_{i=1}^{n} \alpha_i y_i x_i, \] \[ 0 \leq \alpha_i \leq C \] and \[ \sum_{i=1}^{n} \alpha_i y_i = 0 \].
Lagrangian of the SVM

\[ \mathcal{L}(w, \xi, \alpha, \nu) = \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} \xi_i + \sum_{i=1}^{n} \alpha_i (1 - \xi_i - y_i (w^\top x_i + b)) - \sum_{i=1}^{n} \nu_i \xi_i \]

\[ \mathcal{L}(w, \xi, \alpha, \nu) = \frac{1}{2} \|w\|^2 - w^\top \left( \sum_{i=1}^{n} \alpha_i y_i x_i \right) + \sum_{i=1}^{n} \xi_i (C - \alpha_i - \nu_i) - \sum_{i=1}^{n} \alpha_i y_i b + \sum_{i=1}^{n} \alpha_i \]

Stationarity of the Lagrangian

\[ \nabla_w \mathcal{L} = w - \sum_{i=1}^{n} \alpha_i y_i x_i, \quad \frac{\partial \mathcal{L}}{\partial \xi_i} = C - \alpha_i - \nu_i \quad \text{and} \quad \frac{\partial \mathcal{L}}{\partial b} = \sum_{i=1}^{n} \alpha_i y_i. \]
Lagrangian of the SVM

\[ \mathcal{L}(w, \xi, \alpha, \nu) \]

\[
= \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} \xi_i + \sum_{i=1}^{n} \alpha_i \left(1 - \xi_i - y_i (w^\top x_i + b)\right) - \sum_{i=1}^{n} \nu_i \xi_i \\
= \frac{1}{2} \|w\|^2 - w^\top \left(\sum_{i=1}^{n} \alpha_i y_i x_i\right) + \sum_{i=1}^{n} \xi_i \left(C - \alpha_i - \nu_i\right) - \sum_{i=1}^{n} \alpha_i y_i b + \sum_{i=1}^{n} \alpha_i \\

\]

Stationarity of the Lagrangian

\[ \nabla_w \mathcal{L} = w - \sum_{i=1}^{n} \alpha_i y_i x_i, \quad \frac{\partial \mathcal{L}}{\partial \xi_i} = C - \alpha_i - \nu_i \quad \text{and} \quad \frac{\partial \mathcal{L}}{\partial b} = \sum_{i=1}^{n} \alpha_i y_i. \]

So that \( \nabla \mathcal{L} = 0 \) leads to

\[ w = \sum_{i=1}^{n} \alpha_i y_i x_i, \quad 0 \leq \alpha_i \leq C \quad \text{and} \quad \sum_{i=1}^{n} \alpha_i y_i = 0. \]
Dual of the SVM

\[
\max_{\alpha} - \frac{1}{2} \left\| \sum_{i=1}^{n} \alpha_i y_i x_i \right\|^2 + \sum_{i=1}^{n} \alpha_i \\
\text{s.t.} \quad \sum_{i=1}^{n} \alpha_i y_i = 0, \quad \forall i, \ 0 \leq \alpha_i \leq C.
\]

\[
\max_{\alpha} - \frac{1}{2} \alpha^\top D_y K D_y \alpha + \alpha^\top 1 \\
\text{s.t.} \quad \alpha^\top y = 0, \quad 0 \leq \alpha \leq C.
\]

with

- \( y^\top = (y_1, \ldots, y_n) \) the vector of labels
- \( D_y = \text{Diag}(y) \) a diagonal matrix with the label \( y \)
- \( K \) the Gram matrix with \( K_{ij} = x_i^\top x_j \)
Dual of the SVM

$$\max_{\alpha} - \frac{1}{2} \sum_{1 \leq i, j \leq n} \alpha_i \alpha_j y_i y_j x_i^\top x_j + \sum_{i=1}^{n} \alpha_i$$

s.t. $$\sum_{i=1}^{n} \alpha_i y_i = 0, \quad \forall i, \ 0 \leq \alpha_i \leq C.$$
Dual of the SVM

\[
\max_{\alpha} - \frac{1}{2} \left\| \sum_{i=1}^{n} \alpha_i y_i x_i \right\|^2 + \sum_{i=1}^{n} \alpha_i \\
\text{s.t.} \quad \sum_{i=1}^{n} \alpha_i y_i = 0, \quad \forall i, \ 0 \leq \alpha_i \leq C.
\]

\[
\max_{\alpha} - \frac{1}{2} \alpha^\top D_y K D_y \alpha + \alpha^\top 1 \\
\text{s.t.} \quad \alpha^\top y = 0, \quad 0 \leq \alpha \leq C.
\]

with

- \( y^\top = (y_1, \ldots, y_n) \) the vector of labels
- \( D_y = \text{Diag}(y) \) a diagonal matrix with the label
- \( K \) the Gram matrix with \( K_{ij} = x_i^\top x_j \)
KKT conditions for the SVM

\[ w = \sum_{i=1}^{n} \alpha_i y_i x_i \quad \text{(LS)} \]

\[ \alpha_i + \nu_i = C \quad \text{(LS)} \]

\[ \sum_{i=1}^{n} \alpha_i y_i = 0 \quad \text{(LS)} \]

\[ 1 - \xi_i - y_i f(x_i) \geq 0 \quad \text{(PF)} \]

\[ \xi_i \geq 0 \quad \text{(PF)} \]

\[ \alpha_i \geq 0 \quad \text{(DF)} \]

\[ \nu_i \geq 0 \quad \text{(DF)} \]

\[ \alpha_i (1 - \xi_i - y_i f(x_i)) = 0 \quad \text{(CS)} \]

\[ \nu_i \xi_i = 0 \quad \text{(CS)} \]

with

\[ f(x_i) = w^\top x_i + b \]

Let

\[ I = \{ i | \xi_i > 0 \} \]

\[ M = \{ i | y_i f(x_i) = 1 \} \]

\[ S = \{ i | \alpha_i \neq 0 \} \]

\[ W = (I \cup M) \quad c \小编 \in I \Rightarrow \nu_i = 0 \Rightarrow \alpha_i = C \Rightarrow i \in S \]

\[ i \in W \Rightarrow \alpha_i = 0 \iff i \notin S \]

We have \( 0 \leq \alpha_i \leq C \).

The set \( S \) of support vectors is therefore composed of some points on the margin and all incorrectly placed points.
KKT conditions for the SVM

\[ \mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i \quad (LS) \]
\[ \alpha_i + \nu_i = C \quad (LS) \]
\[ \sum_{i=1}^{n} \alpha_i y_i = 0 \quad (LS) \]
\[ 1 - \xi_i - y_i f(\mathbf{x}_i) \geq 0 \quad (PF) \]
\[ \xi_i \geq 0 \quad (PF) \]
\[ \alpha_i \geq 0 \quad (DF) \]
\[ \nu_i \geq 0 \quad (DF) \]
\[ \alpha_i (1 - \xi_i - y_i f(\mathbf{x}_i)) = 0 \quad (CS) \]
\[ \nu_i \xi_i = 0 \quad (CS) \]

with \( f(\mathbf{x}_i) = \mathbf{w}^\top \mathbf{x}_i + b \)
KKT conditions for the SVM

Let

\[ I = \{ i \mid \xi_i > 0 \} \]

\[ W = (I \cup M) \]

\[ i \in I \Rightarrow \nu_i = 0 \Rightarrow \alpha_i = C \Rightarrow i \in S \]

\[ i \in W \Rightarrow \alpha_i = 0 \iff i \notin S \]

\[ 0 \leq \alpha_i \leq C. \]

The set \( S \) of support vectors is therefore composed of some points on the margin and all incorrectly placed points.

\[ \mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i \quad (\text{LS}) \]

\[ \alpha_i + \nu_i = C \quad (\text{LS}) \]

\[ \sum_{i=1}^{n} \alpha_i y_i = 0 \quad (\text{LS}) \]

\[ 1 - \xi_i - y_i f(\mathbf{x}_i) \geq 0 \quad (\text{PF}) \]

\[ \xi_i \geq 0 \quad (\text{PF}) \]

\[ \alpha_i \geq 0 \quad (\text{DF}) \]

\[ \nu_i \geq 0 \quad (\text{DF}) \]

\[ \alpha_i (1 - \xi_i - y_i f(\mathbf{x}_i)) = 0 \quad (\text{CS}) \]

\[ \nu_i \xi_i = 0 \quad (\text{CS}) \]

with \( f(\mathbf{x}_i) = \mathbf{w}^\top \mathbf{x}_i + b \)
KKT conditions for the SVM

\[ w = \sum_{i=1}^{n} \alpha_i y_i x_i \]  \hspace{1cm} \text{(LS)}

\[ \alpha_i + \nu_i = C \]  \hspace{1cm} \text{(LS)}

\[ \sum_{i=1}^{n} \alpha_i y_i = 0 \]  \hspace{1cm} \text{(LS)}

\[ 1 - \xi_i - y_i f(x_i) \geq 0 \]  \hspace{1cm} \text{(PF)}

\[ \xi_i \geq 0 \]  \hspace{1cm} \text{(PF)}

\[ \alpha_i \geq 0 \]  \hspace{1cm} \text{(DF)}

\[ \nu_i \geq 0 \]  \hspace{1cm} \text{(DF)}

\[ \alpha_i (1 - \xi_i - y_i f(x_i)) = 0 \]  \hspace{1cm} \text{(CS)}

\[ \nu_i \xi_i = 0 \]  \hspace{1cm} \text{(CS)}

Let

- \( I = \{ i \mid \xi_i > 0 \} \)
- \( M = \{ i \mid y_i f(x_i) = 1 \} \)

with \( f(x_i) = w^\top x_i + b \)
KKT conditions for the SVM

\[
\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i \quad \text{(LS)}
\]

\[
\alpha_i + \nu_i = C \quad \text{(LS)}
\]

\[
\sum_{i=1}^{n} \alpha_i y_i = 0 \quad \text{(LS)}
\]

\[
1 - \xi_i - y_i f(\mathbf{x}_i) \geq 0 \quad \text{(PF)}
\]

\[
\xi_i \geq 0 \quad \text{(PF)}
\]

\[
\alpha_i \geq 0 \quad \text{(DF)}
\]

\[
\nu_i \geq 0 \quad \text{(DF)}
\]

\[
\alpha_i (1 - \xi_i - y_i f(\mathbf{x}_i)) = 0 \quad \text{(CS)}
\]

\[
\nu_i \xi_i = 0 \quad \text{(CS)}
\]

with \(f(\mathbf{x}_i) = \mathbf{w}^\top \mathbf{x}_i + b\)

Let

- \(I = \{ i \mid \xi_i > 0 \}\)
- \(M = \{ i \mid y_i f(\mathbf{x}_i) = 1 \}\)
- \(S = \{ i \mid \alpha_i \neq 0 \}\)
KKT conditions for the SVM

\[ \mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i \quad \text{(LS)} \]
\[ \alpha_i + \nu_i = C \quad \text{(LS)} \]
\[ \sum_{i=1}^{n} \alpha_i y_i = 0 \quad \text{(LS)} \]
\[ 1 - \xi_i - y_i f(\mathbf{x}_i) \geq 0 \quad \text{(PF)} \]
\[ \xi_i \geq 0 \quad \text{(PF)} \]
\[ \alpha_i \geq 0 \quad \text{(DF)} \]
\[ \nu_i \geq 0 \quad \text{(DF)} \]
\[ \alpha_i (1 - \xi_i - y_i f(\mathbf{x}_i)) = 0 \quad \text{(CS)} \]
\[ \nu_i \xi_i = 0 \quad \text{(CS)} \]

Let

- \( I = \{ i \mid \xi_i > 0 \} \)
- \( M = \{ i \mid y_i f(\mathbf{x}_i) = 1 \} \)
- \( S = \{ i \mid \alpha_i \neq 0 \} \)
- \( W = (I \cup M)^c \)

with \( f(\mathbf{x}_i) = \mathbf{w}^\top \mathbf{x}_i + b \)
KKT conditions for the SVM

\[ \mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i \quad (LS) \]
\[ \alpha_i + \nu_i = C \quad (LS) \]
\[ \sum_{i=1}^{n} \alpha_i y_i = 0 \quad (LS) \]
\[ 1 - \xi_i - y_i f(\mathbf{x}_i) \geq 0 \quad (PF) \]
\[ \xi_i \geq 0 \quad (PF) \]
\[ \alpha_i \geq 0 \quad (DF) \]
\[ \nu_i \geq 0 \quad (DF) \]
\[ \alpha_i (1 - \xi_i - y_i f(\mathbf{x}_i)) = 0 \quad (CS) \]
\[ \nu_i \xi_i = 0 \quad (CS) \]

Let
- \( I = \{ i \mid \xi_i > 0 \} \)
- \( M = \{ i \mid y_i f(\mathbf{x}_i) = 1 \} \)
- \( S = \{ i \mid \alpha_i \neq 0 \} \)
- \( W = (I \cup M)^c \)

with \( f(\mathbf{x}_i) = \mathbf{w}^\top \mathbf{x}_i + b \)
KKT conditions for the SVM

\[ \mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i \]  
\[ \alpha_i + \nu_i = C \]  
\[ \sum_{i=1}^{n} \alpha_i y_i = 0 \]

1. \[1 - \xi_i - y_i f(\mathbf{x}_i) \geq 0 \]
2. \[\xi_i \geq 0 \]
3. \[\alpha_i \geq 0 \]
4. \[\nu_i \geq 0 \]
5. \[\alpha_i (1 - \xi_i - y_i f(\mathbf{x}_i)) = 0 \]
6. \[\nu_i \xi_i = 0 \]

with \( f(\mathbf{x}_i) = \mathbf{w}^\top \mathbf{x}_i + b \)

Let

- \( I = \{ i \mid \xi_i > 0 \} \)
- \( M = \{ i \mid y_i f(\mathbf{x}_i) = 1 \} \)
- \( S = \{ i \mid \alpha_i \neq 0 \} \)
- \( W = (I \cup M)^c \)

\[ i \in I \Rightarrow \nu_i = 0 \Rightarrow \alpha_i = C \Rightarrow i \in S \]
KKT conditions for the SVM

\[ w = \sum_{i=1}^{n} \alpha_i y_i x_i \quad \text{(LS)} \]
\[ \alpha_i + \nu_i = C \quad \text{(LS)} \]
\[ \sum_{i=1}^{n} \alpha_i y_i = 0 \quad \text{(LS)} \]
\[ 1 - \xi_i - y_i f(x_i) \geq 0 \quad \text{(PF)} \]
\[ \xi_i \geq 0 \quad \text{(PF)} \]
\[ \alpha_i \geq 0 \quad \text{(DF)} \]
\[ \nu_i \geq 0 \quad \text{(DF)} \]
\[ \alpha_i (1 - \xi_i - y_i f(x_i)) = 0 \quad \text{(CS)} \]
\[ \nu_i \xi_i = 0 \quad \text{(CS)} \]

with \( f(x_i) = w^T x_i + b \)

Let

- \( I = \{ i \mid \xi_i > 0 \} \)
- \( M = \{ i \mid y_i f(x_i) = 1 \} \)
- \( S = \{ i \mid \alpha_i \neq 0 \} \)
- \( W = (I \cup M)^c \)

\[ i \in I \Rightarrow \nu_i = 0 \Rightarrow \alpha_i = C \Rightarrow i \in S \]
\[ i \in W \Rightarrow \alpha_i = 0 \Leftrightarrow i \notin S \]

We have \( 0 \leq \alpha_i \leq C \).
KKT conditions for the SVM

\[ w = \sum_{i=1}^{n} \alpha_i y_i x_i \quad \text{(LS)} \]

\[ \alpha_i + \nu_i = C \quad \text{(LS)} \]

\[ \sum_{i=1}^{n} \alpha_i y_i = 0 \quad \text{(LS)} \]

\[ 1 - \xi_i - y_i f(x_i) \geq 0 \quad \text{(PF)} \]

\[ \xi_i \geq 0 \quad \text{(PF)} \]

\[ \alpha_i \geq 0 \quad \text{(DF)} \]

\[ \nu_i \geq 0 \quad \text{(DF)} \]

\[ \alpha_i (1 - \xi_i - y_i f(x_i)) = 0 \quad \text{(CS)} \]

\[ \nu_i \xi_i = 0 \quad \text{(CS)} \]

with \( f(x_i) = w^\top x_i + b \)

Let

- \( I = \{ i \mid \xi_i > 0 \} \)
- \( M = \{ i \mid y_i f(x_i) = 1 \} \)
- \( S = \{ i \mid \alpha_i \neq 0 \} \)
- \( W = (I \cup M)^c \)

\[ i \in I \Rightarrow \nu_i = 0 \Rightarrow \alpha_i = C \Rightarrow i \in S \]

\[ i \in W \Rightarrow \alpha_i = 0 \Leftrightarrow i \notin S \]

We have \( 0 \leq \alpha_i \leq C \).

The set \( S \) of support vectors is therefore composed of some points on the margin and all incorrectly placed points.
SVM summary so far

- Optimization problem formulated as a strongly convex QP

The dual solution $\alpha^*$ is not necessarily unique $\Rightarrow$ there might be several possible sets of support vectors.

How do we determine $b$?
SVM summary so far

- Optimization problem formulated as a strongly convex QP
- whose dual is also a QP

Remarks:
1. the dual solution $\alpha^*$ is not necessarily unique $\Rightarrow$ there might be several possible sets of support vectors.
2. How do we determine $b$?
SVM summary so far

- Optimization problem formulated as a strongly convex QP
- whose dual is also a QP
- The support vectors are the points that have a non zero optimal weight $\alpha_i$

Remarks:
1. The dual solution $\alpha^*$ is not necessarily unique $\Rightarrow$ there might be several possible sets of support vectors.
2. How do we determine $b$?
SVM summary so far

- Optimization problem formulated as a strongly convex QP
- whose dual is also a QP
- The support vectors are the points that have a non zero optimal weight $\alpha_i$
- The optimal solution is $w^* = \sum_{i \in S} \alpha_i^* y_i x_i$, i.e. a weighted combination of the support vectors
- The solution does not depend on the well-classified points
  → Leads to working set strategies.

Remarks:
1. the dual solution $\alpha^*$ is not necessarily unique
   $\Rightarrow$ there might be several possible sets of support vectors.
2. How do we determine $b$?
SVM summary so far

- Optimization problem formulated as a strongly convex QP
- whose dual is also a QP
- The support vectors are the points that have a non zero optimal weight $\alpha_i$
- The optimal solution is $\mathbf{w}^* = \sum_{i \in S} \alpha_i^* y_i \mathbf{x}_i$, i.e. a weighted combination of the support vectors
- The solution does not depend on the well-classified points
  → Leads to working set strategies.
  → Computational gain

Remarks:
1. the dual solution $\alpha^*$ is not necessarily unique $\Rightarrow$ there might be several possible sets of support vectors.
2. How do we determine $b$?
SVM summary so far

- Optimization problem formulated as a strongly convex QP
- whose dual is also a QP
- The support vectors are the points that have a non zero optimal weight $\alpha_i$
- The optimal solution is $w^* = \sum_{i \in S} \alpha_i^* y_i x_i$, i.e. a weighted combination of the support vectors
- The solution does not depend on the well-classified points
  $\rightarrow$ Leads to working set strategies.
  $\rightarrow$ Computational gain

Remarks:
- the dual solution $\alpha^*$ is not necessarily unique $\Rightarrow$ there might be several possible sets of support vectors.
SVM summary so far

- Optimization problem formulated as a strongly convex QP
- whose dual is also a QP
- The support vectors are the points that have a non zero optimal weight $\alpha_i$
- The optimal solution is $w^* = \sum_{i \in S} \alpha_i^* y_i x_i$, i.e. a weighted combination of the support vectors
- The solution does not depend on the well-classified points
  \rightarrow Leads to working set strategies.
  \rightarrow Computational gain

Remarks:

1. the dual solution $\alpha^*$ is not necessarily unique $ \Rightarrow $ there might be several possible sets of support vectors.
2. How do we determine $b$?
SVM summary so far

- Optimization problem formulated as a strongly convex QP
- whose dual is also a QP
- The support vectors are the points that have a non zero optimal weight $\alpha_i$
- The optimal solution is $w^* = \sum_{i \in S} \alpha_i^* y_i x_i$, i.e. a weighted combination of the support vectors
- The solution does not depend on the well-classified points
  → Leads to working set strategies.
  → Computational gain

Remarks:
1. the dual solution $\alpha^*$ is not necessarily unique $\Rightarrow$ there might be several possible sets of support vectors.
2. How do we determine $b$?
Representer property for the SVM

\[ f^*(x) = w^* \mathbf{x} + b \]
\[ = \sum_{i \in S} \alpha_i^* y_i \mathbf{x}_i^\top \mathbf{x} + b \]
\[ = \sum_{i \in S} \alpha_i^* y_i \ k(\mathbf{x}_i, \mathbf{x}) + b \]

with \( k(\mathbf{x}, \mathbf{x}^\prime) = \mathbf{x}^\top \mathbf{x}^\prime \).
Representer property for the SVM

\[ f^*(x) = \mathbf{w}^* \mathbf{x} + b \]
\[ = \sum_{i \in S} \alpha_i^* y_i \mathbf{x}_i^\top \mathbf{x} + b \]
\[ = \sum_{i \in S} \alpha_i^* y_i \cdot k(\mathbf{x}_i, \mathbf{x}) + b \]

with \( k(\mathbf{x}, \mathbf{x}') = \mathbf{x}^\top \mathbf{x}' \).

Eventually, this whole formulation depends only on the dot product between points.
Representer property for the SVM

\[ f^*(x) = w^* \mathbf{x} + b \]
\[ = \sum_{i \in S} \alpha_i^* y_i x_i^T \mathbf{x} + b \]
\[ = \sum_{i \in S} \alpha_i^* y_i k(x_i, \mathbf{x}) + b \]

with \( k(x, x') = x^T x' \).

- Eventually, this whole formulation depends only on the dot product between points
- → Can we use another dot product than the one associated to the usual Euclidean distance in \( \mathbb{R}^p \)?
Hinge loss interpretation of the SVM

\[ \min_{w, \xi} \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{n} \xi_i \]

s.t. \( \forall i, \begin{cases} y_i (w^T x_i + b) \geq 1 - \xi_i \\ \xi_i \geq 0 \end{cases} \)

Define the hinge loss \( \ell(a, y) = (1 - ya)^+ \) with \((u)^+ = \max(u, 0)\).
Hinge loss interpretation of the SVM

\[ \min_{\mathbf{w}, \xi} \frac{1}{2} \| \mathbf{w} \|^2 + C \sum_{i=1}^{n} \xi_i \]

\[ \text{s.t. } \forall i, \begin{cases} 
\xi_i \geq 1 - y_i (\mathbf{w}^\top \mathbf{x}_i + b) \\
\xi_i \geq 0
\end{cases} \]

\[ \min_{\mathbf{w}} \frac{1}{2} \| \mathbf{w} \|^2 + C \sum_{i=1}^{n} \max (1 - y_i (\mathbf{w}^\top \mathbf{x}_i + b), 0) \]

Define the hinge loss \( \ell(a, y) = (1 - ya)_+ \) with \( (u)_+ = \max(u, 0) \).
Hinge loss interpretation of the SVM

\[
\min_{w, \xi} \quad \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} \xi_i \\
\text{s.t.} \quad \forall i, \begin{cases} y_i (w^\top x_i + b) \geq 1 - \xi_i \\
\quad \xi_i \geq 0 \end{cases}
\]

\[
\min_w \quad \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} \max(1 - y_i (w^\top x_i + b), 0)
\]

Define the hinge loss \( \ell(a, y) = (1 - ya)_+ \) with \( (u)_+ = \max(u, 0) \).

Our problem is now of the form

\[
\min_w \sum_{i=1}^{n} \ell(f(x_i), y_i) + \frac{1}{2C} \|w\|^2 \quad \text{with} \quad f(x) = w^\top x + b.
\]
The hinge loss is the “least convex” loss which upper bounds the 0-1 loss and equals 0 for large scores.
SVM with the quadratic hinge loss

Quadratic hinge loss: \( \ell(a, y) = (1 - ya)^2_+ \).
SVM with the quadratic hinge loss

Quadratic hinge loss: $\ell(a, y) = (1 - ya)_+^2$.

Quadratic SVM

$$\min_w \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} \max (1 - y_i (w^\top x_i + b), 0)^2$$
SVM with the quadratic hinge loss

Quadratic hinge loss: \[ \ell(a, y) = \frac{1}{2} (1 - ya)^2. \]

Quadratic SVM

\[
\min_w \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} \max \left(1 - y_i (w^T x_i + b), 0\right)^2
\]

\[
\min_{w, \xi} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} \xi_i^2
\]

s.t. \quad \forall i, \quad \begin{cases} y_i (w^T x_i + b) \geq 1 - \xi_i \\ \xi_i \geq 0 \end{cases} \]
SVM with the quadratic hinge loss

Quadratic hinge loss: \( \ell(a, y) = (1 - ya)_+^2 \).

Quadratic SVM

\[
\min_w \quad \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} \max (1 - y_i (w^\top x_i + b), 0)^2
\]

\[
\min_{w, \xi} \quad \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} \xi_i^2
\]

s.t. \( \forall i, \begin{cases} y_i (w^\top x_i + b) \geq 1 - \xi_i \\ \xi_i \geq 0 \end{cases} \)

→ Penalizes more strongly misclassified points
→ Less robust to outliers
→ Tends to be less sparse
→ Score in \([0, 1]\) for \(n\) large, interpretable as a probability.
Imbalanced classification

Learn a binary classifier from \((x_i, y_i)\) pairs with \(P = \{i | y_i = 1\}\), \(N = \{i | y_i = -1\}\), \(n_+ = |P|\), \(n_- = |N|\), and with \(n_+ \ll n_-\).

Problem: to minimize the number of mistakes the classifier learned might classify all points as negatives.

Some ways to address the issue:
- Subsample the negatives, and learn an ensemble of classifiers.
- Introduce different costs for the positives and negatives.

Minimize
\[
\min_{w \in \mathbb{R}^p} \frac{1}{2} \|w\|^2_2 + C_+ \sum_{i \in P} \xi_i + C_- \sum_{i \in N} \xi_i
\]
s.t.
\[
\forall i, y_i (w^\top x_i + b) \geq 1 - \xi_i
\]

Naive choice:
\[
C_+ = C / n_+ \quad \text{and} \quad C_- = C / n_-
\]

Is suboptimal in theory and in practice! → Better to search for the optimal hyperparameter pair \((C_+, C_-)\).

SVM, kernel methods and multiclass
Imbalanced classification

Learn a binary classifier from \((x_i, y_i)\) pairs with
\[
\mathcal{P} = \{i \mid y_i = 1\} \quad \mathcal{N} = \{i \mid y_i = -1\},
\]

\[n_+ = |\mathcal{P}|, \quad n_- = |\mathcal{N}| \quad \text{and with} \quad n_+ \ll n_-.
\]
Imbalanced classification

Learn a binary classifier from \((x_i, y_i)\) pairs with

\[ P = \{ i \mid y_i = 1 \} \quad N = \{ i \mid y_i = -1 \}, \]

\[ n_+ = |P|, \quad n_- = |N| \quad \text{and with} \quad n_+ \ll n_. \]

**Problem:** to minimize the number of mistakes the classifier learnt might classify all points as negatives.

Some ways to address the issue

Subsample the negatives, and learn an ensemble of classifiers.

Introduce different costs for the positives and negatives

\[
\begin{align*}
\min_{w \in \mathbb{R}^p} & \quad \frac{1}{2} \|w\|_2^2 + C^+ \sum_{i \in P} \xi_i + C^- \sum_{i \in N} \xi_i \\
\text{s.t.} & \quad \forall i, y_i (w^\top x_i + b) \geq 1 - \xi_i
\end{align*}
\]

Naive choice:

\[ C^+ = \frac{C}{n_+}, \quad C^- = \frac{C}{n_-} \]

Is suboptimal in theory and in practice!!

→ Better to search for the optimal hyperparameter pair \((C^+, C^-)\).

SVM, kernel methods and multiclass 23/23
Imbalanced classification

Learn a binary classifier from \((x_i, y_i)\) pairs with
\[\mathcal{P} = \{i \mid y_i = 1\} \quad \mathcal{N} = \{i \mid y_i = -1\},\]
\[n_+ = |\mathcal{P}|, \quad n_- = |\mathcal{N}| \quad \text{and with} \quad n_+ \ll n_-.

**Problem:** to minimize the number of mistakes the classifier learnt might classify all points as negatives.

**Some ways to address the issue**

- Subsample the negatives, and learn an *ensemble* of classifiers.
Imbalanced classification

Learn a binary classifier from \((x_i, y_i)\) pairs with
\[ \mathcal{P} = \{i \mid y_i = 1\} \quad \mathcal{N} = \{i \mid y_i = -1\}, \]
\[ n_+ = |\mathcal{P}|, \quad n_- = |\mathcal{N}| \quad \text{and with} \quad n_+ \ll n_. \]

**Problem:** to minimize the number of mistakes the classifier learnt might classify all points as negatives.

**Some ways to address the issue**
- Subsample the negatives, and learn an *ensemble* of classifiers.
- Introduce different costs for the positives and negatives

\[
\min_{w \in \mathbb{R}^p} \frac{1}{2} \|w\|_2^2 + C_+ \sum_{i \in \mathcal{P}} \xi_i + C_- \sum_{i \in \mathcal{N}} \xi_i
\]

s.t. \( \forall i, \quad y_i (w^\top x_i + b) \geq 1 - \xi_i \)

Naive choice:
\[
C_+ = C / n_+ \quad \text{and} \quad C_- = C / n_-.
\]

Is suboptimal in theory and in practice!!

→ Better to search for the optimal hyperparameter pair \((C_+, C_-)\).
Imbalanced classification

Learn a binary classifier from \((x_i, y_i)\) pairs with 
\[ \mathcal{P} = \{ i \mid y_i = 1 \}, \quad \mathcal{N} = \{ i \mid y_i = -1 \}, \]

\[ n_+ = |\mathcal{P}|, \quad n_- = |\mathcal{N}| \quad \text{and with} \quad n_+ \ll n_- . \]

**Problem:** to minimize the number of mistakes the classifier learnt might classify all points as negatives.

**Some ways to address the issue**

- Subsample the negatives, and learn an *ensemble* of classifiers.
- Introduce different costs for the positives and negatives

\[
\min_{\mathbf{w} \in \mathbb{R}^p} \quad \frac{1}{2} \|\mathbf{w}\|^2_2 + C_+ \sum_{i \in \mathcal{P}} \xi_i + C_- \sum_{i \in \mathcal{N}} \xi_i \\
\text{s.t.} \quad \forall i, \quad y_i (\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i
\]

- Naive choice: \( C_+ = C/n_+ \) and \( C_- = C/n_- \)

⚠️ Is suboptimal in theory and in practice !!

→ Better to search for the optimal hyperparameter pair \((C_+, C_-)\).