Optimization Theory and Duality for SVMs

Outline:
• Tools for designing learning (training) algorithms.
• How to make the optimization problem more tractable?
• A dual representation of the optimal hyperplane in terms of the training examples.
• What insight do we gain from the dual representation?
• What are the properties of the dual optimization problem?

Quadratic Program

\[
\begin{align*}
\text{minimize} & \quad P(w) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j H_{ij} \\
\text{s.t.} & \quad \sum_{i=1}^{n} \alpha_i y_i = 0, \quad \alpha_i \geq 0, \quad i = 1, \ldots, n
\end{align*}
\]

• \( k \) linear inequality constraints
• \( m \) linear equality constraints
• Gram Matrix \( H=H_{(i,j)} \) is pos. semi-definite \( \forall w_1 \ldots w_n: \sum_{i=1}^{n} \alpha_i \alpha_j w_i w_j H_{ij} \geq 0 \)
  \( \Rightarrow \) convex, no local optima
• \( \hat{\alpha} \) is feasible, if it fulfills constraints

Fermat Theorem

Given an unconstrained optimization problem

\[ \text{minimize } P(w) \]

with \( P(w) \) convex, a necessary and sufficient conditions for a point \( \hat{w} \) to be an optimum is that

\[ \frac{\delta P(\hat{w})}{\delta w} = 0 \]

Lagrange Function

Given an optimization problem

\[ \begin{align*}
\text{minimize} & \quad P(w) \\
\text{s.t.} & \quad g_1(w) \leq 0, \quad \ldots, \quad g_k(w) \leq 0, \quad \alpha_i \geq 0, \quad \beta_i \geq 0, \\
& \quad h_1(w) = 0, \quad \ldots, \quad h_m(w) = 0
\end{align*} \]

the Lagrangian function is defined as

\[ L(\hat{w}, \hat{\alpha}, \hat{\beta}) = P(\hat{w}) + \sum_{i=1}^{k} \alpha_i g_i(\hat{w}) + \sum_{i=1}^{m} \beta_i h_i(\hat{w}) \]

• \( \hat{\alpha} \) and \( \hat{\beta} \) are called Lagrange Multipliers
**Lagrange Theorem**

Given an optimization problem

\[
\text{minimize } P(w) \\
\text{s.t. } h_1(w) = 0 \ldots h_m(w) = 0
\]

with \( P(w) \) convex and all \( h \) affine \((w^* x + b)\), necessary and sufficient conditions for a point \( \hat{w}^o \) to be an optimum are the existence of \( \hat{\beta}^o \) such that

\[
\frac{\delta L(w^o, \hat{\beta}^o)}{\delta w} = 0 \quad \frac{\delta L(w^o, \hat{\beta}^o)}{\delta \beta} = 0 \quad L(w, \hat{\beta}) = P(w) + \sum_{i=1}^m \hat{\beta}_i h_i(w)
\]

\[\Rightarrow L(w^o, \hat{\beta}) \leq L(w^o, \hat{\beta}^o) \leq L(w, \hat{\beta})\]

**Karush-Kuhn-Tucker Theorem**

Given an optimization problem

\[
\text{minimize } P(\vec{\alpha}) \\
\text{s.t. } g_j(\vec{\alpha}) \leq 0 \ldots g_k(\vec{\alpha}) \leq 0 \\
\quad h_1(\vec{\alpha}) = 0 \ldots h_m(\vec{\alpha}) = 0
\]

with \( P(\vec{\alpha}) \) convex and all \( g \) and \( h \) affine, necessary and sufficient conditions for a point \( \vec{\alpha}^o \) to be an optimum are the existence of \( \hat{\alpha}^o \) and \( \hat{\beta}^o \) such that

\[
\frac{\delta L(\vec{\alpha}^o, \hat{\alpha}^o, \hat{\beta}^o)}{\delta \alpha} = 0 \quad \frac{\delta L(\vec{\alpha}^o, \hat{\alpha}^o, \hat{\beta}^o)}{\delta \beta} = 0
\]

\[\alpha_i^o g_i(\vec{\alpha}^o) = 0, i = 1, \ldots, k \]

\[g_j(\vec{\alpha}^o) \leq 0, i = 1, \ldots, k \]

\[\alpha_i^o \geq 0, i = 1, \ldots, k \]

Sufficient for convex QP: \( \max_{\vec{\alpha} \geq 0, \beta} \left[ \min_w L(w, \vec{\alpha}^o, \hat{\beta}^o) \right] \)

---

**Dual Optimization Problem**

**Primal OP:** minimize \( P(\vec{w}, b) = \frac{1}{2} \vec{w} \cdot \vec{w}, \text{ with } \forall i \left[ y_i (\vec{w} \cdot \vec{x}_i + b) \right] \geq 1 \)

**Lemma:** The solution \( \vec{w}^o \) can always be written as a linear combination

\[\vec{w}^o = \sum_{i=1}^n \alpha_i y_i \vec{x}_i \quad \alpha_i \geq 0 \]

of the training data.

\[\Rightarrow\] Lagrange multipliers

**Dual OP:** maximize \( D(\vec{\alpha}) = \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{j=1}^n \sum_{i=1}^n \alpha_i \alpha_j y_i y_j \vec{x}_i \cdot \vec{x}_j \)

\[\text{s.t. } \sum_{i=1}^n \alpha_i y_i = 0 \quad \text{and} \quad 0 \leq \alpha_i \]

\[\Rightarrow\] positive semi-definite quadratic program

**Primal \(\Rightarrow\) Dual**

**Theorem:** The primal OP and the dual OP have the same solution. Given the solution \( \alpha_i^o \) of the dual OP,

\[
\vec{w}^o = \sum_{i=1}^n \alpha_i^o y_i \vec{x}_i \quad b^o = \frac{1}{2} (\vec{w}^o \cdot \vec{x}^{\text{pos}} + \vec{w}^o \cdot \vec{x}^{\text{neg}})
\]

is the solution of the primal OP.

**Theorem:** For any feasible points \( P(\vec{w}, b) \geq D(\vec{\alpha}) \).

\[\Rightarrow\] two alternative ways to represent the learning result

- weight vector and threshold \( \vec{a}^o, \beta \)
- vector of “influences” \( \alpha_1, \ldots, \alpha_n \)
**Properties of the Dual OP**

**Dual OP:** maximize  \( D(\alpha) = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) \)

s.t.  \( \sum_{i=1}^{n} \alpha_i = 0 \) and  \( 0 \leq \alpha_i \)

- single solution (i.e. \( \hat{w}, \hat{b} \) is unique)
- one factor \( \alpha_i \) for each training example
  - describes the “influence” of training examples \( i \) on the result
  - \( \alpha_i > 0 \) \( \iff \) training example is a support vector
  - \( \alpha_i = 0 \) else
- depends exclusively on inner product between training examples

**Properties of the Soft-Margin Dual OP**

**Dual OP:** maximize  \( D(\alpha) = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) \)

s.t.  \( \sum_{i=1}^{n} \alpha_i y_i = 0 \) and  \( 0 \leq \alpha_i \leq C \)

- (mostly) single solution (i.e. \( \hat{w}, \hat{b} \) is almost always unique)
- one factor \( \alpha_i \) for each training example
  - “influence” of single training example limited by \( C \)
  - \( 0 < \alpha_i < C \) \( \iff \) SV with \( \hat{\xi}_i = 0 \)
  - \( \alpha_i = C \) \( \iff \) SV with \( \hat{\xi}_i > 0 \)
  - \( \alpha_i = 0 \) else
- based exclusively on inner product between training examples