Lagrangian duality in nonlinear programming

Martin Branda

Charles University in Prague
Faculty of Mathematics and Physics
Department of Probability and Mathematical Statistics

Computational Aspects of Optimization
Primal problem (P):

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

Lagrangian function, \( u \in \mathbb{R}^m_+ \), \( v \in \mathbb{R}^l \):

\[
L(x, u, v) = f(x) + \sum_{j=1}^m u_j g_j(x) + \sum_{i=1}^l v_i h_i(x)
\]
Dual problem

Dual function:

$$\theta(u, v) = \inf_{x \in X} L(x, u, v)$$  \hspace{1cm} (1)

Dual problem (D):

$$(D) = \sup_{u \geq 0, v} \theta(u, v)$$  \hspace{1cm} (2)$$
Theorem

Let \( x \) be feasible for problem \((P)\) and \((u, v)\) be feasible for problem \((D)\). Then

\[
\theta(u, v) \leq f(x).
\]

Proof.

\[
\theta(u, v) = \inf_y L(y, u, v) \leq L(x, u, v) \leq f(x),
\]

where the last inequality follows from feasibility of \( x \) and \((u, v)\), when \( u_j g_j(x) \leq 0 \) and \( v_i h_i(x) = 0 \).
Weak Duality Theorem – Consequences

1. We obtain

\((P) \geq (D)\).

2. If for some primal feasible \(\bar{x}\) and dual feasible \((\bar{u}, \bar{v})\) holds

\[ f(\bar{x}) = \theta(\bar{u}, \bar{v}), \]

then \(\bar{x}\) is optimal solution of \((P)\) and \((\bar{u}, \bar{v})\) is optimal solution of \((D)\).

3. If \((P) = -\infty\) (unbounded primal problem), then \(\theta(u, v) = -\infty\) for all \((u, v) \in \mathbb{R}_+^m \times \mathbb{R}^l\).

4. If \((D) = \infty\), then \((P)\) is infeasible.
Strong Duality Theorem

**Theorem**

Let

- $X$ be a nonempty convex set
- $f, g_j$ be convex
- $h_i$ be affine
- **Slater condition** be satisfied, i.e. there is $\hat{x} \in X$ such that $g_j(\hat{x}) < 0, \forall j$ and $h_i(\hat{x}) = 0, \forall i$, and $0 \in \text{int}\{(h_1(x), \ldots, h_l(x)) : x \in X\} := h(X)$.

Then $(P) = (D)$.

Moreover, if $(P)$ is finite, then sup in $(D)$ is achieved at $(\bar{u}, \bar{v}) \in \mathbb{R}^m_+ \times \mathbb{R}^l$.

If inf in $(P)$ is achieved at $\bar{x}$, then $\sum_{j=1}^m \bar{u}_j g_j(\bar{x}) = 0$. 
A counterexample

Convexity alone is not sufficient. Consider

\[ p^* = \min_{x, y} e^{-x} \]
\[ \text{s.t. } x^2/y \leq 0, \]
\[ y > 0 \text{ (or } y \geq \varepsilon). \]

The optimal value is \( p^* = 1 \). The dual function is equal to

\[ \theta(u) = \inf_{x, y > 0} e^{-x} + u x^2/y = \begin{cases} 0 & u \geq 0, \\ -\infty & u < 0. \end{cases} \]

The dual problem is

\[ d^* = \max_{u \geq 0} \theta(u) \]

with optimal value \( d^* = 0 \). Slater condition is not satisfied since \( x = 0 \) for any feasible \((x, y)\), i.e. \( x^2/y = 0 \).
Bazaraa et al. (2006), Lemma 6.2.3:

**Lemma**

Let $X \subseteq \mathbb{R}^n$ be a convex set, $f, g_j : \mathbb{R}^n \to \mathbb{R}$ be convex, $h_i : \mathbb{R}^n \to \mathbb{R}$ be affine. If System 1 has no solution, then System 2 has a solution $(u_0, u, v)$. The converse holds true if $u_0 > 0$.

**System 1:** $f(x) < 0$, $g_j(x) \leq 0$, $h_i(x) = 0$ for some $x \in X$.

**System 2:** $u_0 f(x) + \sum_{j=1}^{m} u_j g_j(x) + \sum_{i=1}^{l} v_i h_i(x) \geq 0$ for all $x \in X$, $(u_0, u) \geq 0$, $(u_0, u, v) \neq 0$. 
Let $\gamma$ be a (finite) optimal value of (P) and consider the following system:

\[ f(x) - \gamma < 0, \quad g_j(x) \leq 0, \quad j = 1, \ldots, m, \quad h_i(x) = 0, \quad i = 1, \ldots, l, \quad x \in X. \]

By the definition of $\gamma$ the system has no solution. Hence, there exists $(u_0, u, v) \neq 0$ with $(u_0, u) \geq 0$ such that

\[ u_0(f(x) - \gamma) + \sum_{j=1}^{m} u_j g_j(x) + \sum_{i=1}^{l} v_i h_i(x) \geq 0, \quad \forall x \in X. \]
Suppose that $u_0 = 0$. By assumption there is an $\hat{x} \in X$ such that $g_j(\hat{x}) < 0$, $\forall j$ and $h_i(\hat{x}) = 0$, $\forall i$. Substituting into the inequality we obtain $\sum_{j=1}^m u_j g_j(\hat{x}) \geq 0$. Since $g_j(\hat{x}) < 0$, $\forall j$, we have $u_j = 0$, $\forall j$, and $u_0 = 0$. This implies that $\sum_{i=1}^l v_i h_i(x) \geq 0$ for all $x \in X$. Since $0 \in h(X)$, we can pick a $x \in X$ such that $h_i(x) = -\lambda v_i$, where $\lambda > 0$ (small). Therefore

$$\sum_{i=1}^l v_i h_i(x) = -\lambda \sum_{i=1}^l v_i^2 \geq 0,$$

which implies that $v_i = 0$, $\forall i$. But this is a contradiction with $(u_0, u, v) \neq 0$. Hence $u_0 > 0$. ...
Hence $u_0 > 0$. Thus, if we set $\tilde{u}_j = u_j/u_0$ and $\tilde{v}_i = v_i/u_0$, we get

$$f(x) + \sum_{j=1}^{m} \tilde{u}_j g_j(x) + \sum_{i=1}^{l} \tilde{v}_i h_i(x) \geq \gamma, \quad \forall x \in X.$$ 

This shows that

$$\theta(\tilde{u}, \tilde{v}) = \inf_{x \in X} L(x, \tilde{u}, \tilde{v}) \geq \gamma.$$ 

Together with the Weak Duality Theorem we obtain that

$$\gamma = \theta(\tilde{u}, \tilde{v}) = \sup_{u \geq 0, \nu} \theta(u, \nu).$$
Example: Linear programming duality

\[ \begin{align*}
\min & \quad c^T x \\
\text{s.t.} & \quad Ax = b, \\
& \quad x \geq 0.
\end{align*} \]
Example: Ordinary least squares with equality constraints

\[ \min \| Ax - b \|_2^2 \]
\[ \text{s.t. } Fx = g. \]
Hastie et al. (2009): Training data: \( N \) pairs \((x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\), \( x_i \in \mathbb{R}^p, y_i \in \{-1, 1\} \) (classes).
A linear classification rule with \( \|\beta\| = 1 \)

\[
G(x) = \text{sign}[x^T \beta + \beta_0].
\]

Assume first that the data are separable. We would like to find the biggest margin between the training points for class 1 and \(-1\):

\[
\max_{\beta_0, \beta} \ M \\
\text{s.t. } y_i(x_i^T \beta + \beta_0) \geq M, \ i = 1, \ldots, N, \\
\|\beta\| = 1.
\]
Example: The support vector classifier

Hastie et al. (2009)
Example: The support vector classifier

By setting $M = 1/\|\beta\|$:

$$\begin{align*}
\min_{\beta_0, \beta} & \quad \|\beta\| \\
\text{s.t.} & \quad y_i(x_i^T \beta + \beta_0) \geq 1, \quad i = 1, \ldots, N.
\end{align*}$$

If the classes overlap:

$$\begin{align*}
\min_{\beta_0, \beta, \xi} & \quad \frac{1}{2} \|\beta\|^2 + C \sum_{i=1}^{N} \xi_i \\
\text{s.t.} & \quad y_i(x_i^T \beta + \beta_0) \geq 1 - \xi_i, \quad i = 1, \ldots, N, \\
& \quad \xi_i \geq 0,
\end{align*}$$

where we penalize the overall overlap.
Example: The support vector classifier

Lagrange function

\[ L(\beta_0, \beta, \xi, \alpha, \mu) = \frac{1}{2} \|\beta\|^2 + C \sum_{i=1}^{N} \xi_i - \sum_{i=1}^{N} \mu_i \xi_i \]

\[ - \sum_{i=1}^{N} \alpha_i (y_i (x_i^T \beta + \beta_0) - 1 + \xi_i), \quad \alpha_i \geq 0, \mu_i \geq 0. \]

The dual function

\[ \theta(\alpha, \mu) = \inf_{\beta_0, \beta, \xi} L(\beta_0, \beta, \xi, \alpha, \mu). \]
Example: The support vector classifier

\[ L(\beta_0, \beta, \xi, \alpha, \mu) = \frac{1}{2} \|\beta\|^2 + C \sum_{i=1}^{N} \xi_i - \sum_{i=1}^{N} \mu_i \xi_i \]

\[- \sum_{i=1}^{N} \alpha_i(y_i(x_i^T \beta + \beta_0) - 1 + \xi_i), \alpha_i \geq 0, \mu_i \geq 0 \]

Use the derivatives to obtain the dual function:

\[ \frac{\partial L}{\partial \beta_0} = \sum_{i=1}^{N} \alpha_i y_i = 0, \]

\[ \frac{\partial L}{\partial \beta} = \beta - \sum_{i=1}^{N} \alpha_i y_i x_i = 0, \]

\[ \frac{\partial L}{\partial \xi_i} = C - \alpha_i - \mu_i = 0. \]
Example: The support vector classifier

We can express the dual function

\[
\theta(\alpha, \mu) = \frac{1}{2} \sum_{i=1}^{N} \sum_{i'=1}^{N} \alpha_i \alpha_{i'} y_i y_{i'} x_i^T x_{i'} + C \sum_{i=1}^{N} \xi_i - \sum_{i=1}^{N} \sum_{i'=1}^{N} \alpha_i \alpha_{i'} y_i y_{i'} x_i^T x_{i'} - \beta_0 \sum_{i=1}^{N} \alpha_i y_i - \sum_{i=1}^{N} \alpha_i - \sum_{i=1}^{N} \alpha_i \xi_i - \sum_{i=1}^{N} \mu_i \xi_i
\]

\[
= -\frac{1}{2} \sum_{i=1}^{N} \sum_{i'=1}^{N} \alpha_i \alpha_{i'} y_i y_{i'} x_i^T x_{i'} + \sum_{i=1}^{N} \alpha_i,
\]

subject to \(0 \leq \alpha_i \leq C\), \(\sum_{i=1}^{N} \alpha_i y_i = 0\).
