Optimization for Machine Learning (CSL4010)

Dr. Md Abu Talhamainuddin Ansary
Department of Mathematics, IIT Jodhpur

PPT 4

Duality Theory

- General form of an optimization problem:

$$
\begin{aligned}
(P) : \min_{x \in \mathbb{R}^n} f(x) & & & \\
s.\ t.\ \ g_i(x) &\leq& 0 & \quad i = 1, 2, \ldots, m \\
h_j(x) &=& 0 & \quad j = 1, 2, \ldots, p
\end{aligned}
$$

- Dual problem corresponding to (P) is defined as

$$
\begin{aligned}
(D) : \max_{(\theta, \mu) \in \mathbb{R}^m \times \mathbb{R}^p} \quad & \theta(\lambda, \mu) \\
s.\ t.\ \ & \lambda \geq 0
\end{aligned}
$$

where

$$
theta(\lambda, \mu) = \inf_x \; L(x; \lambda, \mu)
$$

$$
L(x; \lambda, \mu) = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x) + \sum_{j=1}^{p} \mu_j h_j(x)
$$

- x is called primal variable and (λ, μ) is called dual variables.
- The function $L(x; \lambda, \mu)$ is known as Lagrangian function.

- One can observe that the dimension of decision variable in dual problem is same as the number of constraints in primal problem.

- One can observe that the dimension of decision variable in dual problem is same as the number of constraints in primal problem.
- **Weak duality theory:** Suppose $\bar{x}$ and $(\bar{\lambda}, \bar{\mu})$ are feasible points of (P) and (D) respectively. Then $f(\bar{x}) \geq \theta(\bar{\lambda}, \bar{\mu})$.

- One can observe that the dimension of decision variable in dual problem is same as the number of constraints in primal problem.
- **Weak duality theory:** Suppose $\bar{x}$ and $(\bar{\lambda}, \bar{\mu})$ are feasible points of (P) and (D) respectively. Then $f(\bar{x}) \geq \theta(\bar{\lambda}, \bar{\mu})$.
  Proof: Since $\bar{x}$ is a feasible point of (P), $g_i(\bar{x}) \leq 0$ for all i and $h_j(\bar{x}) = 0$ for all j. Similarly, as $\bar{\lambda}$ is feasible for (D) implies $\bar{\lambda}_i \geq 0$. Now

$$\begin{aligned} \theta(\bar{\lambda}, \bar{\mu}) &= \inf_x \{f(x) + \sum_{i=1}^m \bar{\lambda}_i g_i(x) + \sum_{j=1}^p \bar{\mu}_j h_j(x)\} \\ &\leq f(\bar{x}) + \sum_{i=1}^m \bar{\lambda}_i g_i(\bar{x}) + \sum_{j=1}^p \bar{\mu}_j h_j(\bar{x}) \\ &\leq f(\bar{x}). \end{aligned}$$

  The last inequality follows since $h_j(\bar{x}) = 0$ for all j and $\bar{\lambda}_i g_i(\bar{x}) \leq 0$ for all i.
  Hence $f(\bar{x}) \geq \theta(\bar{\lambda}, \bar{\mu})$.

- One can observe that the dimension of decision variable in dual problem is same as the number of constraints in primal problem.
- **Weak duality theory:** Suppose $\bar{x}$ and $(\bar{\lambda}, \bar{\mu})$ are feasible points of (P) and (D) respectively. Then $f(\bar{x}) \geq \theta(\bar{\lambda}, \bar{\mu})$.
  Proof: Since $\bar{x}$ is a feasible point of (P), $g_i(\bar{x}) \leq 0$ for all i and $h_j(\bar{x}) = 0$ for all j. Similarly, as $\bar{\lambda}$ is feasible for (D) implies $\bar{\lambda}_i \geq 0$. Now

$$\begin{aligned}\theta(\bar{\lambda}, \bar{\mu}) &= \inf_x \{f(x) + \sum_{i=1}^{m} \bar{\lambda}_i g_i(x) + \sum_{j=1}^{p} \bar{\mu}_j h_j(x)\} \\ &\leq f(\bar{x}) + \sum_{i=1}^{m} \bar{\lambda}_i g_i(\bar{x}) + \sum_{j=1}^{p} \bar{\mu}_j h_j(\bar{x}) \\ &\leq f(\bar{x}).\end{aligned}$$

  The last inequality follows since $h_j(\bar{x}) = 0$ for all j and $\bar{\lambda}_i g_i(\bar{x}) \leq 0$ for all i.
  Hence $f(\bar{x}) \geq \theta(\bar{\lambda}, \bar{\mu})$.
- **Strong duality theory:** Suppose x^* and (λ^*, μ^*) are feasible points of (P) and (D) respectively and $f(x^*) = \theta(\lambda^*, \mu^*)$. Then x^* is the optimal solution of (P) and (λ^*, μ^*) is the optimal solution of (D).

- Consider the LP:

$$\begin{aligned} \min_x \quad & c^T x \\ s.t. \;\; Ax &\leq b \\ Hx &= h \\ x \geq 0 \end{aligned}$$

- The Lagrangian function is

$$\begin{aligned} L(x;\lambda^1,\lambda^2,\mu) &= c^T x + {\lambda^1}^T(Ax-b) + \mu^T(Hx-h) - {\lambda^2}^T x \\ &= -b^T\lambda^1 - h^T\mu + \left(c + A^T\lambda^1 + H^T\mu - \lambda^2\right)^T x \end{aligned}$$

- So

$$\begin{aligned} \theta(\lambda^1,\lambda^2,\mu) &= \inf_x L(x;\lambda^1,\lambda^2,\mu) \\ &= \begin{cases} -b^T\lambda^1 - h^T\mu \text{ if } c + A^T\lambda^1 + H^T\mu - \lambda^2 = 0 \\ -\infty \text{otherwise} \end{cases} \end{aligned}$$

- So the dual problem is

$$\max_{\lambda^1,\lambda^2,\mu} \quad \theta(\lambda^1,\lambda^2,\mu)$$
$$s.t.\ \lambda^1,\lambda^2 \quad \geq \quad 0$$

- This is equivalent to the following LP:

$$\max_{\lambda^1,\mu} \quad -b^T x - h^T \mu$$
$$s.t.\ A^T\lambda^1 + H^T\mu \quad \geq \quad -c$$
$$\lambda^1 \geq 0$$

- Suppose f^* is the optimum value of (P) and θ^* is the optimum value of (D), then $f^* - \theta^*$ is the duality gap.
- If $f^* = \theta^*$ then we say zero duality gap.
- Suppose f^* is the optimum value of (P) and θ^* is the optimum value of (D) and $f^* = \theta^*$ (i.e. zeros duality gap) iff there exists x^* feasible for (P) and (λ^*, μ^*) feasible for (D) such that

$$L(x^*, \lambda, \mu) \leq L(x^*, \lambda^*, \mu^*) \leq L(x, \lambda^*, \mu^*) \tag{1}$$

  for all (λ, μ) feasible for (D) and x feasible for (P).
- (x^*, λ^*, μ^*) satisfying (**??**) is said to be a Lagrangian saddle point.

- Suppose x^* and (λ^*, μ^*) is the optimal solution of (P) and (D) respectively and $f(x^*) = \theta(\lambda^*, \mu^*)$
- Then

$$\begin{aligned}
f(x^*) &= \theta(\lambda^*, \mu^*) \\
&= \inf_x \{f(x) + \sum_{i=1}^{m} \lambda_i^* g_i(x) + \sum_{j=1}^{p} \mu_j^* h_j(x)\} \\
&\leq f(x^*) + \sum_{i=1}^{m} \lambda_i^* g_i(x^*) + \sum_{j=1}^{p} \mu_j^* h_j(x^*) \\
&= f(x^*) + \sum_{i=1}^{m} \lambda_i^* g_i(x^*) \ \ (since \ \ h_j(x^*) = 0 \ \ \forall j). \qquad (2)
\end{aligned}$$

- Note that $g_i(x^*) \leq 0$ and $\lambda_i^* \geq 0$ for all i. So $\sum_{i=1}^m \lambda_i^* g_i(x^*) \leq 0$.
- If $\sum_{i=1}^m \lambda_i^* g_i(x^*) < 0$ then (**??**) implies $f(x^*) < f(x^*)$, a contradiction. Hence $\sum_{i=1}^m \lambda_i^* g_i(x^*) = 0$.
- Since $\lambda_i^* g_i(x^*) \leq 0$ for all i, $\sum_{i=1}^m \lambda_i^* g_i(x^*) = 0$ implies $\lambda_i^* g_i(x^*) = 0$ for all i.
- The condition $\lambda_i^* g_i(x^*) = 0$ for all i is known as complementary slackness condition.
- From complementary slackness condition, if $g_{\bar{i}}(x^*) < 0$ for some $\bar{i}$ then $\lambda_{\bar{i}}^* = 0$.
- Similarly if $\lambda_{\bar{i}}^* > 0$ for some $\bar{i}$ then $g_{\bar{i}}(x^*) = 0$.

- From second inequality of (**??**), x^* is a minima of $L(x; \lambda^*, \mu^*)$.
- If the primal problem is a convex optimization problem, so does $L(x; \lambda^*, \mu^*)$. Then $x^* = arg\ \min_{x \in \mathbb{R}^n}\ L(x; \lambda^*, \mu^*)$ implies

$$\nabla_x L(x^*, \lambda^*, \mu^*) = 0.$$

- This implies

$$\nabla f(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla g_i(x^*) + \sum_{i=1}^{p} \mu_j^* \nabla h_j(x^*) = 0$$

- Since x^* is a feasible point of (P), $g_i(x^*) \leq 0\ \forall\ i$ and $h_j(x^*) = 0\ \forall\ j$.
- Since (λ^*, μ^*) is feasible for (D), $\lambda_i^* \geq 0$ for all i.
- Also x^*, λ^* satisfy complementary slackness conditions. i.e. $\lambda_i^* g_i(x^*) = 0\ \forall\ i$.

KKT optimality conditions

- Above conditions can be written in together as

$$
\begin{aligned}
\nabla f(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla g_i(x^*) + \sum_{i=1}^{p} \mu_j^* \nabla h_j(x^*) &= 0 && (3)\\
g_i(x^*) &\leq 0 \quad \forall\ i && (4)\\
h_j(x^*) &= 0 \quad \forall\ j && (5)\\
\lambda_i^* \geq 0 \quad \lambda_i^* g_i(x^*) &= 0 \quad \forall\ i && (6)
\end{aligned}
$$

- The conditions (**??**)-(**??**) are known as Karush-Kuhn-Tucker (KKT) optimality condition.
- If (P) is a convex optimization problem and satisfies Slater condition, then x^* is a local minima of (P) iff there exists (λ^*, μ^*) such that $(x^*; \lambda^*, \mu^*)$ satisfies KKT optimality conditions (**??**)-(**??**).

- We need to find $\lambda^* = (\lambda_1^*, \lambda_2^*, \lambda_3^*)^T \geq 0$ and μ_1^* such that

$$\nabla f(x^*) + \sum_{i=1}^{3} \lambda_i^* \nabla g_i(x^*) + \mu_1^* \nabla h_j(x^*) = 0 \quad (7)$$

$$\lambda_i^* g_i(x^*) = 0 \ \forall \ \ i = 1, 2, 3. \quad (8)$$

  where $g_1(x) = x_1^2 + x_2^2 - 5$, $g_2(x) = -x_1$, $g_3(x) = -x_2$.
- Note that $g_2(x^*) = -2 < 0$ and $g_3(x^*) = -1 < 0$. Hence from (**??**), $\lambda_2^* = 0 = \lambda_3^*$.
- So from (**??**),

$$\nabla f(x^*) + \lambda_1^* \nabla g_i(x^*) + \mu_1^* \nabla h_j(x^*) = 0$$

- This implies
$$\begin{bmatrix} -2 \\ -2 \end{bmatrix} + \lambda_1^* \begin{bmatrix} 4 \\ 2 \end{bmatrix} + \mu_1^* \begin{bmatrix} 1 \\ 2 \end{bmatrix} = 0$$
- Solution of the above system of equations is $\lambda_1^* = 1/3 > 0$ and $\mu_1^* = 2/3$.
- Hence we can find $\lambda^* = (1/3, 0/0)^T$ and $\mu^* = 2/3$ satisfying (**??**)-(**??**). So x^* is a KKT point of this problem.
- Since the given problem is a convex optimization problem (objective function and inequality constraint functions are convex and equality constraint is affine). $x^* = (2, 1)^T$ is a global minima of this problem.

Duality theory contd...

- Consider the equality constrained problem:

$$\begin{aligned} \min \quad & f(x) \\ s.t. \;\; h_j(x) = 0 \text{ `}j = & 1, 2, \ldots, p \end{aligned}$$

- x^* is a KKT point of this problem if

$$\begin{aligned} \nabla f(x^*) + \sum_{j=1}^{p} \mu_j \nabla g_j(x^*) &= 0 \\ h_j(x^*) &= 0 \quad j = 1, 2, \ldots, p \end{aligned}$$

- This is a system of $n + p$ equations with $n + p$ unknowns.
- If the problem is convex then solving this we can find the optima. This technique is known as Lagrange multiplier method.