# Class 6: Applications of duality theory

In this class, we use the duality theory in two different contexts: nonlinear integer programming and support vector regression.

### Example 1. Nonlinear integer programming

Consider the following *nonlinear integer optimization* problem:

$$\begin{array}{ll} \min & f(x) \\ \text{s.t.} & x \in X, \\ & x_i \in \{0,1\}, \quad i = 1, \ldots, n, \end{array}$$

where $X \subseteq \mathbb{R}^n$, $f\colon \mathbb{R}^n \to \mathbb{R}$ is nonlinear, and the variables $x_i$ are either zero or one. If $f$ is linear and $X$ is a polyhedron, then the problem is called *0-1 integer programming*, which is known to be NP-complete. This means that the above more general problem is also difficult. One way to solve these kind of problems is by using the so-called *branch-and-bound* technique. This topic will be seen in the second part of the class. For now, we only say that in such a method, it is important to find good lower bounds of the original problem. Thus, define $f^*_{\text{orig}}$ as the optimal value of the above problem (assuming that it is finite). If we consider the following relaxation of the above problem:

$$\begin{array}{ll} \min & f(x) \\ \text{s.t.} & x \in X, \\ & 0 \leq x_i \leq 1, \quad i = 1, \ldots, m, \end{array}$$

and defining its optimal value as $f^*_{\text{relax}}$, we have

$$f^*_{\text{relax}} \leq f^*_{\text{orig}},$$

because the feasible set of the relaxation problem contains the feasible set of the original problem. Now, consider the dual of the relaxation problem. From the weak duality theorem, if we have a feasible dual point, then we get a lower bound $\hat{\omega}$ of the relaxation problem, i.e.,

$$\hat{\omega} \leq f^*_{\text{relax}}.$$

Trivially, $\hat{\omega}$ is also a lower bound for the original problem. Since we only need a feasible point of the dual (instead of the optimal solution of the relaxation problem), it tends to be easier to find the required lower bound in this case.

### Example 2. Support vector regression

Given $m$ data $x^i \in \mathbb{R}^n$ (input) and target values $y_i \in \mathbb{R}$, the *support vector regression* consists in finding the best estimate function $f\colon \mathbb{R}^n \to \mathbb{R}$ such that $f(x^i) = y_i$, $i = 1, \ldots, m$. In other

words, it tries to find a model that can explain the output by given the input data. Assume, for instance, that the model, that is called *regression function*, is linear, i.e.,

$$f(x) = w^\top x + b,$$

for some $w \in \mathbb{R}^n$ and $b \in \mathbb{R}$. Now, given $\varepsilon > 0$, define $\ell_\varepsilon \colon \mathbb{R} \to \mathbb{R}$ as

$$\ell_\varepsilon(t) := \begin{cases} 0, & \text{if } t \in [-\varepsilon, \varepsilon], \\ |t| - \varepsilon, & \text{otherwise.} \end{cases} \tag{1}$$

To find the best fit, we want to minimize the difference between the estimate $f(x^i)$ and the real output $y^i$, by assuming that the an estimate error exits if the difference is greater than $\varepsilon$. More precisely, we want to solve the following problem:

$$\begin{aligned} \min \quad & \sum_{i=1}^{m} \ell_\varepsilon(w^\top x^i + b - y^i) \\ \text{s.t.} \quad & w \in \mathbb{R}^n,\ b \in \mathbb{R}. \end{aligned}$$

By solving this, we find the optimal solutions $w^*$ and $b^*$, and consequently the regression $f$. However, it is possible that there exist some noises in the data, which can disturb this problem's solution considerably. This is called overfitting, and to avoid it, we can consider instead the following problem:

$$\begin{aligned} \min \quad & \sum_{i=1}^{m} \ell_\varepsilon(w^\top x^i + b - y^i) + C\|w\|^2 \\ \text{s.t.} \quad & w \in \mathbb{R}^n,\ b \in \mathbb{R}, \end{aligned}$$

where $C > 0$ is a scalar. The term $C\|w\|^2$ is called *regularization*. Since the function $\ell_\varepsilon$ is nondifferentiable, the above problem is not the usual nonlinear programming that we were considering during the class. However, by adding extra variables $z_i$, $i = 1, \ldots, m$, we can make it differentiable. In fact, first, observe that the above problem is equivalent to

$$\begin{aligned} \min \quad & \sum_{i=1}^{m} z_i + C\|w\|^2 \\ \text{s.t.} \quad & \ell_\varepsilon(w^\top x^i + b - y^i) \le z_i, \quad i = 1, \ldots, m, \\ & w \in \mathbb{R}^n,\ b \in \mathbb{R},\ z \in \mathbb{R}^m. \end{aligned}$$

From the definition (1), we obtain another equivalent problem that is differentiable:

$$\begin{aligned} \min \quad & \sum_{i=1}^{m} z_i + C\|w\|^2 \\ \text{s.t.} \quad & -z_i - \varepsilon \le w^\top x^i + b - y^i \le z_i + \varepsilon, \quad i = 1, \ldots, m, \\ & w \in \mathbb{R}^n,\ b \in \mathbb{R},\ z \in \mathbb{R}^m. \end{aligned}$$

Note that it is in fact a quadratic programming problem. Now, observe that we can replace the linear regression by an arbitrary nonlinear estimate function, by assuming that

$$f(x) = w^\top \phi(x) + b,$$

with some $\phi\colon \mathbb{R}^n \to \mathbb{R}^n$. In this case, the support vector regression problem is given by

$$\begin{array}{ll} \min & \displaystyle\sum_{i=1}^m z_i + C\|w\|^2 \\ \text{s.t.} & -z_i - \varepsilon \le w^\top \phi(x^i) + b - y^i \le z_i + \varepsilon,\ i = 1, \dots, m, \\ & w \in \mathbb{R}^n,\ b \in \mathbb{R},\ z \in \mathbb{R}^m, \end{array} \tag{2}$$

which is also a quadratic programming problem. Now, let us find its dual problem. The Lagrange function associated to (2) is given as follows:

$$\begin{aligned} L(w, b, z, \lambda^{(1)}, \lambda^{(2)}) &= \sum_{i=1}^m z_i + C\|w\|^2 + \sum_{i=1}^m \lambda_i^{(1)}\big(w^\top \phi(x^i) + b - y^i - z_i - \varepsilon\big) \\ &\quad + \sum_{i=1}^m \lambda_i^{(2)}\big(-w^\top \phi(x^i) - b + y^i - z_i - \varepsilon\big) \\ &= C\|w\|^2 + \sum_{i=1}^m (\lambda_i^{(1)} - \lambda_i^{(2)})(w^\top \phi(x^i) + b - y_i) \\ &\quad + \sum_{i=1}^m (1 - \lambda_i^{(1)} - \lambda_i^{(2)}) z_i - \sum_{i=1}^m (\lambda_i^{(1)} + \lambda_i^{(2)})\varepsilon, \end{aligned}$$

where $\lambda^{(1)} \in \mathbb{R}^m$ and $\lambda^{(2)} \in \mathbb{R}^m$ are the Lagrange multipliers. Let us recall that the dual function is given by

$$\omega(\lambda^{(1)}, \lambda^{(2)}) = \inf_{(w,b,z) \in \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^m} L(w, b, z, \lambda^{(1)}, \lambda^{(2)}), \tag{3}$$

and that the dual feasible set is given by

$$D = \big\{(\lambda^{(1)}, \lambda^{(2)}) \in \mathbb{R}_+^m \times \mathbb{R}_+^m \mid \omega(\lambda^{(1)}, \lambda^{(2)}) > -\infty\big\}.$$

Let us analyze the three variables $w, b, z$ separately.

(a) If $\sum_{i=1}^m (\lambda_i^{(1)} - \lambda_i^{(2)}) \neq 0$, then we can take $b$ sufficiently small, and so the Lagrangian function goes to $-\infty$.

(b) Similarly, if $1 - \lambda_i^{(1)} - \lambda_i^{(2)} \neq 0$, then we can take $z_i$ sufficiently small, and so the Lagrangian function goes to $-\infty$.

(c) The function $L$ is quadratic and convex with respect to the variable $w$. Then, from the optimality conditions, the solution $(w^*, b^*, z^*)$ of (3) satisfies

$$\nabla_w L(w^*, b^*, z^*, \lambda^{(1)}, \lambda^{(2)}) = 2Cw^* + \sum_{i=1}^m (\lambda_i^{(1)} - \lambda_i^{(2)})\phi(x^i) = 0,$$

which gives

$$w^* = -\frac{1}{2C} \sum_{i=1}^m (\lambda_i^{(1)} - \lambda_i^{(2)})\phi(x^i).$$

From (a) and (b), we obtain

$$D = \left\{ (\lambda^{(1)}, \lambda^{(2)}) \in \mathbb{R}^m_+ \times \mathbb{R}^m_+ \;\middle|\; \sum_{i=1}^m (\lambda_i^{(1)} - \lambda_i^{(2)}) = 0, \text{ and } 1 - \lambda_i^{(1)} - \lambda_i^{(2)} = 0, i = 1, \ldots, m \right\}.$$

Moreover, from (c), we get

$$L(w^*, b^*, z^*, \lambda^{(1)}, \lambda^{(2)}) = -\frac{1}{4C} \left\| \sum_{i=1}^m (\lambda_i^{(1)} - \lambda_i^{(2)}) \phi(x^i) \right\|^2 - \sum_{i=1}^m (\lambda_i^{(1)} - \lambda_i^{(2)}) y_i - \sum_{i=1}^m (\lambda_i^{(1)} + \lambda_i^{(2)}) \varepsilon.$$

Thus, the dual problem can be written as follows:

$$\begin{aligned} \max \quad & -\frac{1}{4C} \left\| \sum_{i=1}^m (\lambda_i^{(1)} - \lambda_i^{(2)}) \phi(x^i) \right\|^2 - \sum_{i=1}^m (\lambda_i^{(1)} - \lambda_i^{(2)}) y_i - \sum_{i=1}^m (\lambda_i^{(1)} + \lambda_i^{(2)}) \varepsilon \\ \text{s.t.} \quad & \sum_{i=1}^m (\lambda_i^{(1)} - \lambda_i^{(2)}) = 0, \\ & 1 - \lambda_i^{(1)} - \lambda_i^{(2)} = 0, \;\; i = 1, \ldots, m, \\ & \lambda^{(1)} \geq 0, \;\; \lambda^{(2)} \geq 0. \end{aligned} \tag{4}$$

Observe that this dual problem is also a quadratic concave programming (reformulating as a minimization problem, we say convex programming). Furthermore, if the optimal solution of the dual is given by $(\hat{\lambda}^{(1)}, \hat{\lambda}^{(2)})$, the primal solution can be written easily as

$$w^* = -\frac{1}{2C} \sum_{i=1}^m (\hat{\lambda}_i^{(1)} - \hat{\lambda}_i^{(2)}) \phi(x^i).$$

Also, by finding $b^*$ such that $|f(x^i) - y^i| = \varepsilon$ for all the support vectors $(x^i, y_i)$, with $f(x) = (w^*)^\top \phi(x) + b^*$, we obtain the whole estimate function $f$. Let us now define

$$K(x^i, x^j) := \phi(x^i)^\top \phi(x^j) \quad \text{for all } i, j = 1, \ldots, m.$$

Then, the first term of the dual objective function (4) can be written as

$$-\frac{1}{4C} \left\| \sum_{i=1}^m (\lambda_i^{(1)} - \lambda_i^{(2)}) \phi(x^i) \right\|^2 = -\frac{1}{4C} \sum_{i=1}^m \sum_{j=1}^m (\lambda_i^{(1)} - \lambda_i^{(2)})(\lambda_j^{(1)} - \lambda_j^{(2)}) K(x^i, x^j).$$

This means that the dual objective function can be written in terms of $K$ instead of $\phi$. Similarly, since

$$f(x) = (w^*)^\top \phi(x) + b^* = -\frac{1}{2C} \sum_{i=1}^m (\hat{\lambda}_i^{(1)} - \hat{\lambda}_i^{(2)}) K(x^i, x) + b^*$$

holds, the regression function $f$ can also be written in terms of $K$. Because of this, many approaches use this function $K$, called *Kernel function*. There exist many Kernel functions in the literature, but in general it is chosen to make the dual objective function concave.

---

Operations Research, Advanced (Graduate School of Informatics, Kyoto University)
1st part by Ellen H. Fukuda (e-mail: ellen(at)i.kyoto-u.ac.jp, where (at) = @)