

Optimization

Màster de Fonaments de Ciència de Dades

Gerard Gómez

Lecture X. Duality

1. Duality in linear programming
2. Duality in convex programming

Duality in linear programming

X.1 Duality in linear programming

Definition.

Duality is a concept that links the two following linear programming problems (LPP):

Minimize $c^T x$ subject to $Ax \geq b$ $x \geq 0$ **Primal problem**

Maximize $y^T b$ subject to $y^T A \leq c$ $y \geq 0$ **Dual problem**

where A is a $n \times m$ matrix, $x, c \in \mathbb{R}^m$ and $y, b \in \mathbb{R}^n$

- ▶ Is easy to check that *the dual of the dual is the primal* (by transforming minima to maxima and reversing inequalities by appropriately using minus sign).
- ▶ It must be noted that the primal and dual problems **are different, but equivalent**, ways of looking at the same underlying problem.

Duality in linear programming

Assume that the primal problem is formulated in the **standard form**

$$\text{Minimize } c^T x \quad \text{subject to } Ax = b \quad x \geq 0 \quad \textbf{Primal problem}$$

Clearly

$$Ax = b \quad \text{is equivalent to} \quad Ax \geq b, \quad -Ax \geq -b$$

so, if we write

$$\bar{A} = (A \mid -A) \in \mathbb{R}^{n \times 2m}, \quad \bar{b} = \begin{pmatrix} b \\ -b \end{pmatrix} \in \mathbb{R}^{2n}$$

the initial LPP becomes

$$\text{Minimize } c^T x \quad \text{subject to } \bar{A} \begin{pmatrix} x \\ x \end{pmatrix} \geq \bar{b} \quad x \geq 0$$

Therefore, according to the definition, its dual will have the form

$$\text{Maximize } \bar{y}^T \bar{b} \quad \text{subject to } \bar{y}^T \bar{A} \leq c \quad \bar{y} \geq 0$$

Introducing

$$\bar{y} = \begin{pmatrix} y^{(1)} \\ y^{(2)} \end{pmatrix}$$

we get

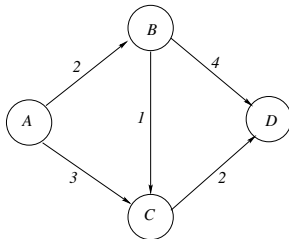
$$\text{Maximize } (y^{(1)} - y^{(2)})^T b \quad \text{subject to } (y^{(1)} - y^{(2)})^T A \leq c \quad \bar{y} \geq 0$$

and letting $y = y^{(1)} - y^{(2)}$, there is no restriction on the sign of y , and we have

$$\text{Maximize } y^T b \quad \text{subject to } y^T A \leq c \quad \textbf{Dual problem}$$

Duality in linear programming. Example

We wish to send a certain product from node A to node D in the network of the figure. There are 5 different channels with different associated costs



If we use variables x_{PQ} to denote the fraction of the product transferred through channel PQ , we want to **minimize the total cost**

$$2x_{AB} + 3x_{AC} + x_{BC} + 4x_{BD} + 2x_{CD}$$

subject to the restrictions

$$x_{AB} = x_{BC} + x_{BD} \quad (\text{no part of the product is lost at node } B)$$

$$x_{AC} + x_{BC} = x_{CD} \quad (\text{no part of the product is lost at node } C)$$

$$x_{BD} + x_{CD} = 1 \quad (\text{the total amount of the product reaches node } D)$$

$$x_{AB}, x_{AC}, x_{BC}, x_{BD}, x_{CD} \geq 0$$

Duality in linear programming. Example

Notice that, as a consequence of the restrictions, it is easy to obtain

$$x_{AB} + x_{AC} = 1$$

so that the total amount of the product departs from node A . This is the **primal formulation of the problem**.

We can also think in terms of prices per unit of product at the different nodes of the network y_A, y_B, y_C, y_D and consider the differences between these prices as the profit when a particular channel is used.

In this context, we are seeking the **maximum profit** $y_D - y_A$ in transferring the good from A to D .

The profits of the five channels will be

$$y_B - y_A, \quad y_C - y_A, \quad y_C - y_B, \quad y_D - y_B, \quad y_D - y_C$$

If we take as a normalization rule $y_A = 0$, then we must demand that these profits not exceed the prices for the use of each channel

$$y_B - y_A = y_B \leq 2, \quad y_C - y_A = y_C \leq 3, \quad y_C - y_B \leq 1, \quad y_D - y_B \leq 4, \quad y_D - y_C \leq 2$$

This will be the **dual formulation of the problem**.

Duality in linear programming. Example

To somehow suspect that these two problems are equivalent and that their optimal solutions must be related to each other. The connection is the dual link.

Using

$$c = \begin{pmatrix} 2 \\ 3 \\ 1 \\ 4 \\ 2 \end{pmatrix} \quad b = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$
$$A = \begin{pmatrix} 1 & 0 & -1 & -1 & 0 \\ 0 & 1 & 1 & 0 & -1 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix}$$

the two problems can be formulated in compact form as

$$\begin{array}{llll} \text{Minimize} & c^T x & \text{subject to} & Ax \geq b \quad x \geq 0 \\ \text{Maximize} & y^T b & \text{subject to} & y^T A \leq c \end{array}$$

where

$$x = (x_{AB}, x_{AC}, x_{BC}, x_{BD}, x_{CD})^T, \quad y = (y_B, y_C, y_D)^T$$

Duality in linear programming

Next we describe the relationship between the solutions of a primal problem (P) and its dual (D), where we assume that (P) is given in standard form.

Lemma

(Weak duality) *If x and y are feasible for (P) and (D), respectively, then*

$$y^T b \leq c^T x$$

Moreover, if the equality holds, $y^T b = c^T x$, then x is an optimal solution for (P) and y for (D),

Proof: From

$$Ax = b, \quad x \geq 0, \quad y^T A \leq c \quad \Rightarrow \quad y^T b = y^T Ax \leq c^T x.$$

In particular, we have

$$\max\{y^T b \mid y^T A \leq c\} \leq \min\{c^T x \mid Ax = b, x \geq 0\}.$$

If $y^T b = c^T x$, this number must be at the same time the previous maximum and minimum, and this in turn implies that y is optimal for (D) and x for (P).



Duality in linear programming

Theorem

(Duality theorem) *Either both problems (P) and (D) are solvable simultaneously, or one of the two is degenerate, in the sense that it does not admit feasible vectors.*

Remark 1. What this statement amounts to is that if x is optimal for (P) , then there exists an optimal vector y for (D) , and the common value $y^t b = c^T x$ is at the same time the minimum for (P) and the maximum for (D) . Conversely, if y is optimal for (D) , there exists an optimal vector x for (P) with the common value $y^T b = c^T x$ being at the same time the minimum for and the maximum.

Remark 2. From the proof of the theorem, which is based on the simplex method, it follows that if $x = (x_B \ 0)$ is optimal for (P) , with $x_B = B^{-1}b$, then $y = c_B^T B^{-1}$ is optimal for (D) , with $c = (c_B \ c_N)^T$. (See the next slides for the definitions of B , b , x_B, \dots)

The simplex method in linear programming

Consider the problem of finding a vector x solving

$$\text{Minimize } c^T x$$

subject to

$$Ax = b, \quad x \geq 0$$

We assume, without loss of generality, that the rank of the $m \times n$ ($m \leq n$) matrix A is m and that the linear system $Ax = b$ is solvable.

Example

$$\begin{array}{ll} \text{Minimize} & 3x_1 + x_2 + 9x_3 + x_4 \\ \text{subject to} & x_1 + 2x_3 + x_4 = 4 \\ & x_2 + x_3 - x_4 = 2 \\ & x_i \geq 0 \end{array}$$

so

$$A = \begin{pmatrix} 1 & 0 & 2 & 1 \\ 0 & 1 & 1 & -1 \end{pmatrix}, \quad b = \begin{pmatrix} 4 \\ 2 \end{pmatrix}, \quad c = \begin{pmatrix} 1 \\ 1 \\ 9 \\ 1 \end{pmatrix}.$$

The simplex method in linear programming

The dual problem is:

$$\begin{array}{ll}\text{Maximize} & 4y_1 + 2y_2 \\ \text{subject to} & y_1 \leq 1 \\ & y_2 \leq 1 \\ & 2y_1 + y_2 \leq 9 \\ & y_1 - y_2 \leq 1\end{array}$$

The primal and dual problems can be written in compact form as

$$\begin{array}{llll}\text{Minimize} & c^T x & \text{subject to} & Ax \geq b, \quad x \geq 0 \quad (P) \\ \text{Maximize} & y^T b & \text{subject to} & y^T A \leq c \quad (D)\end{array}$$

Recall that we have already said that: if $x = (x_B \ 0)$ is optimal for (P) , with $x_B = B^{-1}b$, then $y = c_B^T B^{-1}$ is optimal for (D) , with $c = (c_B \ c_N)^T$.

Next, let us solve the primal problem using the **simplex method**.

The simplex method in linear programming

The computation of the **basic solutions** is the first step of the Simplex Algorithm.

The basic solutions are the solutions of the linear system $Ax = b$ with nonnegative and at least $n - m$ null components.

1. Initialization step.

Find a square $m \times m$ submatrix B of A (of rank m) such that the solution of the linear system $Bx_B = b$ is such that $x_B \geq 0$. In this way, the basic solution x can be written as

$$x = \begin{pmatrix} x_B \\ 0 \end{pmatrix}, \quad x_B \in \mathbb{R}^m, \quad 0 \in \mathbb{R}^{n-m}, \quad x_B \geq 0$$

Then, write

$$A = (B \mid N) \quad \text{and} \quad c = \begin{pmatrix} c_B \\ c_N \end{pmatrix}$$

The simplex method in linear programming

In the example

$$B = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad N = \begin{pmatrix} 2 & 1 \\ 1 & -1 \end{pmatrix}, \quad c_B = \begin{pmatrix} 3 \\ 1 \end{pmatrix}, \quad c_N = \begin{pmatrix} 9 \\ 1 \end{pmatrix}$$

$$x_B = b = \begin{pmatrix} 4 \\ 2 \end{pmatrix}$$

and the initial basic vertex is

$$x = \begin{pmatrix} x_B \\ 0 \end{pmatrix} = \begin{pmatrix} 4 \\ 2 \\ 0 \\ 0 \end{pmatrix}$$

with cost $c^T x = 14$.

The simplex method in linear programming

2. Stopping criterion.

After solving

$$z^T B = c_B$$

look at the vector

$$r = c_N - z^T N$$

If $r \geq 0$ stop, since we already have an optimal solution. If not, choose the “entering” variable corresponding to the most negative component of r .

In the example,

$$z = c_B = \begin{pmatrix} 3 \\ 1 \end{pmatrix}, \quad r = \begin{pmatrix} 9 \\ 1 \end{pmatrix} - \begin{pmatrix} 3 & 1 \end{pmatrix} \begin{pmatrix} 2 & 1 \\ 1 & -1 \end{pmatrix} = \begin{pmatrix} 2 \\ -1 \end{pmatrix}$$

Since not all components of r are nonnegative, we must go through the iterative process in the simplex method.

We will choose x_4 as the entering variable, since this is the one associated with the negative component of r .

The simplex method in linear programming

3. Iterative step.

Solve

$$Bw = y$$

where y is the entering column of N corresponding to the entering variable.

In the example,

$$Bw = y \Rightarrow \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} w = \begin{pmatrix} 1 \\ -1 \end{pmatrix} \Rightarrow w = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

Look at the ratios x_B/w componentwise.

$$\frac{x_B}{w} = \left\{ \frac{4}{1} = 4, \frac{2}{-1} = -2 \right\}$$

Among these ratios select those with positive denominators. Choose as “leaving” variable the one corresponding to the smallest ratio among the selected ones.

In the example, x_1 is the leaving variable, being the only one with positive denominator.

The simplex method in linear programming

2. Stopping criterion.

Now

$$B = \begin{pmatrix} 0 & 1 \\ 1 & -1 \end{pmatrix}, \quad N = \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix}, \quad c_B = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad c_N = \begin{pmatrix} 3 \\ 9 \end{pmatrix}$$

$$Bx_B = b \quad \Rightarrow \quad x_B = \begin{pmatrix} 6 \\ 4 \end{pmatrix}$$

The new extremal vector is

$$x = \begin{pmatrix} 0 \\ 6 \\ 0 \\ 4 \end{pmatrix}$$

with cost $c^T x = 10$. The new vectors z and r are

$$z = \begin{pmatrix} 2 \\ 1 \end{pmatrix} \quad r = \begin{pmatrix} 3 \\ 9 \end{pmatrix} - (2 \quad 1) \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 4 \end{pmatrix}$$

Since all components of r are nonnegative, the search is finished. The minimum cost is 10, and it is taken on at the vector $(0, 6, 0, 4)^T$

The simplex method in linear programming

Since

$$c_B = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad x_B = \begin{pmatrix} 6 \\ 4 \end{pmatrix}, \quad B^{-1} = \begin{pmatrix} 0 & 1 \\ 1 & -1 \end{pmatrix},$$

the solution of the dual problem

$$\text{Maximize } y^T b \quad \text{subject to } y^T A \leq c \quad (D)$$

is

$$y = c_B^T B^{-1} = \begin{pmatrix} 1 & 1 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & -1 \end{pmatrix} = \begin{pmatrix} 2 & 1 \end{pmatrix}$$

and the value of the cost function of the dual problem is also $4y_1 + 2y_2 = 10$.

Duality in convex programming

Lagrange duality


- ▶ The theory of Lagrange duality is the study of optimal solutions to convex optimization problems.
- ▶ As we have already seen, when minimizing (without constraints) a differentiable convex function $f(x)$ with respect to $x \in \mathbb{R}^n$, a necessary and sufficient condition for x^* to be globally optimal is that $\nabla f(x^*) = 0$.
- ▶ The primary goal of duality theory is to characterize the optimal points of convex programs.

Duality in convex programming

We will consider a generic differentiable convex optimization problems of the form

$$\begin{cases} \text{minimize} & f(x) \\ \text{subject to} & g_i(x) \leq 0, \quad i=1,\dots,m \\ & h_i(x) = 0, \quad i=1,\dots,p \end{cases}$$

where $x \in \mathbb{R}^n$ is the optimization variable, $f : \mathbb{R}^n \rightarrow \mathbb{R}$ and $g_i : \mathbb{R}^n \rightarrow \mathbb{R}$ are differentiable convex functions, and $h_i : \mathbb{R}^n \rightarrow \mathbb{R}$ are affine functions¹

¹An affine function is a function of the form $f(x) = a^T x + b$ for some $a \in \mathbb{R}^n$, $b \in \mathbb{R}$. Since the Hessian of an affine function is equal to the zero matrix (i.e., it is both positive semidefinite and negative semidefinite), an affine function is both convex and concave. 

The Lagrangian

- ▶ Given a convex constrained minimization problem of the above form (OPT), the (generalized) **Lagrangian** is a function $L : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p \rightarrow \mathbb{R}$ defined as

$$L(x, \alpha, \beta) = f(x) + \sum_{i=1}^m \alpha_i g_i(x) + \sum_{i=1}^p \beta_i h_i(x).$$

- ▶ The first argument of the Lagrangian is a vector $x \in \mathbb{R}^n$ whose dimensionality matches that of the optimization variable in the optimization problem.
- ▶ We will refer to x as the **primal variables of the Lagrangian**.
- ▶ The second argument of the Lagrangian is a vector $\alpha \in \mathbb{R}^m$ with one variable α_i for each of the m convex inequality constraints in the original optimization problem.
- ▶ The third argument of the Lagrangian is a vector $\beta \in \mathbb{R}^p$ with one variable β_i for each of the p affine equality constraints in the original optimization problem.
- ▶ α and β are collectively known as the **dual variables of the Lagrangian** or **Lagrange multipliers**.

The Lagrangian

- ▶ Intuitively, the Lagrangian can be thought of as a modified version of the objective function to the original convex optimization problem (OPT) which accounts for each of the constraints.
- ▶ The Lagrange multipliers α_i and β_i can be thought of “costs” associated with violating different constraints.
- ▶ The key intuition behind the theory of Lagrange duality is the following:

For any convex optimization problem, there always exist settings of the dual variables such that the unconstrained minimum of the Lagrangian with respect to the primal variables (keeping the dual variables fixed) coincides with the solution of the original constrained minimization problem.

Primal and dual problems

The primal problem

- ▶ Consider the primal optimization problem,

$$\min_x \left[\max_{\alpha, \beta, \alpha_i \geq 0} L(x, \alpha, \beta) \right] \equiv \min_x \theta_P(x) \quad (P)$$

- ▶ In the equation above, the function $\theta_P : \mathbb{R}^n \rightarrow \mathbb{R}$ is called the **primal objective**.
- ▶ Note that the minimization problem $\min_x \theta_P(x)$ is an **unconstrained** minimization problem, that is known as the **primal problem**.
- ▶ We say that a point $x \in \mathbb{R}^n$ is **primal feasible** if $g_i(x) \leq 0$, $i = 1, \dots, m$ and $h_i(x) = 0$, $i = 1, \dots, p$.
- ▶ We will use the vector $x^* \in \mathbb{R}^n$ to denote the solution of (P) , and $p^* = \theta_P(x^*)$ will denote the optimal value of the primal objective.

The dual problem

- ▶ By switching the order of the minimization and maximization above, we obtain an entirely different optimization problem,

$$\max_{\alpha, \beta, \alpha_i \geq 0} \left[\min_x L(x, \alpha, \beta) \right] \equiv \max_{\alpha, \beta, \alpha_i \geq 0} \theta_D(\alpha, \beta) \quad (D)$$

- ▶ In the equation above, the function $\theta_D : \mathbb{R}^n \rightarrow \mathbb{R}$ is called the **dual objective**.
- ▶ The **constrained** minimization problem $\max_{\alpha, \beta, \alpha_i \geq 0} \theta_D(\alpha, \beta)$ is known as the **dual problem**.
- ▶ We say that $(\alpha, \beta) \in \mathbb{R}^m \times \mathbb{R}^p$ are **dual feasible** if $\alpha_i \geq 0$, $i = 1, \dots, m$.
- ▶ We will use the pair of vectors (α^*, β^*) to denote the solution of (D) , and we $d^* = \theta_D(\alpha^*, \beta^*)$ will denote the optimal value of the dual objective.

The primal objective function

The primal objective function, $\theta_P(x)$, is a convex function of x .

$$\theta_P(x) = \max_{\alpha, \beta, \alpha_i \geq 0} \left[f(x) + \sum_{i=1}^m \alpha_i g_i(x) + \sum_{i=1}^m \beta_i h_i(x) \right]$$

- ▶ Each of the $g_i(x)$ are convex functions in x , and since the $\alpha_i \geq 0$, then $\alpha_i g_i(x)$ is convex in x for each i .
- ▶ Similarly, each $\beta_i h_i(x)$ is convex in x (regardless of the sign of β_i) since $h_i(x)$ is linear.
- ▶ Since the sum of convex functions is always convex, the quantity inside the brackets is a convex function of x .
- ▶ Finally, the maximum of a collection of convex functions is again a convex function, so $\theta_P(x)$ is a convex function of x .

Interpreting the primal problem

To interpret the primal problem, note that

$$\begin{aligned}\theta_P(x) &= \max_{\alpha, \beta, \alpha_i \geq 0} L(x, \alpha, \beta) \\ &= f(x) + \max_{\alpha, \beta, \alpha_i \geq 0} \left[\sum_{i=1}^m \alpha_i g_i(x) + \sum_{i=1}^m \beta_i h_i(x) \right]\end{aligned}$$

Considering only the bracketed term, notice that

- ▶ If any $g_i(x) > 0$, then maximizing the bracketed expression involves making the corresponding α_i an arbitrarily large positive number.
- ▶ If $g_i(x) \leq 0$, then the requirement that $\alpha_i \geq 0$ be nonnegative means that the optimal setting of α_i to achieve the maximum is $\alpha_i = 0$, so that the maximum value is 0.
- ▶ Similarly, if any $h_i(x) \neq 0$, then maximizing the bracketed expression involves choosing the corresponding β_i to have the same sign as $h_i(x)$ and arbitrarily large magnitude.
- ▶ If $h_i(x) = 0$, then the maximum value is 0, independent of β_i .

Interpreting the primal problem

Putting these two cases together, we see that if x is primal feasible

$$g_i(x) \leq 0, \quad i = 1, \dots, m, \quad \text{and} \quad h_i(x) = 0, \quad i = 1, \dots, m$$

then the maximum value of the bracketed expression is 0, but if any of the constraints are violated, then the maximum value is ∞ .

From this, we can write,

$$\theta_P(x) = f(x) + \begin{cases} 0 & \text{if } x \text{ is primal feasible} \\ \infty & \text{if } x \text{ is primal infeasible} \end{cases}$$

Therefore, we can interpret the primal objective $\theta_P(x)$ as a modified version of the convex objective function of the original problem, with the difference being that infeasible points have objective value ∞ .

Intuitively, we can consider

$$\max_{\alpha, \beta, \alpha_i \geq 0} \left[\sum_{i=1}^m \alpha_i g_i(x) + \sum_{i=1}^m \beta_i h_i(x) \right] = \begin{cases} 0 & \text{if } x \text{ is feasible for the initial prob.} \\ \infty & \text{if } x \text{ is infeasible for the initial prob.} \end{cases}$$

as a type of “barrier” function which prevents us from considering infeasible points as candidate solutions for the optimization problem.

The dual objective function

The dual objective function $\theta_D(\alpha, \beta)$ is a **concave function** of α and β .

$$\theta_D(\alpha, \beta) = \min_x L(x, \alpha, \beta) = \min_x \left[f(x) + \sum_{i=1}^m \alpha_i g_i(x) + \sum_{i=1}^m \beta_i h_i(x) \right]$$

- ▶ Observe that for any fixed value of x , the quantity inside the brackets is an affine function of α and β , and hence concave.
- ▶ Since the minimum of a collection of concave functions is also concave, we can conclude that $\theta_D(\alpha, \beta)$ is a concave function of α and β .

Interpreting the dual problem

Lemma

If (α, β) are dual feasible, then $\theta_D(\alpha, \beta) \leq p^*$

Proof: Observe that

$$\begin{aligned}\theta_D(\alpha, \beta) &= \min_x L(x, \alpha, \beta) \\ &\leq L(x^*, \alpha, \beta) \\ &= f(x^*) + \sum_{i=1}^m \alpha_i g_i(x^*) + \sum_{i=1}^m \beta_i h_i(x^*) \\ &\leq f(x^*) = p^*\end{aligned}$$

The last inequality follows from the fact that x^* is primal feasible ($g_i(x^*) \leq 0$ and $h_i(x^*) = 0$), (α, β) are dual feasible ($\alpha_i \geq 0$), and hence the latter two terms of in the last minus one line must be nonpositive. \square

This lemma shows that that given any dual feasible (α, β) , the dual objective $\theta_D(\alpha, \beta)$ provides a lower bound on the optimal value p^* of the primal problem.

Interpreting the dual problem

Since the dual problem involves maximizing the dual objective over the space of all dual feasible (α, β) , it follows that **the dual problem can be seen as a search for the tightest possible lower bound on p^* .**

This gives rise to a property of any primal and dual optimization problem pairs known as weak duality:

Lemma

(Weak Duality). For any pair of primal and dual problems, $d^ \leq p^*$.*

For some primal/dual optimization problems, an even stronger result holds, known as strong duality :

Lemma

*(Strong Duality). For any pair of primal and dual problems, which satisfy certain technical conditions called **constraint qualifications**, $d^* = p^*$.*

Constraint qualifications

A number of different **constraint qualifications** exist, of which the most commonly invoked constraint qualification is known as Slater's condition:

A primal/dual problem pair satisfy **Slater's condition** if there exists some feasible primal solution x for which all inequality constraints are strictly satisfied: i.e., $g_i(x) < 0$, $i = 1, \dots, m$.

In practice, **nearly all convex problems satisfy some kind of constraint qualification**, and hence the primal and dual problems have the same optimal value.

Complementary slackness or KKT complementarity

Lemma

(Complementary Slackness). If strong duality holds ($d^* = p^*$), then $\alpha_i^* g_i(x^*) = 0$ for each $i = 1, \dots, m$.

Proof: Suppose that strong duality holds. Then observe that

$$\begin{aligned} p^* = d^* = \theta_D(\alpha^*, \beta^*) &= \min_x L(x, \alpha^*, \beta^*) \leq L(x^*, \alpha^*, \beta^*) \\ &= f(x^*) + \sum_{i=1}^m \alpha_i^* g_i(x^*) + \sum_{i=1}^m \beta_i^* h_i(x^*) \\ &\leq f(x^*) = p^*, \end{aligned}$$

so

$$\sum_{i=1}^m \alpha_i^* g_i(x^*) + \sum_{i=1}^m \beta_i^* h_i(x^*) = 0$$

Recall that each α_i^* is nonnegative, each $g_i(x^*)$ is nonpositive, and each $h_i(x^*)$ is zero, due to the primal and dual feasibility of x^* and (α^*, β^*) , respectively.

As a consequence, we have a summation of all nonpositive terms which equals to zero, so all individual terms in the summation must be zero. \square

Complementary slackness

- ▶ Complementary slackness can be written in many equivalent ways. One way, in particular, is the pair of conditions

$$\begin{aligned}\alpha_i^* > 0 &\Rightarrow g_i(x^*) = 0 \\ g_i(x^*) < 0 &\Rightarrow \alpha_i^* = 0\end{aligned}$$

- ▶ In this form, we can see that whenever any α_i^* is strictly greater than zero, then this implies that the corresponding inequality constraint must hold with equality. We refer to this as an **active constraint**.

The KKT conditions

Finally, we can now characterize the optimal conditions for a primal-dual optimization pair.

Theorem

Suppose that $x^* \in \mathbb{R}^n$, $\alpha^* \in \mathbb{R}^m$, $\beta^* \in \mathbb{R}^p$ satisfy the following conditions:

1. (Primal feasibility) $g_i(x^*) \leq 0$, $i = 1, \dots, m$ and $h_i(x^*) = 0$, $i = 1, \dots, p$,
2. (Dual feasibility) $\alpha_i^* \geq 0$, $i = 1, \dots, m$
3. (Complementary slackness) $\alpha_i^* g_i(x^*) = 0$, $i = 1, \dots, m$, and
4. (Lagrangian stationarity) $\nabla_x L(x^*, \alpha^*, \beta^*) = 0$.

Then, x^* is primal optimal and (α^*, β^*) are dual optimal.

Furthermore, if strong duality holds, then any primal optimal x^* and dual optimal (α^*, β^*) must satisfy the conditions 1 through 4.

The above conditions are known as the **Karush-Kuhn-Tucker (KKT) conditions**.

A simple duality example

Consider the convex optimization problem in \mathbb{R}^2 :

$$\begin{array}{ll}\text{minimize} & x_1^2 + x_2 \\ \text{subject to} & 2x_1 + x_2 \geq 4 \\ & x_2 \geq 1\end{array}$$

We **must** write it as

$$\begin{array}{ll}\text{minimize} & x_1^2 + x_2 \\ \text{subject to} & 4 - 2x_1 - x_2 \leq 0 \\ & 1 - x_2 \leq 0\end{array}$$

The Lagrangian is then

$$L(x, \alpha) = x_1^2 + x_2 + \alpha_1(4 - 2x_1 - x_2) + \alpha_2(1 - x_2),$$

and the objective of the dual problem is defined to be

$$\theta_D(\alpha) = \min_x L(x, \alpha)$$

A simple duality example (cont.)

- ▶ To express the dual objective in a form which depends only on α , but not x , we first observe that the the Lagrangian is differentiable in x , and in fact, is separable in the two components x_1 and x_2 , so we can minimize with respect to each separately.

$$L(x, \alpha) = x_1^2 + x_2 + \alpha_1(4 - 2x_1 - x_2) + \alpha_2(1 - x_2)$$

- ▶ To minimize with respect to x_1 , observe that the Lagrangian is a strictly convex quadratic function of x_1 and hence the minimum with respect to x_1 can be found by setting the derivative to zero

$$\frac{\partial}{\partial x_1} L(x, \alpha) = 2x_1 - 2\alpha_1 = 0 \quad \Rightarrow \quad x_1 = \alpha_1$$

- ▶ To minimize with respect to x_2 , observe that the Lagrangian is an affine function of x_2 , for which the linear coefficient is precisely the derivative of the Lagrangian coefficient with respect to x_2

$$\frac{\partial}{\partial x_2} L(x, \alpha) = 1 - \alpha_1 - \alpha_2$$

A simple duality example (cont.)

Using that

$$L(x, \alpha) = x_1^2 + x_2 + \alpha_1(4 - 2x_1 - x_2) + \alpha_2(1 - x_2),$$

and using the above observations, we have

$$\begin{aligned}\theta_D(\alpha) &= \min_x L(x, \alpha) \\ &= \min_{x_2} [\alpha_1^2 + x_2 + \alpha_1(4 - 2\alpha_1 - x_2) + \alpha_2(1 - x_2)] \\ &= \min_{x_2} [-\alpha_1^2 + 4\alpha_1 + \alpha_2 + x_2(1 - \alpha_1 - \alpha_2)] \\ &= \begin{cases} -\alpha_1^2 + 4\alpha_1 + \alpha_2 & \text{if } 1 - \alpha_1 - \alpha_2 = 0 \\ \infty & \text{otherwise} \end{cases}\end{aligned}$$

- ▶ Since, if the coefficient $1 - \alpha_1 - \alpha_2$ is zero, then the objective function does not depend on x_2 .
- ▶ If the coefficient $1 - \alpha_1 - \alpha_2$ is non-zero, then the objective function can be made arbitrarily small by choosing the x_2 to have the opposite sign of the linear coefficient and arbitrarily large magnitude.

A simple duality example (cont.)

The dual problem is given by

$$\begin{array}{ll}\text{maximize}_{\alpha \in \mathbb{R}^2} & \theta_D(\alpha) \\ \text{subject to} & \alpha_1 \geq 0 \\ & \alpha_2 \geq 0\end{array}$$

We can simplify the dual problem making the dual constraints explicit

$$\begin{array}{ll}\text{maximize}_{\alpha \in \mathbb{R}^2} & -\alpha_1^2 + 4\alpha_1 + \alpha_2 \\ \text{subject to} & \alpha_1 \geq 0 \\ & \alpha_2 \geq 0 \\ & 1 - \alpha_1 - \alpha_2 = 0\end{array}$$

Notice that the dual problem is a concave quadratic program in the variables α

The linear Support Vector Machine

We are given a training dataset of n points of the form

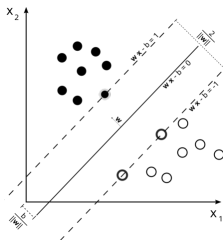
$$(x^{(1)}, y^{(1)}), \dots, (x^{(n)}, y^{(n)})$$

where the $y^{(i)}$ are either 1 or -1 , each indicating the class to which the point $x^{(i)}$ belongs.

We want to find the "maximum-margin hyperplane"

$$\vec{w}^T \vec{x} - b = 0$$

that divides the group of points $x^{(i)}$ for which $y^{(i)} = 1$ from the group of points for which $y^{(i)} = -1$, which is defined so that the distance between the hyperplane and the nearest point $x^{(i)}$ from either group is maximized.



The parameter $b/\|\vec{w}\|$ determines the offset of the hyperplane from the origin along the normal vector \vec{w} .

The hard-margin Support Vector Machine.

If the data $(x^{(i)}, y^{(i)})$ are linearly separable, we can **select two parallel hyperplanes that separate the two classes of data, so that the distance between them is as large as possible.**

The region bounded by these two hyperplanes is called the margin, and the maximum-margin hyperplane is the hyperplane that lies halfway between them.

These hyperplanes can be described by the equations

$$w^T x - b = 1$$

and

$$w^T x - b = -1.$$

Geometrically, the distance between these two hyperplanes is $2/\|w\|$, so **to maximize the distance between the planes we want to minimize $\|w\|$.**

The hard-margin Support Vector Machine.

We also have to prevent data points from falling into the margin; for this we add the following constraint: for each i either

$$w^T x^{(i)} - b \geq 1, \quad \text{if } y^{(i)} = 1$$

or

$$w^T x^{(i)} - b \leq -1, \quad \text{if } y^{(i)} = -1$$

These constraints state that each data point must lie on the correct side of the margin.

The above two conditions can be rewritten as:

$$y^{(i)}(w^T x^{(i)} - b) \geq 1, \quad \text{for all } 1 \leq i \leq n.$$

The optimization problem becomes:

$$\text{Minimize } \|w\| \text{ subject to } y^{(i)}(w^T x^{(i)} - b) \geq 1, \text{ for } i = 1, \dots, n$$

The w and b that solve this problem determine our classifier,

$$x \mapsto \text{sgn}(w^T x - b).$$

The soft-margin Support Vector Machine

To extend SVM to cases in which the data are not linearly separable (soft-margin), we introduce the variables ξ in the loss function, and the soft-margin SVM is defined by

$$\begin{aligned} \text{minimize}_{w,b,\xi} \quad & \frac{1}{2} \|w\|^2 + C \sum_{i=1}^m \xi_i \\ \text{subject to} \quad & 1 - \xi_i - y^{(i)}(w^T x^{(i)} + b) \leq 0 \quad i=1,\dots,m \\ & -\xi_i \leq 0 \quad i=1,\dots,m \end{aligned}$$

whose Lagrangian is

$$L(w, b, \xi, \alpha, \beta) = \frac{1}{2} \|w\|^2 + C \sum_{i=1}^m \xi_i + \sum_{i=1}^m \alpha_i \left(1 - \xi_i - y^{(i)}(w^T x^{(i)} + b) \right) - \sum_{i=1}^m \beta_i \xi_i$$

It must be noted that (w, b, ξ) play the role of the “ x ” primal variables and that (α, β) play the role of the “ α ” dual variables normally used for inequality constraints.

The soft-margin SVM (cont.)

The above Lagrangian gives the primal and dual optimization problems

$$\max_{\alpha, \beta: \alpha_i \geq 0, \beta_i \geq 0} \theta_D(\alpha, \beta) \quad \text{where} \quad \theta_D(\alpha, \beta) = \min_{w, b, \xi} L(w, b, \xi, \alpha, \beta) \quad (SVM-D)$$

$$\min_{w, b, \xi} \theta_P(w, b, \xi) \quad \text{where} \quad \theta_P(w, b, \xi) = \max_{\alpha, \beta: \alpha_i \geq 0, \beta_i \geq 0} L(w, b, \xi, \alpha, \beta) \quad (SVM-P)$$

To get the dual problem in the form shown previously, we still have a little more work to do.

- **Eliminating the primal variables.** To eliminate the primal variables from the dual problem, we compute $\theta_D(\alpha, \beta)$ by noticing that

$$\theta_D(\alpha, \beta) = \min_{w, b, \xi} L(w, b, \xi, \alpha, \beta)$$

is an unconstrained optimization problem, where the objective function $L(w, b, \xi, \alpha, \beta)$ is differentiable. The Lagrangian is a strictly convex quadratic function of w , so for any fixed (α, β) , if $(\hat{w}, \hat{b}, \hat{\xi})$ minimize the Lagrangian, it must be the case that

$$\nabla_w L(\hat{w}, \hat{b}, \hat{\xi}, \alpha, \beta) = \hat{w} - \sum_{i=1}^m \alpha_i y^{(i)} x^{(i)} = 0$$

The soft-margin SVM (cont.)

- Note that the Lagrangian is linear in b and ξ .
- As in the previous duality example, we can set the derivatives with respect to b and ξ to zero, and add the resulting conditions as explicit constraints in the dual optimization problem

$$\frac{\partial}{\partial b} L(\hat{\mathbf{w}}, \hat{b}, \hat{\xi}, \alpha, \beta) = - \sum_{i=1}^m \alpha_i y^{(i)} = 0, \quad \frac{\partial}{\partial \xi_i} L(\hat{\mathbf{w}}, \hat{b}, \hat{\xi}, \alpha, \beta) = C - \alpha_i - \beta_i = 0$$

- Using these conditions, the dual objective function is

$$\begin{aligned} \theta_D(\alpha, \beta) &= L(\hat{\mathbf{w}}, \hat{b}, \hat{\xi}) \\ &= \frac{1}{2} \|\hat{\mathbf{w}}\|^2 + C \sum_{i=1}^m \hat{\xi}_i + \sum_{i=1}^m \alpha_i \left(1 - \hat{\xi}_i - y^{(i)} (\hat{\mathbf{w}}^T \mathbf{x}^{(i)} + \hat{b}) \right) - \sum_{i=1}^m \beta_i \hat{\xi}_i \\ &= \frac{1}{2} \|\hat{\mathbf{w}}\|^2 + C \sum_{i=1}^m \hat{\xi}_i + \sum_{i=1}^m \alpha_i \left(1 - \hat{\xi}_i - y^{(i)} (\hat{\mathbf{w}}^T \mathbf{x}^{(i)}) \right) - \sum_{i=1}^m \beta_i \hat{\xi}_i \\ &= \frac{1}{2} \|\hat{\mathbf{w}}\|^2 + \sum_{i=1}^m \alpha_i \left(1 - y^{(i)} (\hat{\mathbf{w}}^T \mathbf{x}^{(i)}) \right) \end{aligned}$$

where the first equality follows from the optimality of $(\hat{\mathbf{w}}, \hat{b}, \hat{\xi})$, and the third and fourth equalities follow from the last two conditions.

The soft-margin SVM (cont.)

- To use the $\nabla_w L = 0$ condition, observe that

$$\begin{aligned}\frac{1}{2}\|\hat{w}\|^2 + \sum_{i=1}^m \alpha_i \left(1 - y^{(i)}(\hat{w}^T x^{(i)})\right) &= \sum_{i=1}^m \alpha_i + \frac{1}{2}\|\hat{w}\|^2 - \hat{w}^T \sum_{i=1}^m \alpha_i y^{(i)} x^{(i)} \\ &= \sum_{i=1}^m \alpha_i + \frac{1}{2}\|\hat{w}\|^2 - \|\hat{w}\|^2 \\ &= \sum_{i=1}^m \alpha_i - \frac{1}{2}\|\hat{w}\|^2 \\ &= \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j y^{(i)} y^{(j)} \langle x^{(i)}, x^{(j)} \rangle\end{aligned}$$

The soft-margin SVM (cont.)

- Therefore, our dual problem (with no more primal variables and all constraints made explicit) is

$$\text{maximize}_{\alpha, \beta} \quad \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j y^{(i)} y^{(j)} \langle x^{(i)}, x^{(j)} \rangle$$

$$\text{subject to} \quad \alpha_i \geq 0, \quad i = 1, \dots, m$$

$$\beta_i \geq 0, \quad i = 1, \dots, m$$

$$\alpha_i + \beta_i = C, \quad i = 1, \dots, m$$

$$\sum_{i=1}^m \alpha_i y^{(i)} = 0$$

The soft-margin SVM (cont.)

- **KKT complementary.** KKT complementarity requires that for any primal optimal (w^*, b^*, ξ^*) and dual optimal (α^*, β^*)

$$\begin{aligned}\alpha_i^* \left(1 - \xi_i^* - y^{(i)}((w^{*T} x^{(i)} + b^*))\right) &= 0 \\ \beta_i^* \xi_i^* &= 0\end{aligned}$$

for $i = 1, \dots, m$. From the first condition, we see that if $\alpha_i^* > 0$, then in order for the product to be zero, then $1 - \xi_i^* - y^{(i)}((w^{*T} x^{(i)} + b^*)) = 0$. It follows that

$$y^{(i)}((w^{*T} x^{(i)} + b^*)) \leq 1$$

since $\xi_i^* \geq 0$ by primal feasibility. Similarly, if $\beta_i^* > 0$ then $\xi_i^* = 0$ to ensure complementarity.

From the primal constraint, $y^{(i)}((w^{*T} x^{(i)} + b^*)) \geq 1 - \xi_i$, it follows that

$$y^{(i)}((w^{*T} x^{(i)} + b^*)) \geq 1$$

Finally, since $\beta_i^* > 0$ is equivalent to $\alpha_i^* < C$ (since $\alpha_i^* + \beta_i^* = C$), we can summarize the KKT conditions as follows:

$$\alpha_i^* < C \quad \Rightarrow \quad y^{(i)}(w^{*T} x^{(i)} + b^*) \geq 1$$

$$\alpha_i^* > 0 \quad \Rightarrow \quad y^{(i)}(w^{*T} x^{(i)} + b^*) \leq 1$$

The soft-margin SVM (cont.)

► Equivalently

$$\alpha_i^* = 0 \quad \Rightarrow \quad y^{(i)}(w^{*T} x^{(i)} + b^*) \geq 1$$

$$0 < \alpha_i^* < C \quad \Rightarrow \quad y^{(i)}(w^{*T} x^{(i)} + b^*) = 1$$

$$\alpha_i^* = C \quad \Rightarrow \quad y^{(i)}(w^{*T} x^{(i)} + b^*) \leq 1$$

The soft-margin SVM (cont.)

- **Simplification.** We can tidy up our dual problem slightly by observing that each pair of constraints of the form

$$\beta_i \geq 0 \quad \alpha_i + \beta_i = C$$

is equivalent to the single constraint $\alpha_i \leq C$; that is, if we solve the optimization problem

$$\text{maximize}_{\alpha, \beta} \quad \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j y^{(i)} y^{(j)} \langle x^{(i)}, x^{(j)} \rangle$$

$$\text{subject to} \quad 0 \leq \alpha_i \leq C, \quad i = 1, \dots, m$$

$$\sum_{i=1}^m \alpha_i y^{(i)} = 0$$

and subsequently set $\beta_i = C - \alpha_i$, then it follows that (α, β) will be optimal for the previous dual problem