Part 4. KKT Conditions and Duality

Math 126 Winter 18

Date of current version: February 16, 2018

Abstract This note studies duality. Many parts of this note are based on the chapters [1, Chapter 10-12] [2, Chapter 2,5] and their corresponding lecture notes available online by the authors.

Please email me if you find any typos or errors.

1 Linearly Constrained Problems (see [1, Chapter 10] [2, Chapter 2])

Consider the linearly constrained problem:

$$\min_{\boldsymbol{x}} f(\boldsymbol{x}) \quad (1.1)$$
$$\text{subject to } \boldsymbol{a}_i^\top \boldsymbol{x} \le b_i, \quad i = 1, \ldots, m.$$

Recall the optimality condition with the feasible set $\mathcal{X} = \{\boldsymbol{x} \,:\, \boldsymbol{a}_i^\top \boldsymbol{x} \le b_i,\ i = 1, \ldots, m\}$.

Theorem 1.1 *Suppose f is convex and differentiable. Then $\boldsymbol{x}_*$ is optimal if and only if $\boldsymbol{x}_* \in \mathcal{X}$ and*

$$\langle \nabla f(\boldsymbol{x}_*), \boldsymbol{y} - \boldsymbol{x}_* \rangle \ge 0 \text{ for all } \boldsymbol{y} \in \mathcal{X}. \quad (1.2)$$

This is difficult to validate, and this section derives an equivalent optimality condition that is much easier to handle for the linearly constrained problems.

1.1 Separation Theorem

Theorem 1.2 *(Strict separation theorem) Let $\mathcal{C} \subseteq \mathbb{R}^d$ be a nonempty closed and convex set, and let $\boldsymbol{y} \notin \mathcal{C}$. Then there exists $\boldsymbol{p} \in \mathbb{R}^d \backslash \{\boldsymbol{0}\}$ and $\alpha \in \mathbb{R}$ such that*

$$\boldsymbol{p}^\top \boldsymbol{y} > \alpha \quad \text{and} \quad \boldsymbol{p}^\top \boldsymbol{x} \le \alpha \text{ for all } \boldsymbol{x} \in \mathcal{C}. \quad (1.3)$$

1.2 Farkas' Lemma: Alternative Theorem

Lemma 1.1 *(Farkas' lemma) Let $\boldsymbol{A} \in \mathbb{R}^{p \times d}$ and $\boldsymbol{b} \in \mathbb{R}^d$. Then exactly one of the following systems has a solution:*

- $\boldsymbol{A}\boldsymbol{x} \preceq \boldsymbol{0},\ \boldsymbol{b}^\top \boldsymbol{x} > 0$
- $\boldsymbol{A}^\top \boldsymbol{y} = \boldsymbol{b},\ \boldsymbol{y} \succeq 0$

Proof *The proof uses Theorem 1.2.* □

Donghwan Kim
Dartmouth College
E-mail: donghwan.kim@dartmouth.edu

Theorem 1.3 *(KKT conditions for linearly constrained problems; necessary optimality conditions) Consider the problem (1.1) where f is continuously differentiable over $\mathbb{R}^d$. Let $\boldsymbol{x}_*$ be a local minimum point of (1.1). Then there exists $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_m)^\top \succeq \mathbf{0}$ such that*

$$\nabla f(\boldsymbol{x}_*) + \sum_{i=1}^{m} \lambda_i \boldsymbol{a}_i = \mathbf{0} \quad \text{and} \quad \lambda_i(\boldsymbol{a}_i^\top \boldsymbol{x}_* - b_i) = 0,\ i = 1, \ldots, m. \tag{1.4}$$

Proof *Since $\boldsymbol{x}_*$ is a local minimum point, we have*

$$\langle \nabla f(\boldsymbol{x}_*), \boldsymbol{x} - \boldsymbol{x}_* \rangle \geq 0 \text{ for all } \boldsymbol{x} \in \mathbb{R}^d \text{ satisfying } \boldsymbol{a}_i^\top \boldsymbol{x} \leq b_i,\ i = 1, \ldots, m.$$

Denote the set of active constraints by

$$I(\boldsymbol{x}_*) = \{i \ :\ \boldsymbol{a}_i^\top \boldsymbol{x}_* = b_i\}$$

Using the change of variables $\boldsymbol{y} = \boldsymbol{x} - \boldsymbol{x}_*$, we have $\langle \nabla f(\boldsymbol{x}_*), \boldsymbol{y} \rangle \geq 0$ for any $\boldsymbol{y}$ satisfying

$$\begin{cases} \boldsymbol{a}_i^\top \boldsymbol{y} \leq 0, & i \in I(\boldsymbol{x}_*), \\ \boldsymbol{a}_i^\top \boldsymbol{y} \leq b_i - \boldsymbol{a}_i^\top \boldsymbol{x}_*, & i \notin I(\boldsymbol{x}_*). \end{cases}$$

Next shows that the second set of inequalities can be removed. Suppose that $\boldsymbol{y}$ satisfies $\boldsymbol{a}_i^\top \boldsymbol{y} \leq 0$ for all $i \in I(\boldsymbol{x}_*)$. Since $b_i - \boldsymbol{a}_i^\top \boldsymbol{x}_* > 0$ for all $i \notin I(\boldsymbol{x}_*)$, it follows that there exists a small enough $\alpha > 0$, for which $\boldsymbol{a}_i^\top(\alpha \boldsymbol{y}) \leq b_i - \boldsymbol{a}_i^\top \boldsymbol{x}_*$. Thus, since we also have $\boldsymbol{a}_i^\top(\alpha \boldsymbol{y}) \leq 0$ for any $i \in I(\boldsymbol{x}_*)$, it follows that $\langle \nabla f(\boldsymbol{x}_*), \alpha \boldsymbol{y} \rangle \geq 0$, and hence $\langle \nabla f(\boldsymbol{x}_*), \boldsymbol{y} \rangle \geq 0$. This shows that

$$\boldsymbol{a}_i^\top \boldsymbol{y} \leq 0 \text{ for all } i \in I(\boldsymbol{x}_*) \qquad \Longrightarrow \qquad \langle -\nabla f(\boldsymbol{x}_*), \boldsymbol{y} \rangle \leq 0.$$

Thus, by Lemma 1.1, there exist $\lambda_i \geq 0$, $i \in I(\boldsymbol{x}_*)$, such that

$$-\nabla f(\boldsymbol{x}_*) = \sum_{i \in I(\boldsymbol{x}_*)} \lambda_i \boldsymbol{a}_i.$$

Defining $\lambda_i = 0$ for all $i \notin I(\boldsymbol{x}_*)$, we equivalently have that $\lambda_i(\boldsymbol{a}_i^\top \boldsymbol{x}_* - b_i) = 0$ for all $i = 1, \ldots, m$ and

$$\nabla f(\boldsymbol{x}_*) + \sum_{i=1}^{m} \lambda_i \boldsymbol{a}_i = \mathbf{0}$$

are required. □

Theorem 1.4 *(KKT conditions for convex linearly constrained problems; necessary and sufficient optimality conditions) Consider the problem (1.1) where f is convex and continuously differentiable over $\mathbb{R}^d$. Let $\boldsymbol{x}_*$ be a feasible point of (1.1). Then $\boldsymbol{x}_*$ is an optimal solution of (1.1) if and only if there exists $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_m)^\top \succeq \mathbf{0}$ such that*

$$\nabla f(\boldsymbol{x}_*) + \sum_{i=1}^{m} \lambda_i \boldsymbol{a}_i = \mathbf{0} \quad \text{and} \quad \lambda_i(\boldsymbol{a}_i^\top \boldsymbol{x}_* - b_i) = 0,\ i = 1, \ldots, m. \tag{1.5}$$

Proof *Necessity is proven. To prove the sufficiency, suppose that $\boldsymbol{x}_*$ is a feasible point of (1.1) satisfying (1.5). Let $\boldsymbol{x}$ be any feasible point of (1.1). Define the function*

$$\mathcal{L}(\boldsymbol{x}) = f(\boldsymbol{x}) + \sum_{i=1}^{m} \lambda_i(\boldsymbol{a}_i^\top \boldsymbol{x} - b_i).$$

A feasible point $\boldsymbol{x}_*$ is a minimizer of $\mathcal{L}$, since $\nabla \mathcal{L}(\boldsymbol{x}_*) = \mathbf{0}$ by the first condition of (1.5). This with the second condition of (1.5) yields

$$f(\boldsymbol{x}_*) = f(\boldsymbol{x}_*) + \sum_{i=1}^{m} \lambda_i(\boldsymbol{a}_i^\top \boldsymbol{x}_* - b_i) = \mathcal{L}(\boldsymbol{x}_*) \leq \mathcal{L}(\boldsymbol{x}) = f(\boldsymbol{x}) + \sum_{i=1}^{m} \lambda_i(\boldsymbol{a}_i^\top \boldsymbol{x} - b_i) \leq f(\boldsymbol{x}).$$

This proves that $\boldsymbol{x}_*$ is a global optimal solution of (1.1). □

Theorem 1.5 *(KKT conditions for linearly constrained problems) Consider*

$$\min_{\boldsymbol{x}} f(\boldsymbol{x}) \tag{1.6}$$

$$\begin{aligned} \textit{subject to } \boldsymbol{a}_i^\top \boldsymbol{x} &\le b_i, \quad i = 1, \dots, m, \\ \boldsymbol{c}_i^\top \boldsymbol{x} &= d_i, \quad i = 1, \dots, n, \end{aligned} \tag{1.7}$$

where f is a continuously differentiable function over $\mathbb{R}^d$. Then we have

- *(necessity of the KKT conditions) If $\boldsymbol{x}_*$ is a local minimum point, then there exists $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_m)^\top \succeq \mathbf{0}$ and $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)^\top \in \mathbb{R}^n$ such that*

$$\nabla f(\boldsymbol{x}_*) + \sum_{i=1}^m \lambda_i \boldsymbol{a}_i + \sum_{i=1}^n \mu_i \boldsymbol{c}_i = \mathbf{0} \quad \textit{and} \quad \lambda_i(\boldsymbol{a}_i^\top \boldsymbol{x}_* - b_i) = 0,\ i = 1, \dots, m. \tag{1.8}$$

- *(sufficiency in the convex case) If in addition f is convex over $\mathbb{R}^d$ and $\boldsymbol{x}_*$ is a feasible point for which there exist $\boldsymbol{\lambda} \succeq \mathbf{0}$ and $\boldsymbol{\mu} \in \mathbb{R}^n$ such that (1.8) is satisfied, then $\boldsymbol{x}_*$ is an optimal solution.*

Proof – *Consider the equivalent problem*

$$\begin{aligned} &\min_{\boldsymbol{x}} f(\boldsymbol{x}) \\ \textit{subject to } \boldsymbol{a}_i^\top \boldsymbol{x} &\le b_i, \quad i = 1, \dots, m, \\ \boldsymbol{c}_i^\top \boldsymbol{x} &\le d_i, \quad i = 1, \dots, n, \\ -\boldsymbol{c}_i^\top \boldsymbol{x} &\le -d_i, \quad i = 1, \dots, n, \end{aligned}$$

where $\boldsymbol{x}_*$ is also a local minimum point, and thus, by Theorem 1.3 there exist $\boldsymbol{\lambda}, \boldsymbol{\mu}^+, \boldsymbol{\mu}^- \succeq \mathbf{0}$ such that

$$\nabla f(\boldsymbol{x}_*) + \sum_{i=1}^m \lambda_i \boldsymbol{a}_i + \sum_{i=1}^n \mu_i^+ \boldsymbol{c}_i - \sum_{i=1}^n \mu_i^- \boldsymbol{c}_i = \mathbf{0} \quad \textit{and} \quad \begin{cases} \lambda_i^\top(\boldsymbol{a}_i^\top \boldsymbol{x}_* - b_i) = 0,\ i = 1, \dots, m, \\ \mu_i^+(\boldsymbol{c}_i^\top \boldsymbol{x}_* - d_i) = 0,\ i = 1, \dots, n, \\ \mu_i^-(-\boldsymbol{c}_i^\top \boldsymbol{x}_* + d_i) = 0,\ i = 1, \dots, n. \end{cases} \tag{1.9}$$

We thus have (1.8) with $\boldsymbol{\mu} = \boldsymbol{\mu}^+ - \boldsymbol{\mu}^-$.

– *Suppose that $\boldsymbol{x}_*$ satisfies (1.8). Then it also satisfies (1.9) with $\boldsymbol{\mu}^+ = \max\{\boldsymbol{\mu}, \mathbf{0}\}$ and $\boldsymbol{\mu}^- = \max\{-\boldsymbol{\mu}, \mathbf{0}\}$. Then by Theorem 1.4 this implies that $\boldsymbol{x}_*$ is an optimal solution.*

□

2 KKT Conditions (see [1, Chapter 11] [2, Chapter 5])

Consider the linearly constrained problem:

$$\begin{aligned} &\min_{\boldsymbol{x}} f(\boldsymbol{x}) \\ &\text{subject to } g_i(\boldsymbol{x}) \le 0, \quad i = 1, \dots, m, \end{aligned} \tag{2.1}$$

where f, g_i are continuously differentiable functions over $\mathbb{R}^d$.

Definition 2.1 (feasible descent directions). Consider the problem

$$\begin{aligned} &\min f(\boldsymbol{x}) \\ &\text{subject to } \boldsymbol{x} \in \mathcal{C}, \end{aligned} \tag{2.2}$$

where f is continuously differentiable over the set $\mathcal{C} \subseteq \mathbb{R}^d$. Then a vector $\boldsymbol{d} \neq \boldsymbol{0}$ is called a feasible descent direction at $\boldsymbol{x} \in \mathcal{C}$ if $\nabla f(\boldsymbol{x})^\top \boldsymbol{d} < 0$, and there exists $\epsilon > 0$ such that $\boldsymbol{x} + t\boldsymbol{d} \in \mathcal{C}$ for all $t \in [0, \epsilon]$.

Lemma 2.1 *Consider the problem (2.2). If $\boldsymbol{x}_*$ is a local optimal solution of (2.2), then there are no feasible descent directions at $\boldsymbol{x}_*$.*

Lemma 2.2 *Let $\boldsymbol{x}_*$ be a local minimum of (2.1). Let $I(\boldsymbol{x}_*)$ be the set of active constraints at $\boldsymbol{x}_*$:*

$$I(\boldsymbol{x}_*) = \{i \ : \ g_i(\boldsymbol{x}_*) = 0, \ i = 1, \dots, m\}. \tag{2.3}$$

Then there does not exist a vector $\boldsymbol{d} \in \mathbb{R}^d$ such that

$$\nabla f(\boldsymbol{x}_*)^\top \boldsymbol{d} < 0, \quad \nabla g_i(\boldsymbol{x}_*)^\top \boldsymbol{d} < 0, \ i \in I(\boldsymbol{x}_*). \tag{2.4}$$

2.1 The Nonconvex Case

Lemma 2.3 *(Gordan's alternative theorem) Let $\boldsymbol{A} \in \mathbb{R}^{m \times n}$. Then exactly one of the following two systems has a solution:*

- $\boldsymbol{A}\boldsymbol{x} \prec \boldsymbol{0}$.
- $\boldsymbol{p} \neq \boldsymbol{0}$, $\boldsymbol{A}^\top \boldsymbol{p} = \boldsymbol{0}$, $\boldsymbol{p} \succeq \boldsymbol{0}$.

Theorem 2.1 *(Fritz-John conditions for inequality constrained problems). Let $\boldsymbol{x}_*$ be a local minimum of (2.1). Then there exist multipliers $\lambda_0, \dots, \lambda_m \ge 0$, which are not all zeros, such that*

$$\lambda_0 \nabla f(\boldsymbol{x}_*) + \sum_{i=1}^m \lambda_i \nabla g_i(\boldsymbol{x}_*) = \boldsymbol{0}, \quad \lambda_i g_i(\boldsymbol{x}_*) = 0, \ i = 1, \dots, m. \tag{2.5}$$

Proof *By Lemma 2.2 the following system of inequalities does not have a feasible point:*

$$\nabla f(\boldsymbol{x}_*)^\top \boldsymbol{d} < 0, \quad \nabla g_i(\boldsymbol{x}_*)^\top \boldsymbol{d} < 0, \ i \in I(\boldsymbol{x}_*)$$

where $I(\boldsymbol{x}_*) = \{i \ : \ g_i(\boldsymbol{x}_*) = 0, \ i = 1, \dots, m\} = \{i_1, \dots, i_k\}$. This is equivalent to

$$\boldsymbol{A}\boldsymbol{d} < \boldsymbol{0},$$

where $\boldsymbol{A} = [\nabla f(\boldsymbol{x}_*), \nabla g_{i_1}(\boldsymbol{x}_*), \dots, \nabla g_{i_k}(\boldsymbol{x}_*)]^\top$. By Lemma 2.3 the system is infeasible if and only if there exists a vector $\boldsymbol{\eta} = (\lambda_0, \lambda_{i_1}, \dots, \lambda_{i_k})^\top \neq \boldsymbol{0}$ such that

$$\boldsymbol{A}^\top \boldsymbol{\eta} = \boldsymbol{0}, \quad \boldsymbol{\eta} \succeq \boldsymbol{0},$$

which is the same as

$$\lambda_0 \nabla f(\boldsymbol{x}_*) + \sum_{i \in I(\boldsymbol{x}_*)} \lambda_i \nabla g_i(\boldsymbol{x}_*) = \boldsymbol{0}, \quad \lambda_i g_i(\boldsymbol{x}_*) = 0, \ i \in I(\boldsymbol{x}_*).$$

Defining $\lambda_i = 0$ for any $i \notin I(\boldsymbol{x}_*)$ concludes the proof. □

A major drawback of the Fritz-John conditions is that they allow λ_0 to be zero. The case $\lambda_0 = 0$ is not informative since the conditions becomes

$$\sum_{i=1}^{m} \lambda_i \nabla g_i(\boldsymbol{x}_*) = \mathbf{0}, \tag{2.6}$$

which means that the gradients of the active constraints $\{\nabla g_i(\boldsymbol{x}_*)\}_{i \in I(\boldsymbol{x}_*)}$ are linearly dependent. This condition has nothing to do with the objective function, implying that there might be a lot of points satisfying the Fritz-John conditions which are not local minimum points.

Theorem 2.2 *(KKT conditions for inequality constrained problems) Let $\boldsymbol{x}_*$ be a local minimum of (2.1). Let*

$$I(\boldsymbol{x}_*) = \{i \; : \; g_i(\boldsymbol{x}_*) = 0\} \tag{2.7}$$

be the set of active constraints. Suppose that the active constraints $\{\nabla g_i(\boldsymbol{x}_*)\}_{i \in I(\boldsymbol{x}_*)}$ are linearly independent. Then there exist multipliers $\lambda_1, \ldots, \lambda_m \geq 0$ such that

$$\nabla f(\boldsymbol{x}_*) + \sum_{i=1}^{m} \lambda_i \nabla g_i(\boldsymbol{x}_*) = \mathbf{0}, \quad \textit{and} \quad \lambda_i g_i(\boldsymbol{x}_*) = 0, \; i = 1, \ldots, m. \tag{2.8}$$

Proof *By the Fritz-John conditions, there exist $\tilde{\lambda}_0, \ldots, \tilde{\lambda}_m \geq 0$, which are not all zeros, such that*

$$\tilde{\lambda}_0 \nabla f(\boldsymbol{x}_*) + \sum_{i=1}^{m} \tilde{\lambda}_i \nabla g_i(\boldsymbol{x}_*) = \mathbf{0}, \quad \tilde{\lambda}_i g_i(\boldsymbol{x}_*) = 0, \; i = 1, \ldots, m. \tag{2.9}$$

We have that $\tilde{\lambda}_0 \neq 0$, since otherwise ($\tilde{\lambda}_0 = 0$) by (2.9) we have

$$\sum_{i \in I(\boldsymbol{x}_*)} \tilde{\lambda}_i \nabla g_i(\boldsymbol{x}_*) = \mathbf{0},$$

where not all the scalars $\tilde{\lambda}_i$, $i \in I(\boldsymbol{x}_*)$ are zeros, leading to a contradiction. Defining $\lambda_i = \frac{\tilde{\lambda}_i}{\tilde{\lambda}_0}$, the result directly follows from (2.9). □

Theorem 2.3 *(KKT conditions for inequality/equality constrained problems) Let $\boldsymbol{x}_*$ be a local minimum of the problem*

$$\begin{aligned} &\min f(\boldsymbol{x}) \\ &\textit{subject to } g_i(\boldsymbol{x}) \leq 0, \quad i = 1, \ldots, m, \\ &\qquad\qquad h_i(\boldsymbol{x}) = 0, \quad i = 1, \ldots, n, \end{aligned} \tag{2.10}$$

where f, g_i, h_i are continuously differentiable functions over $\mathbb{R}^d$. Suppose that the gradients of the active constraints and the equality constraints

$$\{\nabla g_i(\boldsymbol{x}_*) \; : \; i \in I(\boldsymbol{x}_*)\} \cup \{\nabla h_i(\boldsymbol{x}_*) \; : \; i = 1, \ldots, n\} \tag{2.11}$$

are linearly independent. Then there exist multipliers $\lambda_1, \ldots, \lambda_m \geq 0$ and $\mu_1, \ldots, \mu_n \in \mathbb{R}$ such that

$$\nabla f(\boldsymbol{x}_*) + \sum_{i=1}^{m} \lambda_i \nabla g_i(\boldsymbol{x}_*) + \sum_{i=1}^{n} \mu_i \nabla h_i(\boldsymbol{x}_*) = \mathbf{0}, \quad \lambda_i g_i(\boldsymbol{x}_*) = 0, \; i = 1, \ldots, m. \tag{2.12}$$

Theorem 2.4 *(sufficiency of the KKT conditions for convex optimization problems) Let* $\boldsymbol{x}_*$ *be a feasible point of*

$$\begin{aligned} &\min f(\boldsymbol{x}) \\ &\textit{subject to } g_i(\boldsymbol{x}) \le 0, \quad i = 1, \ldots, m, \\ &\qquad h_i(\boldsymbol{x}) = 0, \quad i = 1, \ldots, n, \end{aligned} \tag{2.13}$$

where f, g_i *are continuously differentiable convex functions over* $\mathbb{R}^d$ *and* h_i *are affine functions. Suppose that there exist multipliers* $\lambda_1, \ldots, \lambda_m \ge 0$ *and* $\mu_1, \ldots, \mu_n \in \mathbb{R}$ *such that*

$$\nabla f(\boldsymbol{x}_*) + \sum_{i=1}^{m} \lambda_i \nabla g_i(\boldsymbol{x}_*) + \sum_{i=1}^{n} \mu_i \nabla h_i(\boldsymbol{x}_*) = \mathbf{0}, \quad \lambda_i g_i(\boldsymbol{x}_*) = 0, \ i = 1, \ldots, m. \tag{2.14}$$

Then $\boldsymbol{x}_*$ *is an optimal solution of* (2.13).

Proof *Let* $\boldsymbol{x}$ *be a feasible point. Define the convex function*

$$\mathcal{L}(\boldsymbol{x}) = f(\boldsymbol{x}) + \sum_{i=1}^{m} \lambda_i g_i(\boldsymbol{x}) + \sum_{j=1}^{n} \mu_j h_j(\boldsymbol{x}).$$

A feasible point $\boldsymbol{x}_*$ *is a minimizer of* $\mathcal{L}$, *since* $\nabla \mathcal{L}(\boldsymbol{x}_*) = \mathbf{0}$ *and in particular* $\mathcal{L}(\boldsymbol{x}_*) \le \mathcal{L}(\boldsymbol{x})$, *we have*

$$\begin{aligned} f(\boldsymbol{x}_*) &= f(\boldsymbol{x}_*) + \sum_{i=1}^{m} \lambda_i g_i(\boldsymbol{x}_*) + \sum_{i=1}^{n} \mu_i h_i(\boldsymbol{x}_*) \\ &= \mathcal{L}(\boldsymbol{x}_*) \le \mathcal{L}(\boldsymbol{x}) = f(\boldsymbol{x}) + \sum_{i=1}^{m} \lambda_i g_i(\boldsymbol{x}) + \sum_{i=1}^{n} \mu_i h_i(\boldsymbol{x}) \le f(\boldsymbol{x}), \end{aligned}$$

showing that $\boldsymbol{x}_*$ *is the optimal solution.* □

Theorem 2.5 *(necessity of the KKT conditions under Slater's condition). Let* $\boldsymbol{x}_*$ *be an optimal solution of the problem*

$$\begin{aligned} &\min f(\boldsymbol{x}) \\ &\textit{subject to } g_i(\boldsymbol{x}) \le 0, \ i = 1, \ldots, m, \end{aligned} \tag{2.15}$$

where f, g_i *are continuously differentiable functions over* $\mathbb{R}^d$. *In addition,* g_i *are convex functions. Suppose that there exists* $\hat{\boldsymbol{x}} \in \mathbb{R}^d$ *such that*

$$g_i(\hat{\boldsymbol{x}}) < 0, \ i = 1, \ldots, m. \tag{2.16}$$

Then there exist multipliers $\lambda_1, \ldots, \lambda_m \ge 0$ *such that*

$$\nabla f(\boldsymbol{x}_*) + \sum_{i=1}^{m} \lambda_i \nabla g_i(\boldsymbol{x}_*) = \mathbf{0}, \quad \lambda_i g_i(\boldsymbol{x}_*) = 0, \ i = 1, \ldots, m. \tag{2.17}$$

Proof *Since* $\boldsymbol{x}_*$ *is an optimal solution, then the Fritz-John conditions are satisfied. That is, there exist* $\tilde{\lambda}_0, \ldots, \tilde{\lambda}_m \ge 0$, *which are not all zeros, such that*

$$\tilde{\lambda}_0 \nabla f(\boldsymbol{x}_*) + \sum_{i=1}^{m} \tilde{\lambda}_i \nabla g_i(\boldsymbol{x}_*) = \mathbf{0}, \quad \tilde{\lambda}_i g_i(\boldsymbol{x}_*) = 0, \ i = 1, \ldots, m. \tag{2.18}$$

All that we need to show is that $\tilde{\lambda}_0 > 0$, *and then the conditions* (2.17) *will be satisfied with* $\lambda_i = \frac{\tilde{\lambda}_i}{\tilde{\lambda}_0}$.

To prove that $\tilde{\lambda}_0 > 0$, *assume in contradiction that it is zero; then*

$$\sum_{i=1}^{m} \tilde{\lambda}_i \nabla g_i(\boldsymbol{x}_*) = \mathbf{0}. \tag{2.19}$$

By the Slater's condition (2.16)

$$0 > g_i(\hat{\boldsymbol{x}}) \geq g_i(\boldsymbol{x}_*) + \langle \nabla g_i(\boldsymbol{x}_*),\, \hat{\boldsymbol{x}} - \boldsymbol{x}_* \rangle, \quad i = 1, \ldots, m.$$

Multiplying the i*th inequality by* $\tilde{\lambda}_i$ *and summing them, we have*

$$0 > \sum_{i=1}^{m} \tilde{\lambda}_i g_i(\boldsymbol{x}_*) + \left\langle \sum_{i=1}^{m} \tilde{\lambda}_i \nabla g_i(\boldsymbol{x}_*),\ \hat{\boldsymbol{x}} - \boldsymbol{x}_* \right\rangle, \tag{2.20}$$

where the inequality is strict since not all the $\tilde{\lambda}_i$ *are zero. Plugging* (2.17) *and* (2.19) *into* (2.20) *yields* $0 > 0$*, thus concluding the proof.* □

Theorem 2.6 *(necessity of the KKT conditions under the generalized Slater's condition) Let* $\boldsymbol{x}_*$ *be an optimal solution of the problem*

$$\begin{aligned} &\min f(\boldsymbol{x}) \\ &\text{subject to } g_i(\boldsymbol{x}) \leq 0,\ i = 1, \ldots, m, \\ &\qquad\qquad \tilde{g}_i(\boldsymbol{x}) \leq 0,\ i = 1, \ldots, l, \\ &\qquad\qquad h_i(\boldsymbol{x}) = 0,\ i = 1, \ldots, n, \end{aligned} \tag{2.21}$$

where f, g_i *are continuously differentiable convex functions over* $\mathbb{R}^d$*, and* $\tilde{g}_i, h_i$ *are affine. Suppose that there exists* $\hat{\boldsymbol{x}} \in \mathbb{R}^d$ *such that*

$$g_i(\hat{\boldsymbol{x}}) < 0,\ i = 1, \ldots, m, \quad \tilde{g}_i(\hat{\boldsymbol{x}}) \leq 0,\ i = 1, \ldots, l, \quad h_i(\hat{\boldsymbol{x}}) = 0,\ i = 1, \ldots, n. \tag{2.22}$$

Then there exists multipliers $\lambda_1, \ldots, \lambda_m, \eta_1, \ldots, \eta_l \geq 0$*,* $\mu_1, \ldots, \mu_n \in \mathbb{R}$ *such that*

$$\begin{aligned} &\nabla f(\boldsymbol{x}_*) + \sum_{i=1}^{m} \lambda_i \nabla g_i(\boldsymbol{x}_*) + \sum_{i=1}^{l} \eta_i \nabla \tilde{g}_i(\boldsymbol{x}_*) + \sum_{i=1}^{n} \mu_i \nabla h_i(\boldsymbol{x}_*) = \mathbf{0}, \\ &\lambda_i g_i(\boldsymbol{x}_*) = 0,\ i = 1, \ldots, m, \quad \eta_i \tilde{g}_i(\boldsymbol{x}_*) = 0,\ i = 1, \ldots, l. \end{aligned} \tag{2.23}$$

3 Lagrange Dual (see [1, Chapter 12] [2, Chapter 3,5])

Consider the following problem (that is not necessarily convex):

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x}) \tag{3.1}$$
$$\text{subject to } g_i(\boldsymbol{x}) \le 0,\ i = 1,\dots,m,$$
$$h_i(\boldsymbol{x}) = 0,\ i = 1,\dots,n.$$

We assume that its domain $\mathcal{D} = \operatorname{dom} f \bigcap (\cap_{i=1}^m \operatorname{dom} g_i) \bigcap (\cap_{i=1}^n \operatorname{dom} h_i)$ is nonempty, and denote the optimal value by p_*.

3.1 Lagrange Dual Function

The Lagrangian is

$$\mathcal{L}(\boldsymbol{x},\boldsymbol{\lambda},\boldsymbol{\mu}) = f(\boldsymbol{x}) + \sum_{i=1}^m \lambda_i g_i(\boldsymbol{x}) + \sum_{i=1}^n \mu_i h_i(\boldsymbol{x}), \tag{3.2}$$

which is a weighted sum of objective and constraint functions. λ_i and μ_i are Lagrange multipliers associated with $g_i(\boldsymbol{x}) \le 0$ and $h_i(\boldsymbol{x}) = 0$ respectively. The vectors $\boldsymbol{\lambda} = (\lambda_1,\dots,\lambda_m)^\top$ and $\boldsymbol{\mu} = (\mu_1,\dots,\mu_n)^\top$ are called the dual variables or Lagrange multipliers.

The Lagrange dual function (or dual function) is

$$q(\boldsymbol{\lambda},\boldsymbol{\mu}) = \inf_{\boldsymbol{x}\in\mathcal{D}} \mathcal{L}(\boldsymbol{x},\boldsymbol{\lambda},\boldsymbol{\mu}) = \inf_{\boldsymbol{x}\in\mathcal{D}} \left(f(\boldsymbol{x}) + \sum_{i=1}^m \lambda_i g_i(\boldsymbol{x}) + \sum_{i=1}^n \mu_i h_i(\boldsymbol{x}) \right). \tag{3.3}$$

When the Lagrangian is unbounded below in $\boldsymbol{x}$, the dual function takes the value $-\infty$. Since the dual function is the pointwise infimum of a family of affine functions of $(\boldsymbol{\lambda},\boldsymbol{\mu})$, it is concave, even when the problem (3.1) is not convex.

Lemma 3.1 *The dual function yields lower bounds on the optimal value p_*. For any $\boldsymbol{\lambda} \succeq \mathbf{0}$ and $\boldsymbol{\mu}$ we have*

$$q(\boldsymbol{\lambda},\boldsymbol{\mu}) \le p_*. \tag{3.4}$$

Proof *Suppose $\tilde{\boldsymbol{x}}$ is a feasible point. Then we have*

$$q(\boldsymbol{\lambda},\boldsymbol{\mu}) = \inf_{\boldsymbol{x}\in\mathcal{D}} \mathcal{L}(\boldsymbol{x},\boldsymbol{\lambda},\boldsymbol{\mu}) \le \mathcal{L}(\tilde{\boldsymbol{x}},\boldsymbol{\lambda},\boldsymbol{\mu}) = f(\tilde{\boldsymbol{x}}) + \sum_{i=1}^m \lambda_i g_i(\tilde{\boldsymbol{x}}) + \sum_{i=1}^n \mu_i h_i(\tilde{\boldsymbol{x}}) \le f(\tilde{\boldsymbol{x}}).$$

Since this inequality holds for every feasible point $\tilde{\boldsymbol{x}}$, the inequality (3.4) follows. □

Example 3.1 Consider the following least-norm solution of linear equations:

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} \boldsymbol{x}^\top \boldsymbol{x}$$
$$\text{subject to } \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b},$$

where $\boldsymbol{A} \in \mathbb{R}^{p\times d}$. The Lagrangian is

$$\mathcal{L}(\boldsymbol{x},\boldsymbol{\mu}) = \boldsymbol{x}^\top\boldsymbol{x} + \boldsymbol{\mu}^\top(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}).$$

Using $\nabla_{\boldsymbol{x}}\mathcal{L}(\boldsymbol{x},\boldsymbol{\mu}) = 2\boldsymbol{x} + \boldsymbol{A}^\top\boldsymbol{\mu} = \mathbf{0}$, we have the corresponding dual function

$$q(\boldsymbol{\mu}) = \inf_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x},\boldsymbol{\mu}) = \mathcal{L}\left(-\frac{1}{2}\boldsymbol{A}^\top\boldsymbol{\mu},\boldsymbol{\mu}\right) = -\frac{1}{4}\boldsymbol{\mu}^\top\boldsymbol{A}\boldsymbol{A}^\top\boldsymbol{\mu} - \boldsymbol{b}^\top\boldsymbol{\mu},$$

which is a concave quadratic function. The lower bound property states that for any $\boldsymbol{\mu}$, we have

$$-\frac{1}{4}\boldsymbol{\mu}^\top\boldsymbol{A}\boldsymbol{A}^\top\boldsymbol{\mu} - \boldsymbol{b}^\top\boldsymbol{\mu} \le p_*.$$

Example 3.2 Consider an LP in standard form:

$$\begin{aligned} &\min_{\boldsymbol{x}} \boldsymbol{c}^\top \boldsymbol{x} \\ &\text{subject to } \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}, \\ &\qquad\qquad \boldsymbol{x} \succeq \boldsymbol{0}. \end{aligned}$$

The Lagrangian is

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \boldsymbol{c}^\top \boldsymbol{x} - \boldsymbol{\lambda}^\top \boldsymbol{x} + \boldsymbol{\mu}^\top (\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}).$$

The dual function is

$$\begin{aligned} q(\boldsymbol{\lambda}, \boldsymbol{\mu}) &= \inf_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = -\boldsymbol{b}^\top \boldsymbol{\mu} + \inf_{\boldsymbol{x}} \left\{ (\boldsymbol{c} + \boldsymbol{A}^\top \boldsymbol{\mu} - \boldsymbol{\lambda})^\top \boldsymbol{x} \right\} \\ &= \begin{cases} -\boldsymbol{b}^\top \boldsymbol{\mu}, & \boldsymbol{c} + \boldsymbol{A}^\top \boldsymbol{\mu} - \boldsymbol{\lambda} = \boldsymbol{0}, \\ -\infty, & \text{otherwise.} \end{cases} \end{aligned}$$

The lower bound property states that

$$-\boldsymbol{b}^\top \boldsymbol{\mu} \le p_* \text{ if } \boldsymbol{A}^\top \boldsymbol{\mu} + \boldsymbol{c} \succeq \boldsymbol{0}.$$

3.2 Lagrange Dual Function and Conjugate Function

The conjugate f^* of a function $f : \mathbb{R}^d \to \mathbb{R}$ is given by

$$f^*(\boldsymbol{y}) = \sup_{\boldsymbol{x} \in \operatorname{dom} f} (\boldsymbol{y}^\top \boldsymbol{x} - f(\boldsymbol{x})).$$

f^* is convex even if f is not convex, since it is the pointwise supremum of a family of affine functions of $\boldsymbol{y}$.

Example 3.3 Examples of conjugate functions.

- $f(x) = ax + b$:

$$f^*(y) = \sup_x (yx - ax - b) = \begin{cases} -b, & y = a, \\ \infty, & \text{otherwise.} \end{cases}$$

- $f(x) = -\log x$ with $\operatorname{dom} f = \mathbb{R}_{++}$:

$$f^*(y) = \sup_{x>0} (yx + \log x) = \begin{cases} -1 - \log(-y), & y < 0, \\ \infty, & \text{otherwise.} \end{cases}$$

- $f(\boldsymbol{x}) = \frac{1}{2} \boldsymbol{x}^\top \boldsymbol{Q} \boldsymbol{x}$ with $\boldsymbol{Q} \in \mathbb{S}^d_{++}$:

$$f^*(\boldsymbol{y}) = \sup_{\boldsymbol{x}} \left(\boldsymbol{y}^\top \boldsymbol{x} - \frac{1}{2} \boldsymbol{x}^\top \boldsymbol{Q} \boldsymbol{x} \right) = \frac{1}{2} \boldsymbol{y}^\top \boldsymbol{Q}^{-1} \boldsymbol{y}$$

The conjugate function and Lagrange dual function are closely related. If conjugate of f is known, the derivation of dual is simplified.

Example 3.4 Consider the following problem with linear inequality and equality constraint:

$$\begin{aligned} &\min_{\boldsymbol{x}} f(\boldsymbol{x}) \\ &\text{subject to } \boldsymbol{A}\boldsymbol{x} \preceq \boldsymbol{b}, \\ &\qquad\qquad \boldsymbol{C}\boldsymbol{x} = \boldsymbol{d}. \end{aligned}$$

Using the conjugate of f, we can write the dual function as

$$
\begin{aligned}
q(\boldsymbol{\lambda}, \boldsymbol{\mu}) &= \inf_{\boldsymbol{x}} \left\{ f(\boldsymbol{x}) + \boldsymbol{\lambda}^\top (\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}) + \boldsymbol{\mu}^\top (\boldsymbol{C}\boldsymbol{x} - \boldsymbol{d}) \right\} \\
&= -\boldsymbol{b}^\top \boldsymbol{\lambda} - \boldsymbol{d}^\top \boldsymbol{\mu} + \inf_{\boldsymbol{x}} \left\{ f(\boldsymbol{x}) + (\boldsymbol{A}^\top \boldsymbol{\lambda} + \boldsymbol{C}^\top \boldsymbol{\mu})^\top \boldsymbol{x} \right\} \\
&= -\boldsymbol{b}^\top \boldsymbol{\lambda} - \boldsymbol{d}^\top \boldsymbol{\mu} - f^*(-\boldsymbol{A}^\top \boldsymbol{\lambda} - \boldsymbol{C}^\top \boldsymbol{\mu}).
\end{aligned}
$$

The domain of q follows from the domain of f^*:

$$
\operatorname{dom} q = \{ (\boldsymbol{\lambda}, \boldsymbol{\mu}) \ : \ -\boldsymbol{A}^\top \boldsymbol{\lambda} - \boldsymbol{C}^\top \boldsymbol{\mu} \in \operatorname{dom} f^* \}.
$$

Example 3.5 Consider the following equality constrained norm minimization:

$$
\begin{aligned}
&\min_{\boldsymbol{x}} \ ||\boldsymbol{x}|| \\
&\text{subject to } \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b},
\end{aligned}
$$

where $||\cdot||$ is any norm. The conjugate of $f = ||\cdot||$ is given by

$$
f^*(\boldsymbol{y}) = \begin{cases} 0, & ||\boldsymbol{y}||_* \le 1, \\ \infty, & \text{otherwise,} \end{cases}
$$

which is the indicator function of the dual norm unit ball. Then the dual function is

$$
q(\boldsymbol{\mu}) = -\boldsymbol{b}^\top \boldsymbol{\mu} - f^*(-\boldsymbol{A}^\top \boldsymbol{\mu}) = \begin{cases} -\boldsymbol{b}^\top \boldsymbol{\mu}, & ||\boldsymbol{A}^\top \boldsymbol{\mu}||_* \le 1, \\ -\infty, & \text{otherwise.} \end{cases}
$$

Example 3.6 Consider the entropy maximization problem:

$$
\begin{aligned}
&\min_{\boldsymbol{x}} \ \sum_{i=1}^{d} x_i \log x_i \\
&\text{subject to } \boldsymbol{A}\boldsymbol{x} \preceq \boldsymbol{b}, \\
&\qquad\qquad \boldsymbol{1}^\top \boldsymbol{x} = 1.
\end{aligned}
$$

The conjugate of a sum of negative entropy functions is

$$
f^*(\boldsymbol{y}) = \sup_{\boldsymbol{x} \in \mathbb{R}^d_{++}} \left(\boldsymbol{y}^\top \boldsymbol{x} - \sum_{i=1}^{d} x_i \log x_i \right) = \sum_{i=1}^{d} e^{y_i - 1}.
$$

Then the dual function is

$$
q(\boldsymbol{\lambda}, \mu) = -\boldsymbol{b}^\top \boldsymbol{\lambda} - \mu - \sum_{i=1}^{d} e^{-\boldsymbol{a}_i^\top \boldsymbol{\lambda} - \mu - 1} = -\boldsymbol{b}^\top \boldsymbol{\lambda} - \mu - e^{-\mu - 1} \sum_{i=1}^{d} e^{-\boldsymbol{a}_i^\top \boldsymbol{\lambda}},
$$

where $\boldsymbol{a}_i$ is the ith column of $\boldsymbol{A}$.

3.3 Lagrange Dual Problem

What is the best lower bound that can be obtained from the Lagrange dual function? This question leads to the following Lagrange dual problem associated with the primal problem (3.1):

$$
\begin{aligned}
&\max_{\boldsymbol{\lambda}, \boldsymbol{\mu}} \ q(\boldsymbol{\lambda}, \boldsymbol{\mu}) \\
&\text{subject to } \boldsymbol{\lambda} \succeq \boldsymbol{0}.
\end{aligned}
\tag{3.5}
$$

A pair $(\boldsymbol{\lambda}, \boldsymbol{\mu})$ is dual feasible if it satisfies $\boldsymbol{\lambda} \succeq \boldsymbol{0}$ and $q(\boldsymbol{\lambda}, \boldsymbol{\mu}) > -\infty$. An optimal solution $(\boldsymbol{\lambda}_*, \boldsymbol{\mu}_*)$ of (3.5) is called dual optimal or optimal Lagrange multipliers. The Lagrange dual problem (3.5) is a convex optimization problem whether or not the primal problem (3.1) is convex.

Example 3.7 Lagrange dual of standard form LP. The Lagrange dual for the standard form LP

$$\begin{aligned}\min_{\boldsymbol{x}}\ & \boldsymbol{c}^\top \boldsymbol{x}\\ \text{subject to } & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}\\ & \boldsymbol{x} \succeq \boldsymbol{0}\end{aligned}$$

is given by

$$q(\boldsymbol{\lambda}, \boldsymbol{\mu}) = \begin{cases} -\boldsymbol{b}^\top \boldsymbol{\mu}, & \boldsymbol{A}^\top \boldsymbol{\mu} - \boldsymbol{\lambda} + \boldsymbol{c} = \boldsymbol{0},\\ -\infty, & \text{otherwise.}\end{cases}$$

The Lagrange dual problem is

$$\begin{aligned}\max_{\boldsymbol{\lambda},\boldsymbol{\mu}}\ & \left[q(\boldsymbol{\lambda}, \boldsymbol{\mu}) \equiv \begin{cases} -\boldsymbol{b}^\top \boldsymbol{\mu}, & \boldsymbol{A}^\top \boldsymbol{\mu} - \boldsymbol{\lambda} + \boldsymbol{c} = \boldsymbol{0},\\ -\infty, & \text{otherwise.}\end{cases}\right]\\ \text{subject to } & \boldsymbol{\lambda} \succeq \boldsymbol{0}.\end{aligned}$$

This is equivalent to

$$\begin{aligned}\max_{\boldsymbol{\lambda},\boldsymbol{\mu}}\ & \left\{-\boldsymbol{b}^\top \boldsymbol{\mu}\right\}\\ \text{subject to } & \boldsymbol{A}^\top \boldsymbol{\mu} - \boldsymbol{\lambda} + \boldsymbol{c} = \boldsymbol{0}\\ & \boldsymbol{\lambda} \succeq \boldsymbol{0},\end{aligned}$$

and

$$\begin{aligned}\max_{\boldsymbol{\mu}}\ & \left\{-\boldsymbol{b}^\top \boldsymbol{\mu}\right\}\\ \text{subject to } & \boldsymbol{A}^\top \boldsymbol{\mu} + \boldsymbol{c} \succeq \boldsymbol{0}.\end{aligned}$$

Theorem 3.1 *(weak duality theorem) Consider the primal problem* (3.1) *and its dual problem* (3.5)*. Then,*

$$d_* \le p_* \tag{3.6}$$

where p_* and d_* are optimal primal and dual values respectively.

The difference $p_* - d_*$ is called the optimal duality gap, which is always nonnegative.

Theorem 3.2 *(strong duality theorem) Consider*

$$\begin{aligned}\min_{\boldsymbol{x}}\ & f(\boldsymbol{x}) \qquad\qquad (3.7)\\ \textit{subject to } & g_i(\boldsymbol{x}) \le 0,\ i = 1, \dots, m,\\ & \tilde{g}_i(\boldsymbol{x}) \le 0,\ i = 1, \dots, l,\\ & h_i(\boldsymbol{x}) = 0,\ i = 1, \dots, n,\end{aligned}$$

where f, g_i are convex functions over $\mathbb{R}^d$, and $\tilde{g}_i, h_i$ are affine. Suppose that there exists $\hat{\boldsymbol{x}} \in \mathbb{R}^d$ such that

$$g_i(\hat{\boldsymbol{x}}) < 0,\ i = 1, \dots, m, \quad \tilde{g}_i(\hat{\boldsymbol{x}}) \le 0,\ i = 1, \dots, l, \quad h_i(\hat{\boldsymbol{x}}) = 0,\ i = 1, \dots, n. \tag{3.8}$$

*Then if the problem has a finite optimal value p_*, the dual optimal value d_* is attained (*i.e.*, there exists a dual feasible $(\boldsymbol{\lambda}_*, \boldsymbol{\mu}_*)$ with $q(\boldsymbol{\lambda}_*, \boldsymbol{\mu}_*) = d_* = p_*$), and the optimal values of the primal and dual problems are same:*

$$d_* = p_*.$$

There exist conditions called constraint qualifications other than Slater's condition.

Example 3.8 Consider

$$\min_{\boldsymbol{x}} \boldsymbol{x}^\top \boldsymbol{x}$$
$$\text{subject to } \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b},$$

and its dual problem

$$\max_{\boldsymbol{\mu}} \left\{ -\frac{1}{4}\boldsymbol{\mu}^\top \boldsymbol{A}\boldsymbol{A}^\top \boldsymbol{\mu} - \boldsymbol{b}^\top \boldsymbol{\mu} \right\}.$$

Slater's condition is simply that the primal problem is feasible, so $p_* = d_*$ if $\boldsymbol{b} \in \mathcal{R}(\boldsymbol{A})$, *i.e.*, $p_* < \infty$.

3.4 Saddle-point Interpretation

For simplicity, assume that there are no equality constraints. Then, we have

$$\sup_{\boldsymbol{\lambda} \succeq \boldsymbol{0}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}) = \sup_{\boldsymbol{\lambda} \succeq \boldsymbol{0}} \left(f(\boldsymbol{x}) + \sum_{i=1}^{m} \lambda_i g_i(\boldsymbol{x}) \right) = \begin{cases} f(\boldsymbol{x}), & g_i(\boldsymbol{x}) \leq 0,\ i = 1, \ldots, m, \\ \infty, & \text{otherwise}, \end{cases} \tag{3.9}$$

which leads to

$$p_* = \inf_{\boldsymbol{x}} \sup_{\boldsymbol{\lambda} \succeq \boldsymbol{0}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}). \tag{3.10}$$

By definition, we also have

$$d_* = \sup_{\boldsymbol{\lambda} \succeq \boldsymbol{0}} \inf_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}). \tag{3.11}$$

Thus, weak duality can be expressed as the inequality

$$\sup_{\boldsymbol{\lambda} \succeq \boldsymbol{0}} \inf_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}) \leq \inf_{\boldsymbol{x}} \sup_{\boldsymbol{\lambda} \succeq \boldsymbol{0}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}) \tag{3.12}$$

and strong duality as the equality

$$\sup_{\boldsymbol{\lambda} \succeq \boldsymbol{0}} \inf_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}) = \inf_{\boldsymbol{x}} \sup_{\boldsymbol{\lambda} \succeq \boldsymbol{0}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}). \tag{3.13}$$

Strong duality means that the order of the minimization over $\boldsymbol{x}$ and the maximization over $\boldsymbol{\lambda} \succeq \boldsymbol{0}$ can be switched without affecting the result. This is called the saddle-point property.

We refer to a pair $\tilde{\boldsymbol{w}} \in \mathcal{W}$ and $\tilde{\boldsymbol{z}} \in \mathcal{Z}$ as a saddle-point for f (and $\mathcal{W}$ and $\mathcal{Z}$) if

$$f(\tilde{\boldsymbol{w}}, \boldsymbol{z}) \leq f(\tilde{\boldsymbol{w}}, \tilde{\boldsymbol{z}}) \leq f(\boldsymbol{w}, \tilde{\boldsymbol{z}}) \tag{3.14}$$

for all $\boldsymbol{w} \in \mathcal{W}$ and $\boldsymbol{z} \in \mathcal{Z}$. In other words, $\tilde{\boldsymbol{w}}$ minimizes $f(\boldsymbol{w}, \tilde{\boldsymbol{z}})$ over $\boldsymbol{w} \in \mathcal{W}$ and $\tilde{\boldsymbol{z}}$ maximizes $f(\tilde{\boldsymbol{w}}, \boldsymbol{z})$ over $\boldsymbol{z} \in \mathcal{Z}$, *i.e.*,

$$f(\tilde{\boldsymbol{w}}, \tilde{\boldsymbol{z}}) = \inf_{\boldsymbol{w} \in \mathcal{W}} f(\boldsymbol{w}, \tilde{\boldsymbol{z}}) = \sup_{\boldsymbol{z} \in \mathcal{Z}} f(\tilde{w}, \boldsymbol{z}), \tag{3.15}$$

which implies that the saddle-point property holds.

If $\boldsymbol{x}_*$ and $\boldsymbol{\lambda}_*$ are primal and dual optimal points for a problem in which strong duality holds, they form a saddle-point for the Lagrangian. The converse is also true.

3.5 Optimality Conditions

3.5.1 Certificate of Suboptimality and Stopping Criteria

If $\boldsymbol{x}$ is primal feasible and $(\boldsymbol{\lambda}, \boldsymbol{\mu})$ is dual feasible, we have

$$f(\boldsymbol{x}) - p_* \leq f(\boldsymbol{x}) - q(\boldsymbol{\lambda}, \boldsymbol{\mu}). \tag{3.16}$$

In particular, this establishes that $\boldsymbol{x}$ is ϵ-suboptimal, with $\epsilon = f(\boldsymbol{x}) - q(\boldsymbol{\lambda}, \boldsymbol{\mu})$. It also establishes that $(\boldsymbol{\lambda}, \boldsymbol{\mu})$ is ϵ-suboptimal for the dual problem.

We refer to the gap between primal and dual objectives

$$f(\boldsymbol{x}) - q(\boldsymbol{\lambda}, \boldsymbol{\mu}), \tag{3.17}$$

as the duality gap associated with the primal feasible point $\boldsymbol{x}$ and dual feasible point $(\boldsymbol{\lambda}, \boldsymbol{\mu})$.

Suppose an algorithm produces a sequence of primal feasible $\boldsymbol{x}_k$ and dual feasible $(\boldsymbol{\lambda}_k, \boldsymbol{\mu}_k)$ for $k = 1, 2, \ldots,$. For given required absolute accuracy $\epsilon > 0$, the stopping criteria

$$f(\boldsymbol{x}_k) - q(\boldsymbol{\lambda}_k, \boldsymbol{\mu}_k) \leq \epsilon \tag{3.18}$$

guarantees that $\boldsymbol{x}_k$ is ϵ-suboptimal.

3.5.2 Complementary Slackness

Suppose that strong duality holds, and we have

$$\begin{aligned} f(\boldsymbol{x}_*) = q(\boldsymbol{\lambda}_*, \boldsymbol{\mu}_*) &= \inf_{\boldsymbol{x}} \left(f(\boldsymbol{x}) + \sum_{i=1}^{m} \lambda_{*,i} g_i(\boldsymbol{x}) + \sum_{i=1}^{n} \mu_{*,i} h_i(\boldsymbol{x}) \right) \\ &\leq f(\boldsymbol{x}_*) + \sum_{i=1}^{m} \lambda_{*,i} g_i(\boldsymbol{x}_*) + \sum_{i=1}^{n} \mu_{*,i} h_i(\boldsymbol{x}_*) \leq f(\boldsymbol{x}_*). \end{aligned} \tag{3.19}$$

We thus have that $\boldsymbol{x}_*$ minimizes $\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}_*, \boldsymbol{\mu}_*)$, and we also have that

$$\lambda_{*,i} g_i(\boldsymbol{x}_*) = 0, \quad i = 1, \ldots, m, \tag{3.20}$$

which is known as complementary slackness.

3.5.3 KKT Optimality Conditions

We already studied this in the previous section!

3.6 Solving the Primal Problem via the Dual

If strong duality holds and a dual optimal solution $(\boldsymbol{\lambda}_*, \boldsymbol{\mu}_*)$ exists, then any primal optimal point is also a minimizer of $\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}_*, \boldsymbol{\mu}_*)$. This fact sometimes allows us to compute a primal optimal solution from a dual optimal solution. This is useful when the dual problem is easier to solve than the primal problem.

Example 3.9 Consider the entropy maximization problem:

$$\begin{aligned} \min_{\boldsymbol{x}} \ & \sum_{i=1}^{d} x_i \log x_i \\ \text{subject to } & \boldsymbol{A}\boldsymbol{x} \preceq \boldsymbol{b}, \\ & \boldsymbol{1}^\top \boldsymbol{x} = 1 \end{aligned}$$

with domain $\mathbb{R}^d_{++}$, and its dual problem

$$\max_{\boldsymbol{\lambda},\mu} \left\{ -\boldsymbol{b}^\top \boldsymbol{\lambda} - \mu - e^{-\mu-1} \sum_{i=1}^{d} e^{-\boldsymbol{a}_i^\top \boldsymbol{\lambda}} \right\}$$
$$\text{subject to } \boldsymbol{\lambda} \succeq \mathbf{0},$$

where $\boldsymbol{a}_i$ is the ith column of $\boldsymbol{A}$.

Assume that the strong duality holds and an dual optimal solution $(\boldsymbol{\lambda}_*, \mu_*)$ exists. Suppose we have solved the dual problem. The Lagrangian at $(\boldsymbol{\lambda}_*, \mu_*)$ is

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}_*, \mu_*) = \sum_{i=1}^{d} x_i \log x_i + \boldsymbol{\lambda}_*^\top (\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}) + \mu_*(\mathbf{1}^\top \boldsymbol{x} - 1),$$

which is strictly convex on its domain $\mathcal{D}$ and bounded below. It has a unique solution $\boldsymbol{x}_*$:

$$x_{*,i} = e^{-\boldsymbol{a}_i^\top \boldsymbol{\lambda}_* - \mu_* - 1}, \quad i = 1, \ldots, d.$$

If $\boldsymbol{x}_*$ is primal feasible, it must be the primal optimal solution. Otherwise, the primal optimum is not attained.

3.7 Lagrange Dual and Problem Reformulation

3.7.1 Introducing new variables and equality constraints

Consider an unconstrained problem of the form

$$\min_{\boldsymbol{x}} f(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}). \tag{3.21}$$

Its Lagrange dual function is the constant p_*. So while we have strong duality, the Lagrange dual is not useful.

Reformulate the problem as

$$\min_{\boldsymbol{x},\boldsymbol{y}} f(\boldsymbol{y}) \quad \text{subject to } \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b} = \boldsymbol{y}. \tag{3.22}$$

The Lagrangian is

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\mu}) = f(\boldsymbol{y}) + \boldsymbol{\mu}^\top (\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b} - \boldsymbol{y}), \tag{3.23}$$

and the dual function is

$$q(\boldsymbol{\mu}) = \begin{cases} \boldsymbol{b}^\top \boldsymbol{\mu} + \inf_{\boldsymbol{y}} (f(\boldsymbol{y}) - \boldsymbol{\mu}^\top \boldsymbol{y}) = \boldsymbol{b}^\top \boldsymbol{\mu} - f^*(\boldsymbol{\mu}), & \boldsymbol{A}^\top \boldsymbol{\mu} = \mathbf{0}, \\ -\infty, & \text{otherwise.} \end{cases} \tag{3.24}$$

Then the dual problem is

$$\max_{\boldsymbol{\mu}} \left\{ \boldsymbol{b}^\top \boldsymbol{\mu} - f^*(\boldsymbol{\mu}) \right\} \quad \text{subject to } \boldsymbol{A}^\top \boldsymbol{\mu} = \mathbf{0}. \tag{3.25}$$

Example 3.10 Consider the unconstrained norm approximation problem

$$\min_{\boldsymbol{x}} ||\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}||$$

where $|| \cdot ||$ is any norm and $||\boldsymbol{v}||_* = \sup_{||\boldsymbol{u}|| \leq 1} \boldsymbol{u}^\top \boldsymbol{v}$ is a dual norm of $|| \cdot ||$.

Reformulate the problem as

$$\min_{\boldsymbol{x},\boldsymbol{y}} ||\boldsymbol{y}||$$
$$\text{subject to } \boldsymbol{A}\boldsymbol{x} - \boldsymbol{b} = \boldsymbol{y}.$$

The Lagrangian is

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\mu}) = ||\boldsymbol{y}|| + \boldsymbol{\mu}^\top(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b} - \boldsymbol{y}),$$

and the dual function is

$$q(\boldsymbol{\mu}) = \begin{cases} -\boldsymbol{b}^\top\boldsymbol{\mu} + \inf_{\boldsymbol{y}}(||\boldsymbol{y}|| - \boldsymbol{\mu}^\top\boldsymbol{y}), & \boldsymbol{A}^\top\boldsymbol{\mu} = \boldsymbol{0}, \\ -\infty, & \text{otherwise.} \end{cases}$$

Using the fact that the conjugate of a norm is the indicator function of the dual norm unit ball, we have the following dual problem:

$$\max_{\boldsymbol{\mu}} \left\{-\boldsymbol{b}^\top\boldsymbol{\mu}\right\}$$
$$\text{subject to } ||\boldsymbol{\mu}||_* \leq 1$$
$$\boldsymbol{A}^\top\boldsymbol{\mu} = \boldsymbol{0}.$$

3.7.2 Implicit constraints

Including some of the constraints in the objective function, by modifying the objective function to be infinite when constraint is violated, can yield different formulation of dual problem.

Example 3.11 Consider the linear program with box constraints:

$$\min_{\boldsymbol{x}} \boldsymbol{c}^\top\boldsymbol{x}$$
$$\text{subject to } \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$$
$$\boldsymbol{l} \preceq \boldsymbol{x} \preceq \boldsymbol{u}$$

where $\boldsymbol{l} \prec \boldsymbol{u}$. The dual function is

$$\begin{aligned} q(\boldsymbol{\lambda}, \boldsymbol{\eta}, \boldsymbol{\mu}) &= \inf_{\boldsymbol{x}} \left(\boldsymbol{c}^\top\boldsymbol{x} + \boldsymbol{\lambda}^\top(\boldsymbol{x} - \boldsymbol{u}) + \boldsymbol{\eta}^\top(-\boldsymbol{x} + \boldsymbol{l}) + \boldsymbol{\mu}^\top(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b})\right) \\ &= \begin{cases} -\boldsymbol{b}^\top\boldsymbol{\mu} - \boldsymbol{\lambda}^\top\boldsymbol{u} + \boldsymbol{\eta}^\top\boldsymbol{l}, & \boldsymbol{A}^\top\boldsymbol{\mu} + \boldsymbol{\lambda} - \boldsymbol{\eta} + \boldsymbol{c} = \boldsymbol{0}, \\ -\infty, & \text{otherwise,} \end{cases} \end{aligned}$$

and the dual problem is

$$\max_{\boldsymbol{\lambda},\boldsymbol{\eta},\boldsymbol{\mu}} \left[-\boldsymbol{b}^\top\boldsymbol{\mu} - \boldsymbol{\lambda}^\top\boldsymbol{u} + \boldsymbol{\eta}^\top\boldsymbol{l}\right]$$
$$\text{subject to } \boldsymbol{A}^\top\boldsymbol{\mu} + \boldsymbol{\lambda} - \boldsymbol{\eta} + \boldsymbol{c} = \boldsymbol{0},$$
$$\boldsymbol{\lambda} \succeq \boldsymbol{0},\ \boldsymbol{\eta} \succeq \boldsymbol{0}.$$

Instead reformulate the problem as

$$\min_{\boldsymbol{x}} \left[f(\boldsymbol{x}) \equiv \begin{cases} \boldsymbol{c}^\top\boldsymbol{x}, & \boldsymbol{l} \preceq \boldsymbol{x} \preceq \boldsymbol{u}, \\ \infty, & \text{otherwise.} \end{cases}\right]$$
$$\text{subject to } \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}.$$

The dual function is

$$q(\boldsymbol{\mu}) = \inf_{\boldsymbol{l} \preceq \boldsymbol{x} \preceq \boldsymbol{u}} \left(\boldsymbol{c}^\top\boldsymbol{x} + \boldsymbol{\mu}^\top(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b})\right) = -\boldsymbol{b}^\top\boldsymbol{\mu} + \boldsymbol{u}^\top \min\{\boldsymbol{A}^\top\boldsymbol{\mu} + \boldsymbol{c}, \boldsymbol{0}\} + \boldsymbol{l}^\top \max\{\boldsymbol{A}^\top\boldsymbol{\mu} + \boldsymbol{c}, \boldsymbol{0}\},$$

and the dual problem is

$$\max_{\boldsymbol{\mu}} \left[-\boldsymbol{b}^\top\boldsymbol{\mu} + \boldsymbol{u}^\top \min\{\boldsymbol{A}^\top\boldsymbol{\mu} + \boldsymbol{c}, \boldsymbol{0}\} + \boldsymbol{l}^\top \max\{\boldsymbol{A}^\top\boldsymbol{\mu} + \boldsymbol{c}, \boldsymbol{0}\}\right],$$

which is different from the dual of the original problem.

3.8 Genearalized Inequalities

Consider the problem:

$$\min_{\boldsymbol{x}} f(\boldsymbol{x}) \tag{3.26}$$
$$\text{subject to } g_i(\boldsymbol{x}) \preceq_{\mathcal{K}_i} \mathbf{0}, \quad i = 1, \ldots, m,$$
$$h_i(\boldsymbol{x}) = 0, \quad i = 1, \ldots, n,$$

where $\mathcal{K}_i \subseteq \mathbb{R}^{k_i}$ are proper cones.

3.8.1 Dual Cones

Definition 3.1 Let $\mathcal{K}$ be a cone. The set

$$\mathcal{K}^* = \{\boldsymbol{y} \ : \ \boldsymbol{x}^\top \boldsymbol{y} \geq 0 \text{ for all } \boldsymbol{x} \in \mathcal{K}\} \tag{3.27}$$

is called the *dual cone* of $\mathcal{K}$.

Example 3.12 Examples of dual cones.

- The dual cone of a subspace $\mathcal{V} \subseteq \mathbb{R}^d$ is its orthogonal complement $\mathcal{V}^\perp = \{\boldsymbol{y} \ : \ \boldsymbol{y}^\top \boldsymbol{v} = 0 \text{ for all } \boldsymbol{v} \in \mathcal{V}\}$.
- Nonnegative orthant $\mathbb{R}^d_+$ is self-dual since $\boldsymbol{y}^\top \boldsymbol{x} \geq 0$ for all $\boldsymbol{x} \succeq \mathbf{0} \Longleftrightarrow \boldsymbol{y} \succeq \mathbf{0}$.
- Positive semidefinite cone $\mathbb{S}_+$ is self-dual since $\mathsf{tr}\{\boldsymbol{XY}\} \geq 0$ for all $\boldsymbol{X} \succeq \mathbf{0} \Longleftrightarrow \boldsymbol{Y} \succeq \mathbf{0}$.

3.8.2 Lagrange Dual

The Lagrangian is

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = f(\boldsymbol{x}) + \sum_{i=1}^{m} \boldsymbol{\lambda}_i^\top g_i(\boldsymbol{x}) + \sum_{i=1}^{n} \mu_i h_i(\boldsymbol{x}) \tag{3.28}$$

where $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_1^\top, \ldots, \boldsymbol{\lambda}_m^\top)^\top$ and $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^\top$ are Lagrange multipliers. The dual function is

$$q(\boldsymbol{\lambda}, \boldsymbol{\mu}) = \inf_{\boldsymbol{x} \in \mathcal{D}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \inf_{\boldsymbol{x} \in \mathcal{D}} \left(f(\boldsymbol{x}) + \sum_{i=1}^{m} \boldsymbol{\lambda}_i^\top g_i(\boldsymbol{x}) + \sum_{i=1}^{n} \mu_i h_i(\boldsymbol{x}) \right). \tag{3.29}$$

The nonnegativity requirement on the dual variables is

$$\boldsymbol{\lambda}_i \succeq_{\mathcal{K}_i^*} \mathbf{0}, \tag{3.30}$$

where $\mathcal{K}_i^*$ denotes the dual cone of $\mathcal{K}_i$. In other words, the Lagrange multipliers associated with inequalites must be dual nonnegative.

The Lagrange dual problem is

$$\max_{\boldsymbol{x}} q(\boldsymbol{\lambda}, \boldsymbol{\mu}) \tag{3.31}$$
$$\text{subject to } \boldsymbol{\lambda}_i \succeq_{\mathcal{K}_i^*} \mathbf{0}, \quad i = 1, \ldots, m. \tag{3.32}$$

We always have weak duality, *i.e.*, $d_* \leq p_*$, where p_* and d_* denote the optimal value of primal and dual problems respectively. Strong duality ($d_* = p_*$) holds, for example, when the primal problem is convex and satisfes a generalized version of Slater's condition.

Example 3.13 Lagrange dual of semidefinite program.

$$\min_{\boldsymbol{x}} \boldsymbol{c}^\top \boldsymbol{x}$$
$$\text{subject to } x_1 \boldsymbol{F}_1 + \ldots + x_d \boldsymbol{F}_d + \boldsymbol{G} \preceq \mathbf{0}$$

where $\boldsymbol{F}_1, \ldots, \boldsymbol{F}_d, \boldsymbol{G} \in \mathbb{S}^k$. (Here, g_1 is affine, and $\mathcal{K}_1$ is the positive semidefinite cone $\mathbb{S}^k_+$.)

The Lagrangian is

$$\begin{aligned}\mathcal{L}(\boldsymbol{x}, \boldsymbol{Z}) &= \boldsymbol{c}^\top \boldsymbol{x} + \mathsf{tr}\{\boldsymbol{Z}(x_1\boldsymbol{F}_1 + \ldots + x_d\boldsymbol{F}_d + \boldsymbol{G})\} \\ &= x_1(c_1 + \mathsf{tr}\{\boldsymbol{F}_1\boldsymbol{Z}\}) + \ldots + x_d(c_d + \mathsf{tr}\{\boldsymbol{F}_1\boldsymbol{Z}\}) + \mathsf{tr}\{\boldsymbol{G}\boldsymbol{Z}\}\end{aligned}$$

where $\boldsymbol{Z} \in \mathbb{S}^k$ is a Lagrange multiplier matrix. The dual function is given by

$$q(\boldsymbol{Z}) = \inf_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{Z}) = \begin{cases} \mathsf{tr}\{\boldsymbol{G}\boldsymbol{Z}\}, & \mathsf{tr}\{\boldsymbol{F}_i\boldsymbol{Z}\} + c_i = 0,\ i = 1, \ldots, d, \\ -\infty, & \text{otherwise.} \end{cases}$$

Therefore, the dual problem is

$$\begin{aligned} \max_{\boldsymbol{Z}} \ & \mathsf{tr}\{\boldsymbol{G}\boldsymbol{Z}\} \\ \text{subject to } \ & \mathsf{tr}\{\boldsymbol{F}_i\boldsymbol{Z}\} + c_i = 0, \quad i = 1, \ldots, d, \\ & \boldsymbol{Z} \succeq \boldsymbol{0}. \end{aligned}$$

Strong duality holds if the the primal problem is strictly feasible, *i.e.*, there exists an $\hat{\boldsymbol{x}}$ such that

$$\hat{x}_1\boldsymbol{F}_1 + \ldots + \hat{x}_d\boldsymbol{F}_d + \boldsymbol{G} \prec \boldsymbol{0}.$$

References

1. Beck, A.: Introduction to Nonlinear Optimization: Theory, Algorithms, and Applications with MATLAB. Soc. Indust. Appl. Math., Philadelphia (2014). URL http://epubs.siam.org/doi/abs/10.1137/1.9781611973655
2. Boyd, S., Vandenberghe, L.: Convex optimization. Cambridge, UK (2004). URL http://www.stanford.edu/~boyd/cvxbook.html