Lagrangian duality

VI

This chapter collects some basic results on Lagrangian duality, in particular as it applies to convex programs with a zero duality gap.

6.1 The relaxation theorem

Given the problem to find

$$f^* := \underset{\boldsymbol{x}}{\text{infimum}} \;\; f(\boldsymbol{x}), \tag{6.1a}$$

$$\text{subject to} \;\; \boldsymbol{x} \in S, \tag{6.1b}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a given function and $S \subseteq \mathbb{R}^n$, we define a *relaxation* to (6.1) to be a problem of the following form: find

$$f_R^* := \underset{\boldsymbol{x}}{\text{infimum}} \;\; f_R(\boldsymbol{x}), \tag{6.2a}$$

$$\text{subject to} \;\; \boldsymbol{x} \in S_R, \tag{6.2b}$$

where $f_R : \mathbb{R}^n \to \mathbb{R}$ is a function with the property that $f_R \leq f$ on S, and where $S_R \supseteq S$. For this pair of problems, we have the following basic result.

Theorem 6.1 (Relaxation Theorem) (a) [relaxation] $f_R^* \leq f^*$.

(b) [infeasibility] *If* (6.2) *is infeasible, then so is* (6.1).

(c) [optimal relaxation] *If the problem* (6.2) *has an optimal solution,* $\boldsymbol{x}_R^*$*, for which it holds that*

$$\boldsymbol{x}_R^* \in S \qquad \text{and} \qquad f_R(\boldsymbol{x}_R^*) = f(\boldsymbol{x}_R^*), \tag{6.3}$$

then $\boldsymbol{x}_R^*$ *is an optimal solution to* (6.1) *as well.*

Proof. The result in (a) is obvious, as every solution feasible in (6.1) is both feasible in (6.2) and has a lower objective value in the latter problem.

The result in (b) follows for similar reasons.

For the result in (c), we note that

$$f(\boldsymbol{x}_R^*) = f_R(\boldsymbol{x}_R^*) \leq f_R(\boldsymbol{x}) \leq f(\boldsymbol{x}), \qquad \boldsymbol{x} \in S,$$

from which the result follows. ■

This basic result will be utilized both in this chapter and later on to motivate why Lagrangian relaxation, objective function linearization and penalization constitute relaxations, and to derive optimality conditions and algorithms based on them.

6.2 Lagrangian duality

In this section we formulate the Lagrangian dual problem and establish its convexity. The Weak Duality Theorem is also established, and we introduce the terms "Lagrangian relaxation," "Lagrange multiplier," and "duality gap."

6.2.1 Lagrangian relaxation and the dual problem

Consider the optimization problem to find

$$\begin{aligned} f^* := \operatorname*{infimum}_{\boldsymbol{x}} \ & f(\boldsymbol{x}), \\ \text{subject to} \quad & \boldsymbol{x} \in X, \\ & g_i(\boldsymbol{x}) \leq 0, \qquad i = 1, \dots, m, \end{aligned} \tag{6.4}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ and $g_i : \mathbb{R}^n \to \mathbb{R}$ $(i = 1, 2, \dots, m)$ are given functions, and $X \subseteq \mathbb{R}^n$.

For this problem, we assume that

$$-\infty < f^* < \infty, \tag{6.5}$$

that is, that f is bounded from below on the feasible set and the problem has at least one feasible solution.

Definition 6.2 (Lagrange function, relaxation, multiplier) (a) *For an arbitrary vector* $\boldsymbol{\mu} \in \mathbb{R}^m$*, the* Lagrange function *is*

$$L(\boldsymbol{x}, \boldsymbol{\mu}) := f(\boldsymbol{x}) + \sum_{i=1}^{m} \mu_i g_i(\boldsymbol{x}) = f(\boldsymbol{x}) + \boldsymbol{\mu}^{\mathrm{T}} \boldsymbol{g}(\boldsymbol{x}). \tag{6.6}$$

(b) *Consider the problem to*

$$\begin{aligned} &\text{minimize } L(\boldsymbol{x}, \boldsymbol{\mu}), \\ &\text{subject to } x \in X. \end{aligned} \tag{6.7}$$

Whenever $\boldsymbol{\mu}$ *is non-negative, the problem* (6.7) *is referred to as a* Lagrangian relaxation.

(c) *We call the vector* $\boldsymbol{\mu}^* \in \mathbb{R}^m$ a Lagrange multiplier vector *if it is non-negative and if* $f^* = \inf_{\boldsymbol{x} \in X} L(\boldsymbol{x}, \boldsymbol{\mu}^*)$ *holds.* ■

Note that the Lagrangian relaxation (6.7) is a relaxation, in terms of Section 6.1.

Theorem 6.3 (Lagrange multipliers and global optima) *Let* $\boldsymbol{\mu}^*$ *be a Lagrange multiplier vector. Then,* $\boldsymbol{x}^*$ *is an optimal solution to* (6.4) *if and only if* $\boldsymbol{x}^*$ *is feasible in* (6.4) *and*

$$\boldsymbol{x}^* \in \arg\min_{\boldsymbol{x} \in X} L(\boldsymbol{x}, \boldsymbol{\mu}^*), \quad \textit{and} \quad \mu_i^* g_i(\boldsymbol{x}^*) = 0, \ i = 1, \dots, m. \tag{6.8}$$

Proof. If $\boldsymbol{x}^*$ is an optimal solution to (6.4), then it is in particular feasible, and

$$f^* = f(\boldsymbol{x}^*) \geq L(\boldsymbol{x}^*, \boldsymbol{\mu}^*) \geq \operatorname*{infimum}_{\boldsymbol{x} \in X} L(\boldsymbol{x}, \boldsymbol{\mu}^*),$$

where the first inequality stems from the feasibility of $\boldsymbol{x}^*$ and the definition of a Lagrange multiplier vector. The second part of that definition implies that $f^* = \inf_{\boldsymbol{x} \in X} L(\boldsymbol{x}, \boldsymbol{\mu}^*)$, so that equality holds throughout in the above line of inequalities. Hence, (6.8) follows.

Conversely, if $\boldsymbol{x}^*$ is feasible and (6.8) holds, then by the use of the definition of a Lagrange multiplier vector,

$$f(\boldsymbol{x}^*) = L(\boldsymbol{x}^*, \boldsymbol{\mu}^*) = \operatorname*{minimum}_{\boldsymbol{x} \in X} L(\boldsymbol{x}, \boldsymbol{\mu}^*) = f^*,$$

so $\boldsymbol{x}^*$ is a global optimum. ■

Let

$$q(\boldsymbol{\mu}) := \operatorname*{infimum}_{\boldsymbol{x} \in X} L(\boldsymbol{x}, \boldsymbol{\mu}) \tag{6.9}$$

be the *Lagrangian dual function*, defined by the infimum value of the Lagrange function over X; the *Lagrangian dual problem* is to

$$\begin{aligned} &\operatorname*{maximize}_{\boldsymbol{\mu}} \ q(\boldsymbol{\mu}), \\ &\text{subject to } \boldsymbol{\mu} \geq \mathbf{0}^m. \end{aligned} \tag{6.10}$$

For some $\boldsymbol{\mu}$, $q(\boldsymbol{\mu}) = -\infty$ is possible; if it is true for all $\boldsymbol{\mu} \geq \mathbf{0}^m$, then

$$q^* := \operatorname*{supremum}_{\boldsymbol{\mu} \geq \mathbf{0}^m} q(\boldsymbol{\mu})$$

equals $-\infty$. (We can then say that the dual problem is infeasible.)

The *effective domain* of q is

$$D_q := \{ \boldsymbol{\mu} \in \mathbb{R}^m \mid q(\boldsymbol{\mu}) > -\infty \}.$$

Theorem 6.4 (convex dual problem) *The effective domain D_q of q is convex, and q is concave on D_q.*

Proof. Let $\boldsymbol{x} \in \mathbb{R}^n$, $\boldsymbol{\mu}, \bar{\boldsymbol{\mu}} \in \mathbb{R}^m$, and $\alpha \in [0, 1]$. We have that

$$L(\boldsymbol{x}, \alpha\boldsymbol{\mu} + (1-\alpha)\bar{\boldsymbol{\mu}}) = \alpha L(\boldsymbol{x}, \boldsymbol{\mu}) + (1-\alpha)L(\boldsymbol{x}, \bar{\boldsymbol{\mu}}).$$

Take the infimum over $\boldsymbol{x} \in X$ on both sides; then,

$$\begin{aligned}
\inf_{\boldsymbol{x} \in X} L(\boldsymbol{x}, \alpha\boldsymbol{\mu} + (1-\alpha)\bar{\boldsymbol{\mu}}) &= \inf_{\boldsymbol{x} \in X} \{\alpha L(\boldsymbol{x}, \boldsymbol{\mu}) + (1-\alpha)L(\boldsymbol{x}, \bar{\boldsymbol{\mu}})\} \\
&\geq \inf_{\boldsymbol{x} \in X} \alpha L(\boldsymbol{x}, \boldsymbol{\mu}) + \inf_{\boldsymbol{x} \in X} (1-\alpha)L(\boldsymbol{x}, \bar{\boldsymbol{\mu}}) \\
&= \alpha \inf_{\boldsymbol{x} \in X} L(\boldsymbol{x}, \boldsymbol{\mu}) + (1-\alpha) \inf_{\boldsymbol{x} \in X} L(\boldsymbol{x}, \bar{\boldsymbol{\mu}}),
\end{aligned}$$

since $\alpha \in [0, 1]$, and the sum of infimum values may be smaller than the infimum of the sum, since in the former case we have the possibility to choose different optimal solutions in the two problems. Hence,

$$q(\alpha\boldsymbol{\mu} + (1-\alpha)\bar{\boldsymbol{\mu}}) \geq \alpha q(\boldsymbol{\mu}) + (1-\alpha)q(\bar{\boldsymbol{\mu}})$$

holds. This inequality has two implications: if $\boldsymbol{\mu}$ and $\bar{\boldsymbol{\mu}}$ lie in D_q, then so does $\alpha\boldsymbol{\mu} + (1-\alpha)\bar{\boldsymbol{\mu}}$, so D_q is convex; also, q is concave on D_q. ∎

That the Lagrangian dual problem always is convex (we indeed maximize a concave function!) is good news, because it means that it can be solved efficiently. What remains is to show how a Lagrangian dual optimal solution can be used to generate a primal optimal solution.

Next, we establish that every feasible point in the Lagrangian *dual* problem always underestimates the objective function value of every feasible point in the *primal* problem; hence, also their optimal values have this relationship.

Theorem 6.5 (Weak Duality Theorem) (a) *Let $\boldsymbol{x}$ and $\boldsymbol{\mu}$ be feasible in the problems* (6.4) *and* (6.10), *respectively. Then,*

$$q(\boldsymbol{\mu}) \leq f(\boldsymbol{x}).$$

In particular,

$$q^* \leq f^*.$$

(b) *If* $q(\boldsymbol{\mu}) = f(\boldsymbol{x})$, *then the pair* $(\boldsymbol{x}, \boldsymbol{\mu})$ *is optimal in its respective problem.*

Proof. For all $\boldsymbol{\mu} \geq \mathbf{0}^m$ and $\boldsymbol{x} \in X$ with $\boldsymbol{g}(\boldsymbol{x}) \leq \mathbf{0}^m$,

$$q(\boldsymbol{\mu}) = \operatorname*{infimum}_{\boldsymbol{z} \in X} L(\boldsymbol{z}, \boldsymbol{\mu}) \leq f(\boldsymbol{x}) + \boldsymbol{\mu}^{\mathrm{T}} \boldsymbol{g}(\boldsymbol{x}) \leq f(\boldsymbol{x}),$$

so

$$q^* = \operatorname*{supremum}_{\boldsymbol{\mu} \geq \mathbf{0}^m} q(\boldsymbol{\mu}) \leq \operatorname*{infimum}_{\boldsymbol{x} \in X : \boldsymbol{g}(\boldsymbol{x}) \leq \mathbf{0}^m} f(\boldsymbol{x}) = f^*.$$

The result follows. ∎

Weak duality is also a consequence of the Relaxation Theorem: For any $\boldsymbol{\mu} \geq \mathbf{0}^m$, let

$$S := X \cap \{\, \boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{g}(\boldsymbol{x}) \leq \mathbf{0}^m \,\}, \tag{6.11a}$$

$$S_R := X, \tag{6.11b}$$

$$f_R := L(\boldsymbol{\mu}, \cdot). \tag{6.11c}$$

Then, the weak duality statement is the result in Theorem 6.1(a).

If our initial feasibility assumption (6.5) is false, then what does weak duality imply? Suppose that $f^* = -\infty$. Then, weak duality implies that $q(\boldsymbol{\mu}) = -\infty$ for all $\boldsymbol{\mu} \geq \mathbf{0}^m$, that is, the dual problem is infeasible. Suppose then that $X \neq \emptyset$ but that $X \cap \{\, \boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{g}(\boldsymbol{x}) \leq \mathbf{0}^m \,\}$ is empty. Then, $f^* = \infty$, by convention. The dual function satisfies $q(\boldsymbol{\mu}) < \infty$ for all $\boldsymbol{\mu} \geq \mathbf{0}^m$, but it is possible that $q^* = -\infty$, $-\infty < q^* < \infty$, or $q^* = \infty$ (see [Ber99, Figure 5.1.8]). For linear programs, $-\infty < q^* < \infty$ implies $-\infty < f^* < \infty$; see below.

If $q^* = f^*$, then we say that the *duality gap* (as given by $\Gamma := f^* - q^*$) is zero, or that *there is no duality gap*. If there exists a Lagrange multiplier vector, then by the weak duality theorem, this implies that there is no duality gap. The converse is not true in general: there may be cases where no Lagrange multipliers exist even when there is no duality gap; in that case though, the Lagrangian dual problem cannot have an optimal solution, as implied by the following result.

Proposition 6.6 (duality gap and the existence of Lagrange multipliers) (a) *If there is no duality gap, then the set of Lagrange multiplier vectors equals the set of optimal dual solutions (which however may be empty).*

(b) *If there is a duality gap, then there are no Lagrange multipliers.*

Proof. By definition, a vector $\boldsymbol{\mu}^* \geq \mathbf{0}^m$ is a Lagrange multiplier vector if and only if $f^* = q(\boldsymbol{\mu}^*) \leq q^*$, the equality following from the definition of $q(\boldsymbol{\mu}^*)$ and the inequality from the definition of q^* as the supremum of $q(\boldsymbol{\mu})$ over $\mathbb{R}^m_+$. By weak duality, this relation holds if and only if there is no duality gap and $\boldsymbol{\mu}^*$ is an optimal dual solution. ∎

Before moving on, we remark on the *statement* of the problem (6.4). There are several ways in which the original set of constraints of the problem can be placed either within the definition of the *ground set* X (which is kept intact), or within the explicit constraints defined by the functions g_i (which are Lagrangian relaxed). How to distinguish between the two, that is, how to decide whether a constraint should be kept or be Lagrangian relaxed, depends on several factors. For example, keeping more constraints within X may result in a smaller duality gap, and fewer multipliers also result in a simpler Lagrangian dual problem. On the other hand, the Lagrangian subproblem defining the dual function simultaneously becomes more complex and difficult to solve. There are no immediate rules to follow, but experimentation and experience.

6.2.2 Global optimality conditions

The following result characterizes every optimal primal and dual solution. It is however applicable only in the presence of Lagrange multipliers; in other words, the below system (6.12) is consistent if and only if there exists a Lagrange multiplier vector and there is no duality gap.

Theorem 6.7 (global optimality conditions in the absence of a duality gap) *The vector $(\boldsymbol{x}^*, \boldsymbol{\mu}^*)$ is a pair of primal optimal solution and Lagrange multiplier vector if and only if*

$$\boldsymbol{\mu}^* \geq \mathbf{0}^m, \qquad \textit{(Dual feasibility)} \tag{6.12a}$$

$$\boldsymbol{x}^* \in \arg\min_{\boldsymbol{x} \in X} L(\boldsymbol{x}, \boldsymbol{\mu}^*), \qquad \textit{(Lagrangian optimality)} \tag{6.12b}$$

$$\boldsymbol{x}^* \in X, \quad \boldsymbol{g}(\boldsymbol{x}^*) \leq \mathbf{0}^m, \qquad \textit{(Primal feasibility)} \tag{6.12c}$$

$$\mu_i^* g_i(\boldsymbol{x}^*) = 0, \quad i = 1, \dots, m. \qquad \textit{(Complementary slackness)} \tag{6.12d}$$

Proof. Suppose that the pair $(\boldsymbol{x}^* \boldsymbol{\mu}^*)$ satisfies (6.12). Then, from (6.12a) we have that the Lagrangian problem to minimize $L(\boldsymbol{x}, \boldsymbol{\mu}^*)$ over $\boldsymbol{x} \in X$ is a (Lagrangian) relaxation of (6.4). Moreover, according to (6.12b) $\boldsymbol{x}^*$ solves this problem, (6.12c) shows that $\boldsymbol{x}^*$ is feasible in (6.4), and (6.12d) implies that $L(\boldsymbol{x}^*, \boldsymbol{\mu}^*) = f(\boldsymbol{x}^*)$. The Relaxation Theorem 6.1

then yields that $\boldsymbol{x}^*$ is optimal in (6.4), which in turn implies that $\boldsymbol{\mu}^*$ is a Lagrange multiplier vector.

Conversely, if $(\boldsymbol{x}^*, \boldsymbol{\mu}^*)$ is a pair of optimal primal solution and Lagrange multiplier vector, then they are primal and dual feasible, respectively. The relations (6.12b) and (6.12d) follow from Theorem 6.3. ∎

Theorem 6.8 (global optimality and saddle points) *The vector* $(\boldsymbol{x}^*, \boldsymbol{\mu}^*)$ *is a pair of optimal primal solution and Lagrange multiplier vector if and only if* $\boldsymbol{x}^* \in X$, $\boldsymbol{\mu}^* \geq \mathbf{0}^m$, *and* $(\boldsymbol{x}^*, \boldsymbol{\mu}^*)$ *is a* saddle point *of the Lagrangian function on* $X \times \mathbb{R}^m_+$, *that is,*

$$L(\boldsymbol{x}^*, \boldsymbol{\mu}) \leq L(\boldsymbol{x}^*, \boldsymbol{\mu}^*) \leq L(\boldsymbol{x}, \boldsymbol{\mu}^*), \qquad (\boldsymbol{x}, \boldsymbol{\mu}) \in X \times \mathbb{R}^m_+, \tag{6.13}$$

holds.

Proof. We establish that (6.12) and (6.13) are equivalent; Theorem 6.7 then gives the result. The first inequality in (6.13) is equivalent to

$$-\boldsymbol{g}(\boldsymbol{x}^*)^{\mathrm{T}}(\boldsymbol{\mu} - \boldsymbol{\mu}^*) \geq 0, \qquad \boldsymbol{\mu} \in \mathbb{R}^m_+, \tag{6.14}$$

for the given pair $(\boldsymbol{x}^*, \boldsymbol{\mu}^*) \in X \times \mathbb{R}^m_+$. This variational inequality is equivalent to stating that[1]

$$\mathbf{0}^m \geq \boldsymbol{g}(\boldsymbol{x}^*) \perp \boldsymbol{\mu}^* \geq \mathbf{0}^m, \tag{6.15}$$

where $\perp$ denotes orthogonality: that is, for any vectors $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n$, $\boldsymbol{a} \perp \boldsymbol{b}$ means that $\boldsymbol{a}^{\mathrm{T}}\boldsymbol{b} = 0$. Because of the sign restrictions posed on $\boldsymbol{\mu}$ and $\boldsymbol{g}$, that is, the vectors $\boldsymbol{a}$ and $\boldsymbol{b}$, the relation $\boldsymbol{a} \perp \boldsymbol{b}$ actually means that not only does it hold that $\boldsymbol{a}^{\mathrm{T}}\boldsymbol{b} = 0$ but in fact $a_i b_i = 0$ must hold for all $i = 1, \ldots, n$. This complementarity system is, for the given $\boldsymbol{\mu}^* \in \mathbb{R}^m_+$, the same as (6.12a), (6.12c) and (6.12d). The second inequality in (6.13) is equivalent to (6.12b). ∎

The above two theorems also imply that the set of primal–dual optimal solutions $(\boldsymbol{x}^*, \boldsymbol{\mu}^*)$ is a Cartesian product set, $X^* \times U^*$. For example,

[1] We establish the equivalence between (6.14) and (6.15) as follows. (The proof extends that for line search problems in unconstrained optimization in a footnote in Section 11.3.1.)

First, suppose that (6.15) is fulfilled. Then, $-\boldsymbol{g}(\boldsymbol{x}^*)^{\mathrm{T}}(\boldsymbol{\mu} - \boldsymbol{\mu}^*) = -\boldsymbol{g}(\boldsymbol{x}^*)^{\mathrm{T}}\boldsymbol{\mu} \geq 0$, for all $\boldsymbol{\mu} \geq \mathbf{0}^m$, that is, (6.14) is fulfilled. Conversely, suppose that (6.14) is fulfilled. Setting $\boldsymbol{\mu} = \mathbf{0}^m$ yields that $\boldsymbol{g}(\boldsymbol{x}^*)^{\mathrm{T}}\boldsymbol{\mu}^* \geq 0$. On the other hand, the choice $\boldsymbol{\mu} = 2\boldsymbol{\mu}^*$ yields that $-\boldsymbol{g}(\boldsymbol{x}^*)^{\mathrm{T}}\boldsymbol{\mu}^* \geq 0$. Hence, $\boldsymbol{g}(\boldsymbol{x}^*)^{\mathrm{T}}\boldsymbol{\mu}^* = 0$ holds. Last, let $\boldsymbol{\mu} = \boldsymbol{\mu}^* + \boldsymbol{e}_i$, where $\boldsymbol{e}_i$ is the i^{th} unit vector in $\mathbb{R}^m$. Then, $-\boldsymbol{g}(\boldsymbol{x}^*)^{\mathrm{T}}(\boldsymbol{\mu} - \boldsymbol{\mu}^*) = -g_i(\boldsymbol{x}^*) \geq 0$. Since this is true for all $i \in \{1, 2, \ldots, m\}$ we have obtained that $-\boldsymbol{g}(\boldsymbol{x}^*) \geq \mathbf{0}^m$, that is, $\boldsymbol{g}(\boldsymbol{x}^*) \leq \mathbf{0}^m$. We are done.

given any optimal dual solution $\boldsymbol{\mu}^* \in U^*$, every optimal primal solution $\boldsymbol{x}^* \in X^*$ satisfies (6.12). Hence, we can write, for an *arbitrary* dual vector $\boldsymbol{\mu}^* \in U^*$,

$$\begin{aligned} X^* &= \{\, \boldsymbol{x}^* \in \mathbb{R}^n \mid \boldsymbol{x}^* \text{ satisfies (6.12) for } \boldsymbol{\mu} = \boldsymbol{\mu}^* \,\} \\ &= \left\{ \boldsymbol{x}^* \in \arg\min_{\boldsymbol{x} \in X} L(\boldsymbol{x}, \boldsymbol{\mu}^*) \;\middle|\; \boldsymbol{g}(\boldsymbol{x}^*) \leq \mathbf{0}^m;\ (\boldsymbol{\mu}^*)^{\mathrm{T}} \boldsymbol{g}(\boldsymbol{x}^*) = 0 \right\}. \end{aligned}$$

We note that structurally similar results to the above two theorems which are valid for the general problem (6.4) with any size of the duality gap can be found in [LaP05].[2]

We finally note a *practical* connection between the KKT system (5.9) and the above system (6.12). The practical use of the KKT system is normally to investigate whether a primal vector $\boldsymbol{x}$—obtained perhaps from a solver for our problem—is a candidate for a locally optimal solution; in other words, we have access to $\boldsymbol{x}$ and generate a vector $\boldsymbol{\mu}$ of Lagrange multipliers in the investigation of the KKT system (5.9). In contrast, the system (6.12) is normally investigated in the reverse order; we formulate and solve the Lagrangian dual problem, thereby obtaining an optimal dual vector $\boldsymbol{\mu}$. Starting from that vector, we investigate the global optimality conditions stated in (6.12) to obtain, if possible, an optimal primal vector $\boldsymbol{x}$. In the section to follow, we show when this is possible, and provide strong connections between the systems (5.9) and (6.12) in the convex and differentiable case.

6.2.3 Strong duality for convex programs

So far the results have been rather non-technical to achieve: the convexity of the Lagrangian dual problem comes with very few assumptions on the original, primal problem, and the characterization of the primal–dual set of optimal solutions is simple and also quite easily established. In order to establish *strong duality*, that is, to establish sufficient conditions under which there is no duality gap, however, takes much more. In particular, as is the case with the KKT conditions we need regularity conditions (that is, constraint qualifications), and we also need to utilize separation theorems such as Theorem 4.28. Most importantly, however, is that strong duality is deeply associated with the convexity of the original problem, and it is in particular under convexity that the primal and

[2] The system (6.12) is there appended with two relaxation parameters which measure, respectively, the near-optimality of $\boldsymbol{x}^*$ in the Lagrangian subproblem [that is, the ε-optimality in (6.12b)], and the violation of the complementarity conditions (6.12d). The saddle point condition (6.13) is similarly perturbed, and at an optimal solution, the sum of these two parameters equals the duality gap.

dual optimal solutions are linked through the global optimality conditions provided in the previous section. We begin by concentrating on the inequality constrained case, proving this result in detail. We will also specialize the result to quadratic and linear optimization problems.

Consider the inequality constrained *convex* program (6.4), where $f : \mathbb{R}^n \to \mathbb{R}$ and g_i $(i = 1, \dots, m)$ are convex functions and $X \subseteq \mathbb{R}^n$ is a convex set. For this problem, we introduce the following regularity condition, due to Slater (cf. Definition 5.38):

$$\exists \boldsymbol{x} \in X \text{ with } \boldsymbol{g}(\boldsymbol{x}) < \boldsymbol{0}^m. \tag{6.16}$$

Theorem 6.9 (Strong Duality, inequality constrained convex programs) *Suppose that the feasibility condition* (6.5) *and Slater's constraint qualification* (6.16) *hold for the convex problem* (6.4).

(a) *There is no duality gap and there exists at least one Lagrange multiplier vector* $\boldsymbol{\mu}^*$*. Moreover, the set of Lagrange multipliers is bounded and convex.*

(b) *If the infimum in* (6.4) *is attained at some* $\boldsymbol{x}^*$*, then the pair* $(\boldsymbol{x}^*, \boldsymbol{\mu}^*)$ *satisfies the global optimality conditions* (6.12).

(c) *If further* f *and* $\boldsymbol{g}$ *are differentiable at* $\boldsymbol{x}^*$*, then the condition* (6.12b) *can equivalently be written as the variational inequality*

$$\nabla_{\boldsymbol{x}} L(\boldsymbol{x}^*, \boldsymbol{\mu}^*)^{\mathrm{T}}(\boldsymbol{x} - \boldsymbol{x}^*) \geq 0, \qquad \boldsymbol{x} \in X. \tag{6.17}$$

If, in addition, X *is open (such as is the case when* $X = \mathbb{R}^n$*), then this reduces to the condition that*

$$\nabla_{\boldsymbol{x}} L(\boldsymbol{x}^*, \boldsymbol{\mu}^*) = \nabla f(\boldsymbol{x}^*) + \sum_{i=1}^{m} \mu_i^* \nabla g_i(\boldsymbol{x}^*) = \boldsymbol{0}^n, \tag{6.18}$$

and the global optimality conditions (6.12) *reduce to the Karush–Kuhn–Tucker conditions stated in Theorem 5.25.*

Proof. (a) We begin by establishing the existence of a Lagrange multiplier vector (and the presence of a zero duality gap).[3]

First, we consider the following subset of $\mathbb{R}^{m+1}$:

$$A := \{(z_1, \dots, z_m, w)^{\mathrm{T}} \mid \exists \boldsymbol{x} \in X \text{ with } g_i(\boldsymbol{x}) \leq z_i,\ i = 1, \dots, m;\ f(\boldsymbol{x}) \leq w\}.$$

It is elementary to show that A is convex.

Next, we observe that $((\boldsymbol{0}^m)^{\mathrm{T}}, f^*)^{\mathrm{T}}$ is not an interior point of A; otherwise, for some $\varepsilon > 0$ the point $((\boldsymbol{0}^m)^{\mathrm{T}}, f^* - \varepsilon)^{\mathrm{T}} \in A$ holds, which

[3] This result is [Ber99, Proposition 5.3.1], whose proof we also utilize.

would contradict the definition of f^*. Therefore, by the (possibly non-proper) separation result in Theorem 4.28, we can find a hyperplane passing through $((\mathbf{0}^m)^{\mathrm{T}}, f^*)^{\mathrm{T}}$ such that A lies in one of the two corresponding half-spaces. In particular, there then exists a vector $(\boldsymbol{\mu}^{\mathrm{T}}, \beta)^{\mathrm{T}} \neq ((\mathbf{0}^m)^{\mathrm{T}}, 0)^{\mathrm{T}}$ such that

$$\beta f^* \leq \beta w + \boldsymbol{\mu}^{\mathrm{T}} \boldsymbol{z}, \qquad (\boldsymbol{z}^{\mathrm{T}}, w)^{\mathrm{T}} \in A. \tag{6.19}$$

This implies that

$$\beta \geq 0; \qquad \boldsymbol{\mu} \geq \mathbf{0}^m, \tag{6.20}$$

since for each $(\boldsymbol{z}^{\mathrm{T}}, w)^{\mathrm{T}} \in A$ $(\boldsymbol{z}^{\mathrm{T}}, w+\gamma)^{\mathrm{T}} \in A$ and $(z_1, \dots, z_{i-1}, z_i + \gamma, z_{i+1}, \dots, z_m, w)^{\mathrm{T}} \in A$ for all $\gamma > 0$ and $i = 1, \dots, m$.

We claim that $\beta > 0$ in fact holds. Indeed, if it was not the case, then $\beta = 0$ and (6.19) then implies that $\boldsymbol{\mu}^{\mathrm{T}} \boldsymbol{z} \geq 0$ for every pair $(\boldsymbol{z}^{\mathrm{T}}, w)^{\mathrm{T}} \in A$. But since $(\boldsymbol{g}(\bar{\boldsymbol{x}})^{\mathrm{T}}, f(\bar{\boldsymbol{x}}))^{\mathrm{T}} \in A$ [where $\bar{\boldsymbol{x}}$ is such that it satisfies the Slater condition (6.16)], we would obtain that $0 \leq \sum_{i=1}^m \mu_i g_i(\bar{\boldsymbol{x}})$ which in view of $\boldsymbol{\mu} \geq \mathbf{0}^m$ [cf. (6.20)] and the assumption that $\bar{\boldsymbol{x}}$ satisfies the Slater condition (6.16) implies that $\boldsymbol{\mu} = \mathbf{0}^m$. This means, however, that $(\boldsymbol{\mu}^{\mathrm{T}}, \beta)^{\mathrm{T}} = ((\mathbf{0}^m)^{\mathrm{T}}, 0)^{\mathrm{T}}$—a contradiction. We may therefore claim that $\beta > 0$. We further, with no loss of generality, assume that $\beta = 1$.

Thus, since $(\boldsymbol{g}(\boldsymbol{x})^{\mathrm{T}}, f(\boldsymbol{x}))^{\mathrm{T}} \in A$ for every $\boldsymbol{x} \in X$, (6.19) yields that

$$f^* \leq f(\boldsymbol{x}) + \boldsymbol{\mu}^{\mathrm{T}} \boldsymbol{g}(\boldsymbol{x}), \qquad \boldsymbol{x} \in X.$$

Taking the infimum over $\boldsymbol{x} \in X$ and using the fact that $\boldsymbol{\mu} \geq \mathbf{0}^m$ we obtain

$$f^* \leq \operatorname*{infimum}_{\boldsymbol{x} \in X} \{ f(\boldsymbol{x}) + \boldsymbol{\mu}^{\mathrm{T}} \boldsymbol{g}(\boldsymbol{x}) \} = q(\boldsymbol{\mu}) \leq \operatorname*{supremum}_{\boldsymbol{\mu} \geq \mathbf{0}^m} q(\boldsymbol{\mu}) = q^*.$$

Using the Weak Duality Theorem 6.5 it follows that $\boldsymbol{\mu}$ is a Lagrange multiplier vector, and there is no duality gap. This part of the proof is now done.

Take any vector $\bar{\boldsymbol{x}} \in X$ satisfying (6.16) and a Lagrange multiplier vector $\boldsymbol{\mu}^*$. By the definition of a Lagrange multiplier vector, $f^* \leq L(\bar{\boldsymbol{x}}, \boldsymbol{\mu}^*)$ holds, which implies that

$$\sum_{i=1}^m \mu_i^* \leq \frac{[f(\bar{\boldsymbol{x}}) - f^*]}{\min_{i=1,\dots,m} \{ -g_i(\bar{\boldsymbol{x}}) \}}.$$

Since $\boldsymbol{\mu}^* \geq \mathbf{0}^m$, boundedness follows. As by Proposition 6.6(a) the set of Lagrange multipliers is the set of optimal solutions to the dual problem (6.10), convexity follows from the identification of the dual solution set with the set of vectors $\boldsymbol{\mu} \in \mathbb{R}^m_+$ for which

$$q(\boldsymbol{\mu}) \geq q^*$$

holds. This is the upper level set for q at the level q^*; this set is convex, by the concavity of q (cf. Theorem 6.4 and Proposition 3.44).

(b) The result follows from Theorem 6.7.

(c) The first part follows from Theorem 4.23, as the Lagrangian function $L(\cdot, \boldsymbol{\mu}^*)$ is convex. The second part follows by identification. ■

Consider next the extension of the inequality constrained convex program (6.4) in which we seek to find

$$
\begin{aligned}
f^* := \operatorname*{infimum}_{\boldsymbol{x}} \ & f(\boldsymbol{x}), && && (6.21)\\
\text{subject to} \quad & \boldsymbol{x} \in X, \\
& g_i(\boldsymbol{x}) \leq 0, && i = 1, \dots, m, \\
& \boldsymbol{\varepsilon}_j^{\mathrm{T}} \boldsymbol{x} - d_j = 0, && j = 1, \dots, \ell,
\end{aligned}
$$

under the same conditions as stated following (6.4), and where $\boldsymbol{\varepsilon}_j \in \mathbb{R}^n$, $j = 1, \dots, \ell$. For this problem, we replace the Slater condition (6.16) with the following (cf. [BSS93, Theorem 6.2.4]):

$$
\exists \boldsymbol{x} \in X \text{ with } \boldsymbol{g}(\boldsymbol{x}) < \mathbf{0}^m \text{ and } \mathbf{0}^m \in \operatorname{int} \{\, \boldsymbol{E}\boldsymbol{x} - \boldsymbol{d} \mid \boldsymbol{x} \in X \,\}, \qquad (6.22)
$$

where $\boldsymbol{E} \in \mathbb{R}^{\ell \times n}$ has rows $\boldsymbol{\varepsilon}_j^{\mathrm{T}}$, and $\boldsymbol{d} = (d_j)_{j \in \{1,\dots,\ell\}} \in \mathbb{R}^\ell$.

Note that in the statement (6.22), the "int" can be stricken whenever X is polyhedral, so that the latter part simply states that $\boldsymbol{E}\boldsymbol{x} = \boldsymbol{d}$.

For this problem, the Lagrangian dual problem is to find

$$
\begin{aligned}
q^* := \operatorname*{supremum}_{(\boldsymbol{\mu}, \boldsymbol{\lambda})} \ & q(\boldsymbol{\mu}, \boldsymbol{\lambda}), \qquad (6.23)\\
\text{subject to} \ & \boldsymbol{\mu} \geq \mathbf{0}^m,
\end{aligned}
$$

where

$$
\begin{aligned}
q(\boldsymbol{\mu}, \boldsymbol{\lambda}) := \operatorname*{infimum}_{\boldsymbol{x}} \ & L(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\lambda}) := f(\boldsymbol{x}) + \boldsymbol{\mu}^{\mathrm{T}} \boldsymbol{g}(\boldsymbol{x}) + \boldsymbol{\lambda}^{\mathrm{T}} (\boldsymbol{E}\boldsymbol{x} - \boldsymbol{d}), \\
\text{subject to} \ & \boldsymbol{x} \in X.
\end{aligned}
$$

Theorem 6.10 (Strong Duality, general convex programs) *Suppose that in addition to the feasibility condition* (6.5), *Slater's constraint qualification* (6.22) *holds for the problem* (6.21).

(a) *The duality gap is zero and there exists at least one Lagrange multiplier vector pair* $(\boldsymbol{\mu}^*, \boldsymbol{\lambda}^*)$.

(b) *If the infimum in* (6.21) *is attained at some* $\boldsymbol{x}^*$, *then the triple*

$(\boldsymbol{x}^*, \boldsymbol{\mu}^*, \boldsymbol{\lambda}^*)$ *satisfies the global optimality conditions*

$$
\begin{aligned}
\boldsymbol{\mu}^* &\geq \mathbf{0}^m, && \textit{(Dual feasibility)} && (6.24a)\\
\boldsymbol{x}^* &\in \arg\min_{\boldsymbol{x}\in X} L(\boldsymbol{x}, \boldsymbol{\mu}^*, \boldsymbol{\lambda}^*), && \textit{(Lagrangian optimality)} && (6.24b)\\
\boldsymbol{x}^* \in X,\ \boldsymbol{g}(\boldsymbol{x}^*) &\leq \mathbf{0}^m,\ \boldsymbol{E}\boldsymbol{x}^* = \boldsymbol{d}, && \textit{(Primal feasibility)} && (6.24c)\\
\mu_i^* g_i(\boldsymbol{x}^*) &= 0,\ i = 1, \dots, m. && \textit{(Complementary slackness)} && (6.24d)
\end{aligned}
$$

(c) *If further f and $\boldsymbol{g}$ are differentiable at $\boldsymbol{x}^*$, then the condition* (6.24b) *can equivalently be written as*

$$
\nabla_{\boldsymbol{x}} L(\boldsymbol{x}^*, \boldsymbol{\mu}^*, \boldsymbol{\lambda}^*)^{\mathrm{T}}(\boldsymbol{x} - \boldsymbol{x}^*) \geq 0, \qquad \boldsymbol{x} \in X. \tag{6.25}
$$

If, in addition, X is open (such as is the case when $X = \mathbb{R}^n$), then this reduces to the condition that

$$
\nabla_{\boldsymbol{x}} L(\boldsymbol{x}^*, \boldsymbol{\mu}^*, \boldsymbol{\lambda}^*) = \nabla f(\boldsymbol{x}^*) + \sum_{i=1}^{m} \mu_i^* \nabla g_i(\boldsymbol{x}^*) + \sum_{j=1}^{\ell} \lambda_j^* \boldsymbol{\varepsilon}_j = \mathbf{0}^n, \tag{6.26}
$$

and the global optimality conditions (6.24) *reduce to the Karush–Kuhn–Tucker conditions stated in Theorem 5.33.*

Proof. The proof is similar to that of Theorem 6.9. ■

We finally consider a special case where automatically a regularity condition holds.

Consider the affinely constrained convex program to find

$$
\begin{aligned}
f^* := \operatorname*{infimum}_{\boldsymbol{x}} &\ f(\boldsymbol{x}), && && (6.27)\\
\text{subject to} &\qquad \boldsymbol{x} \in X, &&\\
&\boldsymbol{a}_i^{\mathrm{T}}\boldsymbol{x} - b_i \leq 0, && i = 1, \dots, m,\\
&\boldsymbol{\varepsilon}_j^{\mathrm{T}}\boldsymbol{x} - d_j = 0, && j = 1, \dots, \ell,
\end{aligned}
$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is convex and $X \subseteq \mathbb{R}^n$ is polyhedral.

Theorem 6.11 (Strong Duality, affine constraints) *If the feasibility condition* (6.5) *holds for the problem* (6.27)*, then there is no duality gap and there exists at least one Lagrange multiplier vector.*

Proof. Again, the proof is similar to that of Theorem 6.9, except that no additional regularity conditions are needed.[4] ■

[4]For a detailed proof, see [Ber99, Proposition 5.2.1]. (The special case where f is moreover differentiable is covered in [Ber99, Proposition 3.4.2].)

The existence of a multiplier vector [which by Proposition 6.6 and the absence of a duality gap implies the existence of an optimal solution to the dual problem (6.10)] does not imply the existence of an optimal solution to the primal problem (6.27) without any additional assumptions (take the minimization of $f(x) := 1/x$ over $x \geq 1$ for example). However, when f is either weakly coercive, quadratic or linear, the existence results are stronger; see the primal existence results in Theorems 4.6, 4.7, and 6.12 below, for example.

For convex programs where a Slater CQ holds, the Lagrange multipliers defined in this section, and those that appear in the Karush–Kuhn–Tucker conditions, clearly are identical. Next, we specialize the above to linear and quadratic programs.

6.2.4 Strong duality for linear and quadratic programs

The following result will be established and analyzed in detail in Chapter 10 on linear programming duality (cf. Theorem 10.6), but can in fact also be established similarly to above. (See [BSS93, Theorem 2.7.3] or [Ber99, Proposition 5.2.2], for example.) Its proof will however be relegated to that of Theorem 10.6.

Theorem 6.12 (Strong Duality, linear programs) *Assume, in addition to the conditions of Theorem* 6.11, *that f is linear, so that* (6.27) *is a linear program. Then, the primal and dual problems have optimal solutions and there is no duality gap.* ∎

The above result states a strong duality result for a general linear program. We next develop an explicit Lagrangian dual problem for a linear program.

Let $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, $\boldsymbol{c} \in \mathbb{R}^n$, and $\boldsymbol{b} \in \mathbb{R}^m$; consider the linear program

$$\begin{aligned} \underset{\boldsymbol{x}}{\text{minimize}} \ & \boldsymbol{c}^{\mathrm{T}}\boldsymbol{x}, \\ \text{subject to} \ & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}, \\ & \boldsymbol{x} \geq \boldsymbol{0}^n. \end{aligned} \tag{6.28}$$

If we let $X := \mathbb{R}^n_+$, then the Lagrangian dual problem is to

$$\begin{aligned} \underset{\boldsymbol{\lambda} \in \mathbb{R}^m}{\text{maximize}} \ & \boldsymbol{b}^{\mathrm{T}}\boldsymbol{\lambda}, \\ \text{subject to} \ & \boldsymbol{A}^{\mathrm{T}}\boldsymbol{\lambda} \leq \boldsymbol{c}. \end{aligned} \tag{6.29}$$

The reason why we can write it in this form is that

$$q(\boldsymbol{\lambda}) := \underset{\boldsymbol{x} \geq \boldsymbol{0}^n}{\text{infimum}} \left\{ \boldsymbol{c}^{\mathrm{T}}\boldsymbol{x} + \boldsymbol{\lambda}^{\mathrm{T}}(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}) \right\} = \boldsymbol{b}^{\mathrm{T}}\boldsymbol{\lambda} + \underset{\boldsymbol{x} \geq \boldsymbol{0}^n}{\text{infimum}}\, (\boldsymbol{c} - \boldsymbol{A}^{\mathrm{T}}\boldsymbol{\lambda})^{\mathrm{T}}\boldsymbol{x},$$

so that

$$q(\boldsymbol{\lambda}) = \begin{cases} \boldsymbol{b}^{\mathrm{T}}\boldsymbol{\lambda}, & \text{if } \boldsymbol{A}^{\mathrm{T}}\boldsymbol{\lambda} \leq \boldsymbol{c}, \\ -\infty, & \text{otherwise.} \end{cases}$$

(The infimum is attained at zero if and only if these inequalities are satisfied; otherwise, the inner problem is unbounded below.)

Further, why is it that $\boldsymbol{\lambda}$ here is not restricted in sign? Suppose we were to split the system $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ into an inequality system of the form

$$\begin{aligned} \boldsymbol{A}\boldsymbol{x} &\leq \boldsymbol{b}, \\ -\boldsymbol{A}\boldsymbol{x} &\leq -\boldsymbol{b}. \end{aligned}$$

Let $((\boldsymbol{\mu}^+)^{\mathrm{T}}, (\boldsymbol{\mu}^-)^{\mathrm{T}})^{\mathrm{T}}$ be the corresponding vector of multipliers, and take the Lagrangian dual for this formulation. Then, we would have a Lagrange function of the form

$$(\boldsymbol{x}, \boldsymbol{\mu}^+, \boldsymbol{\mu}^-) \mapsto L(\boldsymbol{x}, \boldsymbol{\mu}^+, \boldsymbol{\mu}^-) := \boldsymbol{c}^{\mathrm{T}}\boldsymbol{x} + (\boldsymbol{\mu}^+ - \boldsymbol{\mu}^-)^{\mathrm{T}}(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}),$$

and since $\boldsymbol{\mu}^+ - \boldsymbol{\mu}^-$ can take on any value in $\mathbb{R}^m$ we can simply replace it with the unrestricted vector $\boldsymbol{\lambda} \in \mathbb{R}^m$. This motivates why the multiplier for an equality constraint never is sign restricted; the same was the case, as we saw in Section 5.6, for the multipliers in the KKT conditions.

As applied to this problem, Theorem 6.12 states that if both the primal or dual problems have feasible solutions, then they both have optimal solutions, satisfying strong duality ($\boldsymbol{c}^{\mathrm{T}}\boldsymbol{x}^* = \boldsymbol{b}^{\mathrm{T}}\boldsymbol{\lambda}^*$). On the other hand, if any of the two problems has an unbounded solution, then the other problem is infeasible.

Consider next the quadratic programming problem to

$$\begin{aligned} &\underset{\boldsymbol{x}}{\text{minimize}} \quad \left\{ \frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{x} + \boldsymbol{c}^{\mathrm{T}}\boldsymbol{x} \right\}, \\ &\text{subject to} \quad \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}, \end{aligned} \tag{6.30}$$

where $\boldsymbol{Q} \in \mathbb{R}^{n \times n}$, $\boldsymbol{c} \in \mathbb{R}^n$, $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, and $\boldsymbol{b} \in \mathbb{R}^m$. We develop an explicit dual problem under the assumption that $\boldsymbol{Q}$ is positive definite.

By Lagrangian relaxing the inequality constraints, we obtain that the inner problem in $\boldsymbol{x}$ is solved by letting

$$\boldsymbol{x} = -\boldsymbol{Q}^{-1}(\boldsymbol{c} + \boldsymbol{A}^{\mathrm{T}}\boldsymbol{\mu}). \tag{6.31}$$

Substituting this expression into the Lagrangian function yields the Lagrangian dual problem to

$$\begin{aligned} &\underset{\boldsymbol{\mu}}{\text{maximize}} \quad \left\{ -\frac{1}{2}\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{Q}^{-1}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\mu} - (\boldsymbol{b} + \boldsymbol{A}\boldsymbol{Q}^{-1}\boldsymbol{c})^{\mathrm{T}}\boldsymbol{\mu} - \frac{1}{2}\boldsymbol{c}^{\mathrm{T}}\boldsymbol{Q}^{-1}\boldsymbol{c} \right\}, \\ &\text{subject to} \quad \boldsymbol{\mu} \geq \boldsymbol{0}^m, \end{aligned} \tag{6.32}$$

Strong duality follows for this convex primal–dual pair of quadratic programs, in much the same way as for linear programs.

Theorem 6.13 (Strong Duality, quadratic programs) *For the primal–dual pair of convex quadratic programs* (6.30), (6.32), *the following holds:*

(a) *If both problems have feasible solutions, then both problems also have optimal solutions, and the primal problem* (6.30) *also has a unique optimal solution, given by* (6.31) *for any optimal Lagrange multiplier vector, and in the two problems the optimal values are equal.*

(b) *If either of the two problems has an unbounded solution, then the other one is infeasible.*

(c) *Suppose that $\boldsymbol{Q}$ is positive semi-definite, and that the feasibility condition* (6.5) *holds. Then, both the problem* (6.30) *and its Lagrangian dual have nonempty, closed and convex sets of optimal solutions, and their optimal values are equal.* ∎

In the result (a) it is important to note that the Lagrangian dual problem (6.32) is not necessarily strictly convex; the matrix $\boldsymbol{AQ}^{-1}\boldsymbol{A}^{\mathrm{T}}$ need not be positive definite, especially so when $\boldsymbol{A}$ does not have full rank. The result (c) extends the strong duality result from linear programming, since $\boldsymbol{Q}$ in (c) can be the zero matrix. In the case of (c) we of course cannot write the Lagrangian dual problem in the form of (6.32) because $\boldsymbol{Q}$ is not necessarily invertible.

6.3 Two illustrative examples

Example 6.14 (an explicit, differentiable dual problem) Consider the problem to

$$
\begin{aligned}
\underset{\boldsymbol{x}}{\operatorname{minimize}} \;\; & f(\boldsymbol{x}) := x_1^2 + x_2^2, \\
\text{subject to} \;\; & x_1 + x_2 \geq 4, \\
& x_j \geq 0, \qquad j = 1, 2.
\end{aligned}
$$

We consider the first constraint to be the complicated one, and hence define $g(\boldsymbol{x}) := -x_1 - x_2 + 4$ and let $X := \{\, (x_1, x_2)^{\mathrm{T}} \mid x_j \geq 0,\ j = 1, 2 \,\}$.

Then, the Lagrangian dual function is

$$
\begin{aligned}
q(\mu) &= \underset{\boldsymbol{x} \in X}{\operatorname{minimum}}\ L(\boldsymbol{x}, \mu) := f(\boldsymbol{x}) - \mu(x_1 + x_2 - 4) \\
&= 4\mu + \underset{\boldsymbol{x} \in X}{\operatorname{minimum}}\ \{x_1^2 + x_2^2 - \mu x_1 - \mu x_2\} \\
&= 4\mu + \underset{x_1 \geq 0}{\operatorname{minimum}}\ \{x_1^2 - \mu x_1\} + \underset{x_2 \geq 0}{\operatorname{minimum}}\ \{x_2^2 - \mu x_2\}, \qquad \mu \geq 0.
\end{aligned}
$$

For a fixed $\mu \geq 0$, the minimum is attained at $x_1(\mu) = \frac{\mu}{2}, x_2(\mu) = \frac{\mu}{2}$.

Substituting this expression into $q(\mu)$, we obtain that $q(\mu) = f(\boldsymbol{x}(\mu)) - \mu(x_1(\mu) + x_2(\mu) - 4) = 4\mu - \frac{\mu^2}{2}$.

Note that q is strictly concave, and it is differentiable everywhere (due to the fact that f, g are differentiable and $\boldsymbol{x}(\mu)$ is unique), by Danskin's Theorem 6.17(d).

We have that $q'(\mu) = 4 - \mu = 0 \iff \mu = 4$. As $\mu = 4 \geq 0$, it is the optimum in the dual problem: $\mu^* = 4; \boldsymbol{x}^* = (x_1(\mu^*), x_2(\mu^*))^{\mathrm{T}} = (2,2)^{\mathrm{T}}$. Also, $f(\boldsymbol{x}^*) = q(\mu^*) = 8$.

This is an example where the dual function is differentiable, and therefore we can utilize Proposition 6.29(c). In this case, the optimum $\boldsymbol{x}^*$ is also unique, so it is automatically given as $\boldsymbol{x}^* = \boldsymbol{x}(\mu)$. ■

Example 6.15 (an implicit, non-differentiable dual problem) Consider the linear programming problem to

$$\begin{aligned} \underset{\boldsymbol{x}}{\text{minimize}} \;\; & f(\boldsymbol{x}) := -x_1 - x_2, \\ \text{subject to} \;\; & 2x_1 + 4x_2 \leq 3, \\ & 0 \leq x_1 \leq 2, \\ & 0 \leq x_2 \leq 1. \end{aligned}$$

The optimal solution is $\boldsymbol{x}^* = (3/2, 0)^{\mathrm{T}}, f(\boldsymbol{x}^*) = -3/2$.

Consider Lagrangian relaxing the first constraint, obtaining

$$\begin{aligned} L(\boldsymbol{x}, \mu) &= -x_1 - x_2 + \mu(2x_1 + 4x_2 - 3); \\ q(\mu) &= -3\mu + \underset{0 \leq x_1 \leq 2}{\text{minimum}} \{(-1 + 2\mu)x_1\} + \underset{0 \leq x_2 \leq 1}{\text{minimum}} \{(-1 + 4\mu)x_2\} \\ &= \begin{cases} -3 + 5\mu, & 0 \leq \mu \leq 1/4, \\ -2 + \ \mu, & 1/4 \leq \mu \leq 1/2, \\ \quad\ - 3\mu, & 1/2 \leq \mu. \end{cases} \end{aligned}$$

Check that $\mu^* = 1/2$, and hence that $q(\mu^*) = -3/2$. For linear programs, we have strong duality, but how do we obtain the optimal primal solution from μ^*? It is clear that q is non-differentiable at μ^*. Let us utilize the characterization given in the system (6.12).

First, at μ^*, it is clear that $X(\mu^*)$ is the set $\{\,(2\alpha, 0)^{\mathrm{T}} \mid 0 \leq \alpha \leq 1\,\}$. Among the subproblem solutions, we next have to find one that is primal feasible as well as complementary.

Primal feasibility means that $2 \cdot 2\alpha + 4 \cdot 0 \leq 3 \iff \alpha \leq 3/4$.

Further, complementarity means that $\mu^* \cdot (2x_1^* + 4x_2^* - 3) = 0 \iff \alpha = 3/4$, since $\mu^* \neq 0$. We conclude that the only primal vector that satisfies the system (6.12) together with the dual optimal solution $\mu^* = 1/2$ is $\boldsymbol{x}^* = (3/2, 0)^{\mathrm{T}}$. ■

In the first example, the Lagrangian dual function is differentiable since $\boldsymbol{x}(\mu)$ is unique. The second one shows that otherwise, there may be kinks in the function q where there are alternative solutions $\boldsymbol{x}(\mu)$; as a result, to obtain a primal optimal solution becomes more complex. The Dantzig–Wolfe algorithm, for example, represents a means by which to automatize the process that we have just shown; the algorithm generates extreme points of $X(\boldsymbol{\mu})$ algorithmically, and constructs the best feasible convex combination thereof, obtaining a primal–dual optimal solution in a finite number of iterations for linear programs.

The above examples motivate a deeper study of the differentiability properties of convex (or, concave) functions in general, and the Lagrangian dual objective function in particular.

6.4 *Differentiability properties of the dual function

We have established that the Lagrangian dual problem (6.10) is a convex one, and further that under some circumstances the primal and dual optimal values are the same. We now turn to study the Lagrangian dual problem in detail, and in particular how it can be solved efficiently. First, we will establish when the dual function q is differentiable. We will see that differentiability holds only in some special cases, in which we can recognize the workings of the *Lagrange multiplier method*; this classic method was illustrated in Example 6.14. Most often, the function q will however be non-differentiable, and then this method will fail. This means that we must devise a more general numerical method which is not based on gradients but rather *subgradients*. This type of algorithm is the topic of the next section; we begin by studying the topic of subgradients of convex functions in general.

6.4.1 Subdifferentiability of convex functions

Throughout this section we suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is a convex function, and study its subdifferentiability properties. We will later on apply our findings to the Lagrangian dual function q, or, rather, its negative $-q$. We first remark that a finite convex function is automatically continuous (cf. Theorem 4.26).

Definition 6.16 (subgradient) *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function. We say that a vector $\boldsymbol{g} \in \mathbb{R}^n$ is a* subgradient *of f at $\boldsymbol{x} \in \mathbb{R}^n$ if*

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \boldsymbol{g}^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{x}), \qquad \boldsymbol{y} \in \mathbb{R}^n. \tag{6.33}$$

The set of such vectors $\boldsymbol{g}$ defines the subdifferential *of f at $\boldsymbol{x}$, and is denoted $\partial f(\boldsymbol{x})$.* ∎

For concave functions, the reverse inequality of course holds; for simplicity we will refer also to such vectors $\boldsymbol{g}$ as subgradients.

Notice the connection to the characterization of a convex function in C^1 in Theorem 3.40(a). The difference between them is that $\boldsymbol{g}$ is not unique at a non-differentiable point. (Just as the gradient has a role in supporting hyperplanes to the graph of a convex function in C^1, the role of a subgradient is the same; at a non-differentiable point there are more then one supporting hyperplane to the graph of f.)

We illustrate this in Figure 6.1.

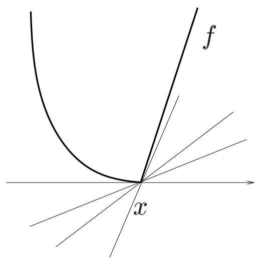

Figure 6.1: Three possible slopes of the convex function f at x.

Notice that a minimum $\boldsymbol{x}^*$ of f over $\mathbb{R}^n$ is characterized by the inclusion $\mathbf{0}^n \in \partial f(\boldsymbol{x}^*)$; recognize, again, the similarity to the C^1 case.

We list some additional basic results for convex functions next. Proofs will not be given here; we refer instead to the convex analysis text by Rockafellar [Roc70].

Proposition 6.17 (properties of a convex function) *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function.*

(a) [boundedness of $\partial f(\boldsymbol{x})$] *For every $\boldsymbol{x} \in \mathbb{R}^n$, $\partial f(\boldsymbol{x})$ is a nonempty, convex, and compact set. If X is bounded then $\cup_{\boldsymbol{x} \in X} \partial f(\boldsymbol{x})$ is bounded.*

(b) [closedness of ∂f] *The subdifferential mapping $\boldsymbol{x} \rightrightarrows \partial f(\boldsymbol{x})$ is* closed*; in other words, if $\{\boldsymbol{x}_k\}$ is a sequence of vectors in $\mathbb{R}^n$ converging to $\boldsymbol{x}$, and $\boldsymbol{g}_k \in \partial f(\boldsymbol{x}_k)$ holds for every k, then the sequence $\{\boldsymbol{g}_k\}$ of subgradients is bounded and every limit point thereof belongs to $\partial f(\boldsymbol{x})$.*

(c) [directional derivative and differentiability] *For every $\boldsymbol{x} \in \mathbb{R}^n$, the directional derivative of f at $\boldsymbol{x}$ in the direction of $\boldsymbol{p} \in \mathbb{R}^n$ satisfies*

$$f'(\boldsymbol{x}; \boldsymbol{p}) = \underset{g \in \partial f(\boldsymbol{x})}{\operatorname{maximum}} \, \boldsymbol{g}^{\mathrm{T}} \boldsymbol{p}. \tag{6.34}$$

In particular, f is differentiable at $\boldsymbol{x}$ with gradient $\nabla f(\boldsymbol{x})$ if and only if it has $\nabla f(\boldsymbol{x})$ as its unique subgradient at $\boldsymbol{x}$; in that case, $f'(\boldsymbol{x};\boldsymbol{p}) = \nabla f(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{p}$.

(d) [Danskin's Theorem—directional derivatives of a convex max function] *Let Z be a compact subset of $\mathbb{R}^m$, and let $\phi : \mathbb{R}^n \times Z \to \mathbb{R}$ be continuous and such that $\phi(\cdot, \boldsymbol{z}) : \mathbb{R}^n \to \mathbb{R}$ is convex for each $\boldsymbol{z} \in Z$. Let the function $f : \mathbb{R}^n \to \mathbb{R}$ be given by*

$$f(\boldsymbol{x}) := \operatorname*{maximum}_{\boldsymbol{z}\in Z} \phi(\boldsymbol{x}, \boldsymbol{z}), \qquad \boldsymbol{x} \in \mathbb{R}^n. \tag{6.35}$$

The function f then is convex on $\mathbb{R}^n$ and has a directional derivative at $\boldsymbol{x}$ in the direction of $\boldsymbol{p}$ equal to

$$f'(\boldsymbol{x};\boldsymbol{p}) = \operatorname*{maximum}_{\boldsymbol{z}\in Z(\boldsymbol{x})} \phi'(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{p}), \tag{6.36}$$

where $\phi'(\boldsymbol{x}, \boldsymbol{z}; \boldsymbol{p})$ is the directional derivative of $\phi(\cdot, \boldsymbol{z})$ at $\boldsymbol{x}$ in the direction of $\boldsymbol{p}$, and $Z(\boldsymbol{x}) := \{\, \boldsymbol{z} \in \mathbb{R}^m \mid \phi(\boldsymbol{x}, \boldsymbol{z}) = f(\boldsymbol{x}) \,\}$.

In particular, if $Z(\boldsymbol{x})$ contains a single point $\bar{\boldsymbol{z}}$ and $\phi(\cdot, \bar{\boldsymbol{z}})$ is differentiable at $\boldsymbol{x}$, then f is differentiable at $\boldsymbol{x}$, and $\nabla f(\boldsymbol{x}) = \nabla_{\boldsymbol{x}}\phi(\boldsymbol{x}, \bar{\boldsymbol{z}})$, where $\nabla_{\boldsymbol{x}}\phi(\boldsymbol{x}, \bar{\boldsymbol{z}})$ is the vector with components $\frac{\partial\phi(\boldsymbol{x},\bar{\boldsymbol{z}})}{\partial x_i}$, $i = 1, \dots, n$.

If further $\phi(\cdot, \boldsymbol{z})$ is differentiable for all $\boldsymbol{z} \in Z$ and $\nabla_{\boldsymbol{x}}\phi(\boldsymbol{x}, \cdot)$ is continuous on Z for each $\boldsymbol{x}$, then

$$\partial f(\boldsymbol{x}) = \operatorname{conv}\{\, \nabla_{\boldsymbol{x}}\phi(\boldsymbol{x}, \boldsymbol{z}) \mid \boldsymbol{z} \in Z(\boldsymbol{x}) \,\}, \qquad \boldsymbol{x} \in \mathbb{R}^n.$$

Proof. (a) This is a special case of [Roc70, Theorem 24.7].

(b) This is [Roc70, Theorem 24.5].

(c) This is [Roc70, Theorem 23.4 and 25.1].

(d) This is [Ber99, Proposition B.25]. ■

Figure 6.2 illustrates the subdifferential of a convex function.

We apply parts of the above results in order to characterize a minimum of a convex function on $\mathbb{R}^n$.

Proposition 6.18 (optimality of a convex function over $\mathbb{R}^n$) *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function. The following three statements are equivalent:*

1. *f is globally minimized at $\boldsymbol{x}^* \in \mathbb{R}^n$;*
2. *$\mathbf{0}^n \in \partial f(\boldsymbol{x}^*)$;*
3. *$f'(\boldsymbol{x}^*;\boldsymbol{p}) \geq 0$ for all $\boldsymbol{p} \in \mathbb{R}^n$.*

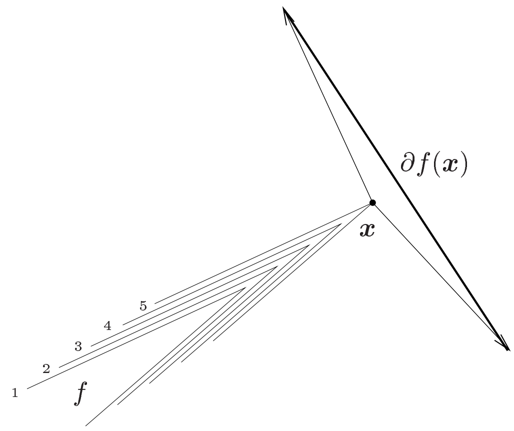

Figure 6.2: The subdifferential of a convex function f at $\boldsymbol{x}$.

Proof. We establish the result thus: $1 \Longrightarrow 2 \Longrightarrow 3 \Longrightarrow 1$.

[$1 \Longrightarrow 2$]: By the statement 1., we have that $f(\boldsymbol{y}) \geq f(\boldsymbol{x}^*)$ for every $\boldsymbol{y} \in \mathbb{R}^n$. This implies that for $\boldsymbol{g} = \mathbf{0}^n$, we satisfy the subgradient inequality (6.33). This establishes the statement 2.

[$2 \Longrightarrow 3$]: We can equivalently write

$$\partial f(\boldsymbol{x}) = \{\, \boldsymbol{g} \in \mathbb{R}^n \mid \boldsymbol{g}^{\mathrm{T}}\boldsymbol{p} \leq f'(\boldsymbol{x};\boldsymbol{p}), \quad \boldsymbol{p} \in \mathbb{R}^n \,\}.$$

With $\boldsymbol{g} = \mathbf{0}^n$ this definition immediately yields the statement 3.

[$3 \Longrightarrow 1$]: By the compactness of the subdifferential [cf. Proposition 6.17(a)] and Weierstrass' Theorem 4.6 the maximum in the expression (6.34) is attained at some $\boldsymbol{g} \in \partial f(\boldsymbol{x}^*)$. It follows that, in the subgradient inequality (6.33), we get that

$$f(\boldsymbol{x}^* + \boldsymbol{p}) \geq f(\boldsymbol{x}^*) + \boldsymbol{g}^{\mathrm{T}}\boldsymbol{p} \geq f(\boldsymbol{x}^*), \qquad \boldsymbol{p} \in \mathbb{R}^n,$$

which is equivalent to the statement 1. ∎

This result implies that a direction $\boldsymbol{p} \in \mathbb{R}^n$ is a descent direction with respect to f at $\boldsymbol{x}$ if and only if $f'(\boldsymbol{x};\boldsymbol{p}) < 0$ holds. This result cannot be extended to non-convex functions, even when the function f is in C^1 or even C^2. [Take $f(x) := x^3$; $x = 0$; $p = -1$; see also the discussion on saddle points in Example 11.2(b).]

6.4.2 Differentiability of the Lagrangian dual function

We consider the inequality constrained problem (6.4), where we make the following standing assumption:

$$f, g_i\,(i = 1, \ldots, m) \in C^0, \; X \text{ is nonempty and compact.} \tag{6.37}$$

Under this assumption, the set of solutions to the Lagrangian subproblem,

$$X(\boldsymbol{\mu}) := \arg \operatorname*{minimum}_{\boldsymbol{x} \in X} \; L(\boldsymbol{x}, \boldsymbol{\mu}), \qquad \boldsymbol{\mu} \in \mathbb{R}^m, \tag{6.38}$$

is nonempty and compact for any choice of dual vector $\boldsymbol{\mu}$ by Weierstrass' Theorem 4.6. We first develop the subdifferentiability properties of the associated dual function q, stated in (6.9). The first result strengthens Theorem 6.4 under these additional assumptions.

Proposition 6.19 (subdifferentiability of the dual function) *Suppose that, in the problem* (6.4), *the compactness condition* (6.37) *holds.*

(a) *The dual function* (6.9) *is finite, continuous and concave on* $\mathbb{R}^m$. *If its supremum over* $\mathbb{R}^m_+$ *is attained, then the optimal solution set therefore is closed and convex.*

(b) *The mapping* $\boldsymbol{\mu} \rightrightarrows X(\boldsymbol{\mu})$ *is closed on* $\mathbb{R}^m$. *If* $X(\bar{\boldsymbol{\mu}})$ *is the singleton set* $\{\bar{\boldsymbol{x}}\}$ *for some* $\bar{\boldsymbol{\mu}} \in \mathbb{R}^m$, *and for some sequence* $\mathbb{R}^m \supset \{\boldsymbol{\mu}_k\} \to \bar{\boldsymbol{\mu}}$, $\boldsymbol{x}_k \in X(\boldsymbol{\mu}_k)$ *for all* k, *then* $\{\boldsymbol{x}_k\} \to \bar{\boldsymbol{x}}$.

(c) *Let* $\boldsymbol{\mu} \in \mathbb{R}^m$. *If* $\boldsymbol{x} \in X(\boldsymbol{\mu})$, *then* $\boldsymbol{g}(\boldsymbol{x})$ *is a subgradient to* q *at* $\boldsymbol{\mu}$, *that is,* $\boldsymbol{g}(\boldsymbol{x}) \in \partial q(\boldsymbol{\mu})$.

(d) *Let* $\boldsymbol{\mu} \in \mathbb{R}^m$. *Then,*

$$\partial q(\boldsymbol{\mu}) = \operatorname{conv}\{\, \boldsymbol{g}(\boldsymbol{x}) \mid \boldsymbol{x} \in X(\boldsymbol{\mu})\,\}.$$

The set $\partial q(\boldsymbol{\mu})$ *is convex and compact. Moreover, if* U *is a bounded set, then* $\cup_{\boldsymbol{\mu} \in U} \partial q(\boldsymbol{\mu})$ *is also bounded.*

(e) *The directional derivative of* q *at* $\boldsymbol{\mu} \in \mathbb{R}^m$ *in the direction of* $\boldsymbol{p} \in \mathbb{R}^m$ *is*

$$q'(\boldsymbol{\mu}; \boldsymbol{p}) = \operatorname*{minimum}_{\boldsymbol{g} \in \partial q(\boldsymbol{\mu})} \boldsymbol{g}^{\mathrm{T}} \boldsymbol{p}.$$

Proof. (a) Theorem 6.4 stated the concavity of q on its effective domain. Weierstrass' Theorem 4.6 states that q is finite on $\mathbb{R}^m$, which is then also its effective domain. The continuity of q follows from that of *any* finite concave function, as we have already seen in Theorem 4.26. The closedness property of the solution set is a direct consequence of the continuity of q (the upper level set then automatically is closed), and complements the result of Theorem 6.9(a).

(b) Let $\{\boldsymbol{\mu}_k\}$ be a sequence of vectors in $\mathbb{R}^m$ converging to $\bar{\boldsymbol{\mu}}$, and let $\boldsymbol{x}_k \in X(\boldsymbol{\mu}_k)$ be arbitrary. Let $\boldsymbol{x}$ be arbitrary in X, and let further $\bar{\boldsymbol{x}} \in X$ be an arbitrary limit point of $\{\boldsymbol{x}_k\}$ (at least one exists by the compactness of X). From the property that for all k,

$$L(\boldsymbol{x}_k, \boldsymbol{\mu}_k) \leq L(\boldsymbol{x}, \boldsymbol{\mu}_k),$$

follows, by the continuity of L, that, in the limit of k in the subsequence in which $\{\boldsymbol{x}_k\}$ converges to $\bar{\boldsymbol{x}}$,

$$L(\bar{\boldsymbol{x}}, \bar{\boldsymbol{\mu}}) \leq L(\boldsymbol{x}, \bar{\boldsymbol{\mu}}),$$

so that $\bar{\boldsymbol{x}} \in X(\bar{\boldsymbol{\mu}})$, as desired. The special case of a singleton set $X(\bar{\boldsymbol{\mu}})$ follows.

(c) Let $\bar{\boldsymbol{\mu}} \in \mathbb{R}^m$ be arbitrary and let $\bar{\boldsymbol{x}} \in X(\bar{\boldsymbol{\mu}})$. We have that

$$\begin{aligned} q(\bar{\boldsymbol{\mu}}) &= \underset{\boldsymbol{y} \in X}{\operatorname{infimum}}\ L(\boldsymbol{y}, \bar{\boldsymbol{\mu}}) = f(\boldsymbol{x}) + \bar{\boldsymbol{\mu}}^{\mathrm{T}} \boldsymbol{g}(\boldsymbol{x}) \\ &= f(\boldsymbol{x}) + \boldsymbol{\mu}^{\mathrm{T}} \boldsymbol{g}(\boldsymbol{x}) + (\bar{\boldsymbol{\mu}} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{g}(\boldsymbol{x}) \geq q(\boldsymbol{\mu}) + (\bar{\boldsymbol{\mu}} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{g}(\boldsymbol{x}), \end{aligned}$$

which implies that $\boldsymbol{g}(\boldsymbol{x}) \in \partial q(\boldsymbol{\mu})$.

(d) The inclusion $\partial q(\boldsymbol{\mu}) \subseteq \operatorname{conv} \{\, \boldsymbol{g}(\boldsymbol{x}) \mid \boldsymbol{x} \in X(\boldsymbol{\mu}) \,\}$ follows from (c) and the convexity of $\partial q(\boldsymbol{\mu})$. The opposite inclusion follows by applying the Separation Theorem 3.24.[5]

(e) See Proposition 6.17(c). ∎

The result in (c) is an independent proof of the concavity of q on $\mathbb{R}^m$.

The result (d) is particularly interesting, because by Carathéodory's Theorem 3.8 every subgradient of q at any point $\boldsymbol{\mu}$ is the convex combination of a finite number (in fact, at most $m+1$) of vectors of the form $\boldsymbol{g}(\boldsymbol{x}^s)$ with $\boldsymbol{x}^s \in X(\boldsymbol{\mu})$. Computationally, this has been utilized to devise efficient (proximal) bundle methods for the Lagrangian dual problem as well as to devise methods to recover primal optimal solutions.

Next, we establish the differentiability of the dual function under additional assumptions.

Proposition 6.20 (differentiability of the dual function) *Suppose that, in the problem* (6.4)*, the compactness condition* (6.37) *holds.*

(a) *Let $\boldsymbol{\mu} \in \mathbb{R}^m$. The dual function q is differentiable at $\boldsymbol{\mu}$ if and only if $\{\, \boldsymbol{g}(\boldsymbol{x}) \mid \boldsymbol{x} \in X(\boldsymbol{\mu}) \,\}$ is a singleton set, that is, if the value of the vector of constraint functions is invariant over the set of solutions $X(\boldsymbol{\mu})$ to the Lagrangian subproblem. Then, we have that*

$$\nabla q(\boldsymbol{\mu}) = \boldsymbol{g}(\boldsymbol{x}),$$

[5] See [BSS93, Theorem 6.3.7] for a detailed proof.

for every $\boldsymbol{x} \in X(\boldsymbol{\mu})$.

(b) *The result in* (a) *holds in particular if the Lagrangian subproblem has a unique solution, that is,* $X(\boldsymbol{\mu})$ *is a singleton set. In particular, this property is satisfied for* $\boldsymbol{\mu} \geq \mathbf{0}^m$ *if further* X *is a convex set,* f *is strictly convex on* X*, and* g_i $(i = 1, \dots, m)$ *are convex, in which case* $q \in C^1$.

Proof. (a) The concave function q is differentiable at the point $\boldsymbol{\mu}$ (where it is finite) if and only if its subdifferential $\partial q(\boldsymbol{\mu})$ there is a singleton, cf. Proposition 6.17(c).

(b) Under either one of the assumptions stated, $X(\boldsymbol{\mu})$ is a singleton, whence the result follows from (a). Uniqueness follows from the convexity of the feasible set and strict convexity of the objective function, according to Proposition 4.10. That $q \in C^1$ follows from the continuity of $\boldsymbol{g}$ and Proposition 6.19(b). ∎

Proposition 6.21 (twice differentiability of the dual function) *Suppose that, in the problem* (6.4), $X = \mathbb{R}^n$*, and* f *and* g_i $(i = 1, \dots, m)$ *are convex functions in* C^2*. Suppose that, at* $\boldsymbol{\mu} \in \mathbb{R}^m$*, the solution* $\boldsymbol{x}$ *to the Lagrangian subproblem not only is unique, but also that the partial Hessian of the Lagrangian is positive definite at the pair* $(\boldsymbol{x}, \boldsymbol{\mu})$*, that is,*

$$\nabla^2_{\boldsymbol{x}\boldsymbol{x}} L(\boldsymbol{x}, \boldsymbol{\mu}) \text{ is positive definite.}$$

Then, the dual function q *is twice differentiable at* $\boldsymbol{\mu}$*, with*

$$\nabla^2 q(\boldsymbol{\mu}) = -\nabla \boldsymbol{g}(\boldsymbol{x})^{\mathrm{T}} [\nabla^2_{\boldsymbol{x}\boldsymbol{x}} L(\boldsymbol{x}, \boldsymbol{\mu})]^{-1} \nabla \boldsymbol{g}(\boldsymbol{x}).$$

Proof. The result follows from the Implicit Function Theorem, which is stated in Chapter 2, applied to the Lagrangian subproblem.[6] ∎

6.5 *Subgradient optimization methods

We begin by establishing the convergence of classic subgradient optimization methods as applied to a general convex optimization problem.

6.5.1 Convex problems

Consider the convex optimization problem to

$$\underset{\boldsymbol{x}}{\text{minimize}} \ f(\boldsymbol{x}), \tag{6.39a}$$

$$\text{subject to } \boldsymbol{x} \in X, \tag{6.39b}$$

[6] See [Ber99, Pages 596–598] for a detailed analysis.

where $f : \mathbb{R}^n \to \mathbb{R}$ is convex and the set $X \subseteq \mathbb{R}^n$ is nonempty, closed and convex.

The subgradient projection algorithm is as follows: select $\boldsymbol{x}_0 \in X$, and for $k = 0, 1, \dots$ generate

$$\boldsymbol{g}_k \in \partial f(\boldsymbol{x}_k), \tag{6.40a}$$

$$\boldsymbol{x}_{k+1} = \operatorname{Proj}_X(\boldsymbol{x}_k - \alpha_k \boldsymbol{g}_k), \tag{6.40b}$$

where the sequence $\{\alpha_k\}$ is generated from one of the following three rules:

The first rule is termed the *divergent series* step length rule, and requires that

$$\alpha_k > 0, \quad k = 0, 1, \dots; \qquad \lim_{k \to \infty} \alpha_k = 0; \qquad \sum_{k=0}^{\infty} \alpha_k = +\infty. \tag{6.41}$$

The second rule adds to the requirements in (6.41) the square-summable restriction

$$\sum_{k=0}^{\infty} \alpha_k^2 < +\infty. \tag{6.42}$$

The conditions in (6.41) allow for convergence to *any* point from *any* starting point, since the total step is infinite, but convergence is therefore also quite slow; the additional condition in (6.42) means fast sequences are selected. An instance of the step length formulas which satisfies both (6.41) and (6.42) is the following:

$$\alpha_k = \gamma + \beta/(k+1), \qquad k = 0, 1, \dots,$$

where $\beta > 0$, $\gamma \geq 0$.

The third step length rule is

$$\alpha_k = \theta_k \frac{f(\boldsymbol{x}_k) - f^*}{\|\boldsymbol{g}_k\|^2}, \qquad 0 < \sigma_1 \leq \theta_k \leq 2 - \sigma_2 < 2, \tag{6.43}$$

where f^* is the optimal value of (6.39). We refer to this step length formula as the *Polyak step*, after the Russian mathematician Boris Polyak who invented the subgradient method in the 1960s together with Ermol'ev and Shor.

How is convergence established for subgradient optimization methods? As shall be demonstrated in Chapters 11 and 12 convergence of algorithms for problems with a *differentiable* objective function is typically based on generating descent directions, and step length rules that result in the sequence $\{\boldsymbol{x}_k\}$ of iterates being strictly descending in the value of f. For the non-differentiable problem at hand, generating descent directions is a difficult task, since it is not true that the negative of an arbitrarily chosen subgradient of f at a non-optimal vector $\boldsymbol{x}$ defines a descent direction.

In *bundle methods* one gathers information from more than one subgradient (hence the term *bundle*) around a current iteration point so that a descent direction can be generated, followed by an inexact line search. We concentrate here on the simpler methodology of subgradient optimization methods, in which we apply the formula (6.40) where the step length α_k is chosen based on very simple rules.

We establish below that if the step length is small enough, an iteration of the subgradient projection method leads to a vector that is closer to the set of optimal solutions. This technical result also motivates the construction of the Polyak step length rule, and hence shows that the convergence of subgradient methods is based on the reduction of the Euclidean distance to the optimal solutions rather than on the reduction of the value of the objective function f.

Proposition 6.22 (decreasing distance to the optimal set) *Suppose that $\boldsymbol{x}_k \in X$ is not optimal in* (6.39), *and that $\boldsymbol{x}_{k+1}$ is given by* (6.40) *for some step length $\alpha_k > 0$.*

Then, for every optimal solution $\boldsymbol{x}^*$ in (6.39),

$$\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\| < \|\boldsymbol{x}_k - \boldsymbol{x}^*\|$$

holds for every step length α_k in the interval

$$\alpha_k \in (0, 2[f(\boldsymbol{x}_k) - f^*]/\|\boldsymbol{g}_k\|^2). \tag{6.44}$$

Proof. We have that

$$\begin{aligned}
\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\|^2 &= \|\mathrm{Proj}_X(\boldsymbol{x}_k - \alpha_k \boldsymbol{g}_k) - \boldsymbol{x}^*\|^2 \\
&= \|\mathrm{Proj}_X(\boldsymbol{x}_k - \alpha_k \boldsymbol{g}_k) - \mathrm{Proj}_X(\boldsymbol{x}^*)\|^2 \\
&\le \|\boldsymbol{x}_k - \alpha_k \boldsymbol{g}_k - \boldsymbol{x}^*\|^2 \\
&= \|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 - 2\alpha_k(\boldsymbol{x}_k - \boldsymbol{x}^*)^{\mathrm{T}}\boldsymbol{g}_k + \alpha_k^2\|\boldsymbol{g}_k\|^2 \\
&\le \|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 - 2\alpha_k[f(\boldsymbol{x}_k) - f^*] + \alpha_k^2\|\boldsymbol{g}_k\|^2 \\
&< \|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2,
\end{aligned}$$

where we have utilized the property that the Euclidean projection is non-expansive (Theorem 4.31), the subgradient inequality (6.33) for convex functions, and the bounds on α_k given by (6.44). ∎

Our first convergence result is based on the divergent series step length formula (6.41), and establishes convergence to the optimal solution set X^* under an assumption on its boundedness. With the other two step length formulas, this condition will be possible to remove.

Recall the definition (3.11) of the minimum distance from a vector to a closed and convex set; our interest is in the distance from an arbitrary vector $\boldsymbol{x} \in \mathbb{R}^n$ to the solution set X^*:

$$\operatorname{dist}_{X^*}(\boldsymbol{x}) := \underset{\boldsymbol{y} \in X^*}{\operatorname{minimum}} \|\boldsymbol{y} - \boldsymbol{x}\|.$$

Theorem 6.23 (convergence of subgradient optimization methods, I) *Let* $\{\boldsymbol{x}_k\}$ *be generated by the method* (6.40), (6.41). *If* X^* *is bounded and the sequence* $\{\boldsymbol{g}_k\}$ *is bounded, then* $\{f(\boldsymbol{x}_k)\} \to f^*$ *and* $\{\operatorname{dist}_{X^*}(\boldsymbol{x}_k)\} \to 0$ *holds.*

Proof. We show that the iterates will eventually belong to an arbitrarily small neighbourhood of the set of optimal solutions to (6.39).

Let $\delta > 0$ and $B^\delta := \{\, \boldsymbol{x} \in \mathbb{R}^n \mid \|\boldsymbol{x}\| \leq \delta \,\}$. Since f is convex, X is nonempty, closed and convex, and X^* is bounded, it follows from [Roc70, Theorem 27.2], applied to the lower semi-continuous, proper[7] and convex function $f + \chi_X$[8] that there exists an $\varepsilon = \varepsilon(\delta) > 0$ such that the level set $\{\, \boldsymbol{x} \in X \mid f(\boldsymbol{x}) \leq f^* + \varepsilon \,\} \subseteq X^* + B^{\delta/2}$; this level set is denoted by X^ε. Moreover, since for all k, $\|\boldsymbol{g}_k\| \leq \sup_s\{\|\boldsymbol{g}_s\|\} < \infty$, and $\{\alpha_k\} \to 0$, there exists an $N(\delta)$ such that $\alpha_k\|\boldsymbol{g}_k\|^2 \leq \varepsilon$ and $\alpha_k\|\boldsymbol{g}_k\| \leq \delta/2$ for all $k \geq N(\delta)$.

The sequel of the proof is based on induction and is organized as follows. In the first part, we show that there exists a finite $k(\delta) \geq N(\delta)$ such that $\boldsymbol{x}_{k(\delta)} \in X^* + B^\delta$. In the second part, we establish that if $\boldsymbol{x}_k$ belongs to $X^* + B^\delta$ for some $k \geq N(\delta)$ then so does $\boldsymbol{x}_{k+1}$, by showing that either $\operatorname{dist}_{X^*}(\boldsymbol{x}_{k+1}) < \operatorname{dist}_{X^*}(\boldsymbol{x}_k)$ holds, or $\boldsymbol{x}_k \in X^\varepsilon$ so that $\boldsymbol{x}_{k+1} \in X^* + B^\delta$ since the step taken is not longer than $\delta/2$.

Let $\boldsymbol{x}^* \in X^*$ be arbitrary. In every iteration k we then have

$$\begin{aligned}
\|\boldsymbol{x}^* - \boldsymbol{x}_{k+1}\|^2 &= \|\boldsymbol{x}^* - \operatorname{Proj}_X(\boldsymbol{x}_k - \alpha_k\boldsymbol{g}_k)\|^2 && \text{(6.45a)}\\
&\leq \|\boldsymbol{x}^* - \boldsymbol{x}_k + \alpha_k\boldsymbol{g}_k\|^2 && \text{(6.45b)}\\
&= \|\boldsymbol{x}^* - \boldsymbol{x}_k\|^2 + \alpha_k\left(2\boldsymbol{g}_k^{\mathrm{T}}(\boldsymbol{x}^* - \boldsymbol{x}_k) + \alpha_k\|\boldsymbol{g}_k\|^2\right), && \text{(6.45c)}
\end{aligned}$$

[7]A *proper* function is a function which is finite at least at some vector and nowhere attains the value $-\infty$. See also Section 1.4.

[8]For any set $S \subset \mathbb{R}^n$ the function χ_S is the *indicator function* of the set S, that is, $\chi_S(\boldsymbol{x}) = 0$ if $\boldsymbol{x} \in S$; and $\chi_S(\boldsymbol{x}) = +\infty$ if $\boldsymbol{x} \notin S$. See also Section 13.1.

where the inequality follows from the projection property. Now, suppose

$$2\boldsymbol{g}_s^{\mathrm{T}}(\boldsymbol{x}^* - \boldsymbol{x}_s) + \alpha_s \|\boldsymbol{g}_s\|^2 < -\varepsilon \tag{6.46}$$

for all $s \geq N(\delta)$. Then, using (6.45) repeatedly, we obtain that for any $k \geq N(\delta)$,

$$\|\boldsymbol{x}^* - \boldsymbol{x}_{k+1}\|^2 < \left\|\boldsymbol{x}^* - \boldsymbol{x}_{N(\delta)}\right\|^2 - \varepsilon \sum_{s=N(\delta)}^{k} \alpha_s,$$

and from (6.40) it follows that the right-hand side of this inequality tends to minus infinity as $k \to \infty$, which clearly is impossible. Therefore,

$$2\boldsymbol{g}_k^{\mathrm{T}}(\boldsymbol{x}^* - \boldsymbol{x}_k) + \alpha_k \|\boldsymbol{g}_k\|^2 \geq -\varepsilon \tag{6.47}$$

for at least one $k \geq N(\delta)$, say $k = k(\delta)$. From the definition of $N(\delta)$, it follows that $\boldsymbol{g}_{k(\delta)}^{\mathrm{T}}(\boldsymbol{x}^* - \boldsymbol{x}_{k(\delta)}) \geq -\varepsilon$. From the definition of a subgradient (cf. Definition 6.16) we have that $f(\boldsymbol{x}^*) - f(\boldsymbol{x}_{k(\delta)}) \geq \boldsymbol{g}_{k(\delta)}^{\mathrm{T}}(\boldsymbol{x}^* - \boldsymbol{x}_{k(\delta)})$, since $\boldsymbol{x}^*, \boldsymbol{x}_{k(\delta)} \in X$. Hence, $f(\boldsymbol{x}_{k(\delta)}) \leq f^* + \varepsilon$, that is, $\boldsymbol{x}_{k(\delta)} \in X^\varepsilon \subseteq X^* + B^{\delta/2} \subset X^* + B^\delta$.

Now, suppose that $\boldsymbol{x}_k \in X^* + B^\delta$ for some $k \geq N(\delta)$. If (6.46) holds for $s = k$, then, by using (6.45), we have that $\|\boldsymbol{x}^* - \boldsymbol{x}_{k+1}\| < \|\boldsymbol{x}^* - \boldsymbol{x}_k\|$ for any $\boldsymbol{x}^* \in X^*$. Hence,

$$\begin{aligned} \operatorname{dist}_{X^*}(\boldsymbol{x}_{k+1}) &\leq \|\operatorname{Proj}_{X^*}(\boldsymbol{x}_k) - \boldsymbol{x}_{k+1}\| < \|\operatorname{Proj}_{X^*}(\boldsymbol{x}_k) - \boldsymbol{x}_k\| \\ &= \operatorname{dist}_{X^*}(\boldsymbol{x}_k) \leq \delta. \end{aligned}$$

Thus, $\boldsymbol{x}_{k+1} \in X^* + B^\delta$. Otherwise, (6.47) must hold and, using the same arguments as above, we obtain that $f(\boldsymbol{x}_k) \leq f^* + \varepsilon$, i.e., $\boldsymbol{x}_k \in X^\varepsilon \subseteq \boldsymbol{x}^* + B^{\delta/2}$. As

$$\begin{aligned} \|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\| &= \|\operatorname{Proj}_X(\boldsymbol{x}_k - \alpha_k \boldsymbol{g}_k) - \boldsymbol{x}_k\| \leq \|\boldsymbol{x}_k - \alpha_k \boldsymbol{g}_k - \boldsymbol{x}_k\| \\ &= \alpha_k \|\boldsymbol{g}_k\| \leq \delta/2 \end{aligned}$$

whenever $k \geq N(\delta)$, it follows that $\boldsymbol{x}_{k+1} \in X^* + B^{\delta/2} + B^{\delta/2} = X^* + B^\delta$.

By induction with respect to $k \geq k(\delta)$, it follows that $\boldsymbol{x}_k \in X^* + B^\delta$ for all $k \geq k(\delta)$. Since this holds for arbitrarily small values of $\delta > 0$ and f is continuous, the theorem follows. ∎

We next introduce the additional requirement (6.42); the resulting algorithm's convergence behaviour is now much more favourable, and the proof is at the same time less technical.

Theorem 6.24 (convergence of subgradient optimization methods, II) *Let $\{\boldsymbol{x}_k\}$ be generated by the method* (6.40), (6.41), (6.42). *If X^* is nonempty and the sequence $\{\boldsymbol{g}_k\}$ is bounded, then $\{f(\boldsymbol{x}_k)\} \to f^*$ and $\{\boldsymbol{x}_k\} \to \boldsymbol{x}^* \in X^*$ holds.*

Proof. Let $\boldsymbol{x}^* \in X^*$ and $k \geq 1$. Repeated application of (6.45) yields

$$\|\boldsymbol{x}^* - \boldsymbol{x}_k\|^2 \leq \|\boldsymbol{x}^* - \boldsymbol{x}_0\|^2 + 2\sum_{s=0}^{k-1} \alpha_s \boldsymbol{g}_s^{\mathrm{T}}(\boldsymbol{x}^* - \boldsymbol{x}_s) + \sum_{s=0}^{k-1} \alpha_s^2 \|\boldsymbol{g}_s\|^2 . \tag{6.48}$$

Since $\boldsymbol{x}^* \in X^*$ and $\boldsymbol{g}_s \in \partial f(\boldsymbol{x}_s)$ for all $s \geq 0$ we obtain that

$$f(\boldsymbol{x}_s) \geq f^* \geq f(\boldsymbol{x}_s) + \boldsymbol{g}_s^{\mathrm{T}}(\boldsymbol{x}^* - \boldsymbol{x}_s), \qquad s \geq 0, \tag{6.49}$$

and hence that $\boldsymbol{g}_s^{\mathrm{T}}(\boldsymbol{x}^* - \boldsymbol{x}_s) \leq 0$ for all $s \geq 0$. Define $c := \sup_k\{\|\boldsymbol{g}_k\|\}$ and $p = \sum_{k=0}^{\infty} \alpha_k^2$, so that $\|\boldsymbol{g}_s\| \leq c$ for any $s \geq 0$ and $\sum_{s=0}^{k-1} \alpha_s^2 < p$. From (6.48) we then conclude that $\|\boldsymbol{x}^* - \boldsymbol{x}_k\|^2 < \|\boldsymbol{x}^* - \boldsymbol{x}_0\|^2 + pc^2$ for any $k \geq 1$, and thus that the sequence $\{\boldsymbol{x}_k\}$ is bounded.

Assume now that there is no subsequence $\{\boldsymbol{x}_{k_i}\}$ of $\{\boldsymbol{x}_k\}$ with $\{\boldsymbol{g}_{k_i}^{\mathrm{T}}(\boldsymbol{x}^* - \boldsymbol{x}_{k_i})\} \to 0$. Then there must exist an $\varepsilon > 0$ with $\boldsymbol{g}_s^{\mathrm{T}}(\boldsymbol{x}^* - \boldsymbol{x}_s) \leq -\varepsilon$ for all sufficiently large values of s. From (6.48) and the conditions on the step lengths it follows that $\{\|\boldsymbol{x}^* - \boldsymbol{x}_s\|\} \to -\infty$, which clearly is impossible. The sequence $\{\boldsymbol{x}_k\}$ must therefore contain a subsequence $\{\boldsymbol{x}_{k_i}\}$ such that $\{\boldsymbol{g}_{k_i}^{\mathrm{T}}(\boldsymbol{x}^* - \boldsymbol{x}_{k_i})\} \to 0$. From (6.49) it follows that $\{f(\boldsymbol{x}_{k_i})\} \to f^*$. The boundedness of $\{\boldsymbol{x}_k\}$ implies the existence of an accumulation point of the subsequence $\{\boldsymbol{x}_{k_i}\}$, say $\boldsymbol{x}^\infty$. From the continuity of f it follows that $\boldsymbol{x}^\infty \in X^*$.

To show that $\boldsymbol{x}^\infty$ is the only accumulation point of $\{\boldsymbol{x}_k\}$, let $\delta > 0$ and choose an $M(\delta)$ such that $\|\boldsymbol{x}^\infty - \boldsymbol{x}_{M(\delta)}\|^2 \leq \delta/2$ and $\sum_{s=M(\delta)}^{\infty} \alpha_s^2 \leq \delta/(2c^2)$. Consider any $k > M(\delta)$. Analogously to the derivation of (6.48), and using (6.49), we then obtain that

$$\|\boldsymbol{x}^\infty - \boldsymbol{x}_k\|^2 \leq \left\|\boldsymbol{x}^\infty - \boldsymbol{x}_{M(\delta)}\right\|^2 + \sum_{s=M(\delta)}^{k-1} \alpha_s^2 \|\boldsymbol{g}_s\|^2 < \frac{\delta}{2} + \frac{\delta}{2c^2}c^2 = \delta.$$

Since this holds for arbitrarily small values of $\delta > 0$, we are done. ∎

Note that the boundedness condition on $\{\boldsymbol{g}_k\}$ is fulfilled whenever we know before-hand that the sequence $\{\boldsymbol{x}_k\}$ is bounded, such as in the case when X itself is bounded; cf. Proposition 6.17(a).

We finally present the even stronger convergence properties of the subgradient projection method using the Polyak step.

Theorem 6.25 (convergence of subgradient optimization methods, III) *Let $\{\boldsymbol{x}_k\}$ be generated by the method* (6.40), (6.43). *If X^* is nonempty then $\{f(\boldsymbol{x}_k)\} \to f^*$ and $\{\boldsymbol{x}_k\} \to \boldsymbol{x}^* \in X^*$ holds.*

Proof. From Proposition 6.22 follows that the sequence $\{\|\boldsymbol{x}_k - \boldsymbol{x}^*\|\}$ is strictly decreasing for every $\boldsymbol{x}^* \in X^*$, and therefore has a limit. By construction of the step length, in which the step lengths are bounded away from zero and $2[f(\boldsymbol{x}_k) - f^*]/\|\boldsymbol{g}_k\|^2$, it follows from the proof of Proposition 6.22 that $\{[f(\boldsymbol{x}_k) - f^*]^2/\|\boldsymbol{g}_k\|^2\} \to 0$ must hold. Since $\{\boldsymbol{g}_k\}$ must be bounded due to the boundedness of $\{\boldsymbol{x}_k\}$ [Proposition 6.17(a)], we have that $\{f(\boldsymbol{x}_k)\} \to f^*$. Further, $\{\boldsymbol{x}_k\}$ is bounded, and due to the continuity property of f every limit point must then belong to X^*.

It remains to show that there can be only one limit point. This property follows from the monotone decrease of the distance $\|\boldsymbol{x}_k - \boldsymbol{x}^*\|$. In detail, the proof is as follows. Suppose two subsequences of $\{\boldsymbol{x}_k\}$ exist, such that they converge to two different vectors in X^*:

$$\{\boldsymbol{x}_{m_i}\} \to \boldsymbol{x}_1^*; \qquad \{\boldsymbol{x}_{l_i}\} \to \boldsymbol{x}_2^*; \qquad \boldsymbol{x}_1^* \neq \boldsymbol{x}_2^*.$$

We must then have $\{\|\boldsymbol{x}_{l_i} - \boldsymbol{x}_1^*\|\} \to \rho > 0$. Since $\boldsymbol{x}_1^* \in X^*$ and the distance to X^* is decreasing, $\{\|\boldsymbol{x}_k - \boldsymbol{x}_1^*\|\} \to \rho$ holds, and in particular $\{\|\boldsymbol{x}_{m_i} - \boldsymbol{x}_1^*\|\} \to \rho$, which is a contradiction. ∎

Contrary to the slow convergence of the subgradient projection algorithms that rely on the divergent series step length rule, under additional conditions on the function f a subgradient algorithm based on the Polyak step length (6.43) is *geometrically convergent*, in the sense that there exist $c > 0$ and $\eta \in (0, 1)$ with

$$\|\boldsymbol{x}_k - \boldsymbol{x}^*\| \leq c\eta^k, \qquad k = 0, 1, \dots .$$

See Section 6.9 for references to other subgradient algorithms than those presented here.

6.5.2 Application to the Lagrangian dual problem

We remind ourselves that the Lagrangian dual problem is a concave maximization problem, and that the appearance of the dual function is similar to that of the following example:

Let $h(x) := \text{minimum}\,\{h_1(x), h_2(x)\}$, where $h_1(x) := 4 - |x|$ and $h_2(x) := 4 - (x - 2)^2$. Then,

$$h(x) = \begin{cases} 4 - x, & \text{if } 1 \leq x \leq 4; \\ 4 - (x-2)^2 & \text{if } x \leq 1,\ x \geq 4; \end{cases}$$

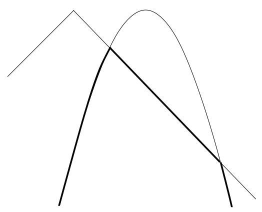

Figure 6.3: A convex min-function with three pieces.

cf. Figure 6.3.

The function h is non-differentiable at $x = 1$ and $x = 4$, since its graph has non-unique supporting hyperplanes there:

$$\partial h(x) = \begin{cases} \{4 - 2x\}, & \text{if } x < 1; \\ [-1, 2], & \text{if } x = 1; \\ \{-1\}, & \text{if } 1 < x < 4; \\ [-4, -1], & \text{if } x = 4; \\ \{4 - 2x\}, & \text{if } x > 4. \end{cases}$$

the subdifferential is here either a singleton (at differentiable points) or a closed interval (at non-differentiable points).

Note the monotonically decreasing nature of the relation $x \mapsto \partial h(x)$. Note also that $0 \in \partial h(1)$, whence $x^* = 1$ defines a maximum over $\mathbb{R}$.

Now, let $\boldsymbol{g} \in \partial q(\bar{\boldsymbol{\mu}})$, and let U^* be the set of optimal solutions to (6.10). Then,

$$U^* \subseteq \{\, \boldsymbol{\mu} \in \mathbb{R}^m \mid \boldsymbol{g}^{\mathrm{T}}(\boldsymbol{\mu} - \bar{\boldsymbol{\mu}}) \geq 0 \,\}.$$

In other words, any subgradient defines a half-space that contains the set of optimal solutions; cf. Figure 6.4. We therefore know that a small enough step in the direction of a subgradient gets us closer to the set of optimal solutions; cf. Proposition 6.22. But again consider Figure 6.4: an arbitrary subgradient, like the on depicted, may not define an ascent direction! As we saw in the previous section, convergence must be based on other arguments, like the decreasing distance to U^* alluded to above and in the previous section. In the next subsection we discuss in brief the generation of ascent directions.

We consider the Lagrangian dual problem (6.10). We suppose, as in the previous section, that X is compact so that the infimum in (6.9) is attained for every $\boldsymbol{\mu} \geq \mathbf{0}^m$ (which is the set over which we wish to maximize q) and q is real-valued over $\mathbb{R}^m_+$.

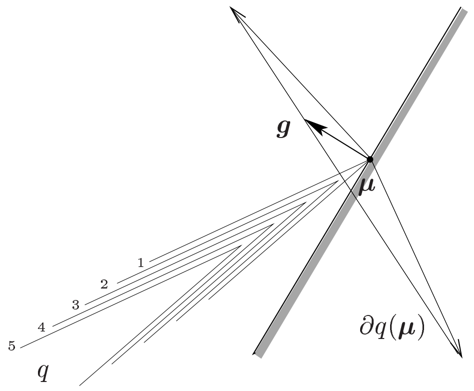

Figure 6.4: The half-space defined by the subgradient $\boldsymbol{g}$ of q at $\boldsymbol{\mu}$. Note that the subgradient is not an ascent direction.

In the case of our special concave maximization problem, the iteration has the form

$$\begin{aligned}\boldsymbol{\mu}_{k+1} &= \operatorname{Proj}_{\mathbb{R}^m_+}[\boldsymbol{\mu}_k + \alpha_k \boldsymbol{g}_k] = [\boldsymbol{\mu}_k + \alpha_k \boldsymbol{g}_k]_+ \\ &= (\text{maximum}\,\{0, (\boldsymbol{\mu}_k)_i + \alpha_k (\boldsymbol{g}_k)_i\})_{i=1}^m, \end{aligned} \tag{6.50}$$

where $\boldsymbol{g}_k \in \partial q(\boldsymbol{\mu}_k)$ is arbitrarily chosen; we would typically use $\boldsymbol{g}_k = \boldsymbol{g}(\boldsymbol{x}_k)$, where $\boldsymbol{x}_k \in \arg\text{minimum}_{\boldsymbol{x} \in X}\ L(\boldsymbol{x}, \boldsymbol{\mu}_k)$. The projection operation onto the first orthant is, as we can see, very simple.

Replacing the Polyak step (6.43) with the corresponding dual form

$$\alpha_k = \theta_k \frac{q^* - q(\boldsymbol{\mu}_k)}{\|\boldsymbol{g}_k\|^2}, \qquad 0 < \sigma_1 \leq \theta_k \leq 2 - \sigma_2 < 2, \tag{6.51}$$

convergence will now be a simple consequence of the above theorems.

The compactness condition (6.37) and the fact that the feasible set of (6.4) is nonempty ensure that the problem (6.4) has an optimal solution; in particular, the feasibility condition (6.5) then holds. Further, if we introduce the Slater condition (6.16), we are ensured that there is no duality gap, and that the dual problem (6.10) has a compact set U^* of optimal solutions. Under these assumptions, we have the following results for subgradient optimization methods.

Theorem 6.26 (convergence of subgradient optimization methods) *Suppose that the problem* (6.4) *is feasible, and that the compactness condition* (6.37) *and the Slater condition* (6.16) *hold.*

(a) *Let* $\{\boldsymbol{\mu}_k\}$ *be generated by the method* (6.50), (6.41). *Then,* $\{q(\boldsymbol{\mu}_k)\} \to q^*$, *and* $\{\operatorname{dist}_{U^*}(\boldsymbol{\mu}_k)\} \to 0$.

(b) *Let* $\{\boldsymbol{\mu}_k\}$ *be generated by the method* (6.50), (6.41), (6.42). *Then,* $\{\boldsymbol{\mu}_k\}$ *converges to an optimal solution to* (6.10).

(c) *Let* $\{\boldsymbol{\mu}_k\}$ *be generated by the method* (6.50), (6.51). *Then,* $\{\boldsymbol{\mu}_k\}$ *converges to an optimal solution to* (6.10).

Proof. The results follow from Theorems 6.23, 6.24, and 6.25, respectively. Note that in the first two cases, boundedness conditions were assumed for X^* and the sequence of subgradients. The corresponding conditions for the Lagrangian dual problem are fulfilled under the CQs imposed, since they imply that the search for an optimal solution is done over a compact set; cf. Theorem 6.9(a) and its proof. ∎

6.5.3 The generation of ascent directions

Proposition 6.18 shows that the *existence* of a descent direction with respect to the convex function $f : \mathbb{R}^n \to \mathbb{R}$ at some $\bar{\boldsymbol{x}} \in \mathbb{R}^n$ hinges on the existence of some vector $\bar{\boldsymbol{p}} \in \mathbb{R}^n$ such that $f'(\bar{\boldsymbol{x}}; \bar{\boldsymbol{p}}) < 0$. According to the definition of the directional derivative and the compactness of $\partial f(\bar{\boldsymbol{x}})$, this is equivalent to the statement that $\boldsymbol{g}^{\mathrm{T}}\bar{\boldsymbol{p}} \leq \varepsilon < 0$ for every $\boldsymbol{g} \in \partial f(\bar{\boldsymbol{x}})$. In the context of Lagrangian duality we show below how we can generate an ascent directions for q at some $\boldsymbol{\mu} \in \mathbb{R}^m$.

Definition 6.27 (steepest ascent direction) *Suppose that the problem* (6.4) *is feasible, and that the compactness condition* (6.37) *holds. Consider the Lagrangian dual problem* (6.10), *and let* $\boldsymbol{\mu} \in \mathbb{R}^m$. *A vector* $\bar{\boldsymbol{p}} \in \mathbb{R}^m$ *is a* steepest ascent *direction if*

$$q'(\boldsymbol{\mu}; \bar{\boldsymbol{p}}) = \underset{\|\boldsymbol{p}\| \leq 1}{\operatorname{maximum}}\; q'(\boldsymbol{\mu}; \boldsymbol{p})$$

holds. ∎

Proposition 6.28 (the shortest subgradient yields the steepest ascent direction) *Suppose that the problem* (6.4) *is feasible, and that the compactness condition* (6.37) *holds. Consider the Lagrangian dual problem* (6.10). *The direction* $\bar{\boldsymbol{p}}$ *of steepest ascent with respect to* q *at* $\boldsymbol{\mu}$ *is given below, where* $\bar{\boldsymbol{g}} \in \partial q(\boldsymbol{\mu})$ *is the shortest subgradient in* $\partial q(\boldsymbol{\mu})$ *with respect to the Euclidean norm:*

$$\bar{\boldsymbol{p}} = \begin{cases} \mathbf{0}^m, & \text{if } \bar{\boldsymbol{g}} = \mathbf{0}^m, \\ \frac{\bar{\boldsymbol{g}}}{\|\bar{\boldsymbol{g}}\|}, & \text{if } \bar{\boldsymbol{g}} \neq \mathbf{0}^m. \end{cases}$$

Proof. By Definition 6.27 and Proposition 6.19(e), the following string of equalities and inequalities can easily be verified:

$$\begin{aligned}
\underset{\|\boldsymbol{p}\|\leq 1}{\text{maximum}}\; q'(\boldsymbol{\mu};\boldsymbol{p}) &= \underset{\|\boldsymbol{p}\|\leq 1}{\text{maximum}}\; \underset{\boldsymbol{g}\in\partial q(\boldsymbol{\mu})}{\text{infimum}}\; \boldsymbol{g}^{\mathrm{T}}\boldsymbol{p} \\
&\leq \underset{\boldsymbol{g}\in\partial q(\boldsymbol{\mu})}{\text{infimum}}\; \underset{\|\boldsymbol{p}\|\leq 1}{\text{maximum}}\; \boldsymbol{g}^{\mathrm{T}}\boldsymbol{p} \\
&= \underset{\boldsymbol{g}\in\partial q(\boldsymbol{\mu})}{\text{infimum}}\; \|\boldsymbol{g}\| \\
&= \|\bar{\boldsymbol{g}}\|. \qquad (6.52)
\end{aligned}$$

If we can construct a direction $\bar{\boldsymbol{p}}$ such that $q'(\boldsymbol{\mu};\bar{\boldsymbol{p}}) = \|\bar{\boldsymbol{g}}\|$ then by (6.52) $\bar{\boldsymbol{p}}$ is the steepest ascent direction. If $\bar{\boldsymbol{g}} = \mathbf{0}^m$ then for $\bar{\boldsymbol{p}} = \mathbf{0}^m$ we obviously have that $q'(\boldsymbol{\mu};\bar{\boldsymbol{p}}) = \|\bar{\boldsymbol{g}}\|$. Suppose then that $\bar{\boldsymbol{p}} \neq \mathbf{0}^m$, and let $\bar{\boldsymbol{p}} := \bar{\boldsymbol{g}}/\|\bar{\boldsymbol{g}}\|$. Note that

$$\begin{aligned}
q'(\boldsymbol{\mu};\boldsymbol{p}) &= \underset{\boldsymbol{g}\in\partial q(\boldsymbol{\mu})}{\text{infimum}}\; \boldsymbol{g}^{\mathrm{T}}\bar{\boldsymbol{p}} = \underset{\boldsymbol{g}\in\partial q(\boldsymbol{\mu})}{\text{infimum}}\; \frac{\bar{\boldsymbol{g}}^{\mathrm{T}}\boldsymbol{g}}{\|\boldsymbol{g}\|} \\
&= \frac{1}{\|\boldsymbol{g}\|}\, \underset{\boldsymbol{g}\in\partial q(\boldsymbol{\mu})}{\text{infimum}}\; \left\{\|\bar{\boldsymbol{g}}\|^2 + \bar{\boldsymbol{g}}^{\mathrm{T}}(\boldsymbol{g}-\bar{\boldsymbol{g}})\right\} \\
&= \|\bar{\boldsymbol{g}}\| + \frac{1}{\|\boldsymbol{g}\|}\, \underset{\boldsymbol{g}\in\partial q(\boldsymbol{\mu})}{\text{infimum}}\; \bar{\boldsymbol{g}}^{\mathrm{T}}(\boldsymbol{g}-\bar{\boldsymbol{g}}). \qquad (6.53)
\end{aligned}$$

Since $\bar{\boldsymbol{g}}$ is the shortest vector in $\partial q(\boldsymbol{\mu})$, then, by the variational inequality characterization of the projection of $\mathbf{0}^m$ onto $\partial q(\boldsymbol{\mu})$ established in Theorem 4.23, we obtain that $\bar{\boldsymbol{g}}^{\mathrm{T}}(\boldsymbol{g}-\bar{\boldsymbol{g}}) \geq 0$ for every $\boldsymbol{g} \in \partial q(\boldsymbol{\mu})$. Hence, $\text{infimum}_{g\in\partial q(\boldsymbol{\mu})}\, \bar{\boldsymbol{g}}^{\mathrm{T}}(\boldsymbol{g}-\bar{\boldsymbol{g}}) = 0$ is achieved at $\bar{g}$. From (6.53) it then follows that $q'(\boldsymbol{\mu},\bar{\boldsymbol{p}}) = \|\bar{\boldsymbol{g}}\|$. We are done. ∎

6.6 *Obtaining a primal solution

It remains for us to show how an optimal dual solution $\boldsymbol{\mu}^*$ can be *translated* into an optimal primal solution $\boldsymbol{x}^*$. Obviously, convexity and strong duality will be needed in general, if we are to be able to utilize the primal–dual optimality characterization in Theorem 6.7. It turns out that the generation of a primal optimum is automatic if q is differentiable at $\boldsymbol{\mu}^*$, which is also the condition under which the famous *Lagrange multiplier method* works. Unfortunately, in many cases, such as for most non-strictly convex optimization problems (like linear programming), this will not be the case, and then the translation work becomes more complex.

We start with the ideal case.

6.6.1 Differentiability at the optimal solution

The following results summarize the optimality conditions for the Lagrangian dual problem (6.10), and their consequences for the availability of a primal optimal solution in the absence of a duality gap.

Proposition 6.29 (optimality conditions for the dual problem) *Suppose that, in the problem* (6.4)*, the compactness condition* (6.37) *holds. Suppose further that the Lagrangian dual problem has an optimal solution,* $\boldsymbol{\mu}^*$*.*

(a) *The dual optimal solution is characterized by the inclusion*

$$\mathbf{0}^m \in -\partial q(\boldsymbol{\mu}^*) + N_{\mathbb{R}^m_+}(\boldsymbol{\mu}^*). \tag{6.54}$$

In other words, there then exists $\boldsymbol{\gamma}^* \in \partial q(\boldsymbol{\mu}^*)$*—an optimality-characterizing subgradient of* q *at* $\boldsymbol{\mu}^*$*—such that*

$$\mathbf{0}^m \leq \boldsymbol{\mu}^* \perp \boldsymbol{\gamma}^* \leq \mathbf{0}^m. \tag{6.55}$$

There exists a finite set of solutions $\boldsymbol{x}^i \in X(\boldsymbol{\mu}^*)$ $(i = 1, \dots, k)$ *where* $k \leq m+1$ *such that*

$$\boldsymbol{\gamma}^* = \sum_{i=1}^{k} \alpha_i \boldsymbol{g}(\boldsymbol{x}^i); \qquad \sum_{i=1}^{k} \alpha_i = 1; \qquad \alpha_i \geq 0,\ i = 1, \dots, k. \tag{6.56}$$

Hence, we have that

$$\sum_{i=1}^{k} \alpha_i \mu_i^* g_i(\boldsymbol{x}^i) = 0, \qquad j = 1, \dots, m. \tag{6.57}$$

(b) *If there is a duality gap, then* q *is non-differentiable at* $\boldsymbol{\mu}^*$*.*

(c) *If* q *is differentiable at* $\boldsymbol{\mu}^*$*, then there is no duality gap. Further, any vector in* $X(\boldsymbol{\mu}^*)$ *then solves the primal problem* (6.4)*.*

Proof. (a) The first result is a direct statement of the optimality conditions of the convex and subdifferentiable program (6.10); the complementarity conditions in (6.55) are an equivalent statement of the inclusion in (6.54).

The second result is an application of Carathéodory's Theorem 3.8 to the compact and convex set $\partial q(\boldsymbol{\mu}^*)$.

(b) The result is established once (c) is.

(c) Let $\bar{\boldsymbol{x}}$ be any vector in $X(\boldsymbol{\mu}^*)$ for which $\nabla q(\boldsymbol{\mu}^*) = \boldsymbol{g}(\bar{\boldsymbol{x}})$ holds, cf. Proposition 6.20(a). We obtain from (6.55) that

$$\mathbf{0}^m \leq \boldsymbol{\mu}^* \perp \boldsymbol{g}(\bar{\boldsymbol{x}}) \leq \mathbf{0}^m.$$

Hence, the pair $(\boldsymbol{\mu}, \bar{\boldsymbol{x}})$ fulfills all the conditions stated in (6.12), so that, by Theorem 6.7, $\bar{\boldsymbol{x}}$ is an optimal solution to (6.4). ∎

Remark 6.30 (the non-coordinability phenomenon and decomposition algorithms) Many interesting problems do not comply with the conditions in (c); for example, linear programming is one where the Lagrangian dual problem often is non-differentiable at every dual optimal solution.[9] This is sometimes called the *non-coordinability phenomenon* (cf. [Las70, DiJ79]). It was in order to cope with this phenomenon that Dantzig–Wolfe decomposition ([DaW60, Las70]) and other column generation algorithms, Benders decomposition ([Ben62, Las70]) and generalized linear programming were developed; noticing that the convex combination of a finite number of candidate primal solutions are sufficient to verify an optimal primal–dual solution [cf. (6.57)], methodologies were developed to generate those vectors algorithmically. See also [LPS99] for overviews on the subject of generating primal optimal solutions from dual optimal ones, and [BSS93, Theorem 6.5.2] for an LP procedure that provides primal feasible solutions for convex programs.

Note that the equation (6.57) in (a) reduces to the complementarity condition that $\mu_i^* g_i(\bar{\boldsymbol{x}}) = 0$ holds, for the *averaged* solution, $\bar{\boldsymbol{x}} := \sum_{i=1}^k \alpha_i \boldsymbol{x}^i$, whenever all the functions g_i are affine. ∎

6.6.2 Everett's Theorem

The next result shows that the solution to the Lagrangian subproblem solves a perturbed version of the original problem. We state the result for the general problem to find

$$
\begin{aligned}
f^* := \operatorname*{infimum}_{\boldsymbol{x}} \ & f(\boldsymbol{x}), && && (6.58)\\
\text{subject to} \quad & \boldsymbol{x} \in X, \\
& g_i(\boldsymbol{x}) \leq 0, && i = 1, \dots, m, \\
& h_j(\boldsymbol{x}) = 0, && j = 1, \dots, \ell,
\end{aligned}
$$

where $f : \mathbb{R}^n \to \mathbb{R}$, $g_i : \mathbb{R}^n \to \mathbb{R}$ $(i = 1, 2, \dots, m)$, and $h_j : \mathbb{R}^n \to \mathbb{R}$ $(j = 1, 2, \dots, \ell)$ are given functions, and $X \subseteq \mathbb{R}^n$.

Theorem 6.31 (Everett's Theorem) *Let $(\boldsymbol{\mu}, \boldsymbol{\lambda}) \in \mathbb{R}^m_+ \times \mathbb{R}^\ell$. Consider the Lagrangian subproblem to*

$$
\operatorname*{minimize}_{\boldsymbol{x} \in X} \left\{ f(\boldsymbol{x}) + \boldsymbol{\mu}^{\mathrm{T}} \boldsymbol{g}(\boldsymbol{x}) + \boldsymbol{\lambda}^{\mathrm{T}} \boldsymbol{h}(\boldsymbol{x}) \right\}. \tag{6.59}
$$

Suppose that $\bar{\boldsymbol{x}}$ is an optimal solution to this problem, and let $\mathcal{I}(\boldsymbol{\mu}) \subseteq \{1, \dots, m\}$ denote the set of indices i for which $\mu_i > 0$.

[9] In other words, even if a Lagrange multiplier vector is known, the Lagrangian subproblem may not identify a primal optimal solution.

(a) $\bar{\boldsymbol{x}}$ *is an optimal solution to the perturbed primal problem to*

$$\begin{aligned} \underset{\boldsymbol{x}}{\text{minimize}} \quad & f(\boldsymbol{x}), && (6.60)\\ \text{subject to} \quad & \boldsymbol{x} \in X, \\ & g_i(\boldsymbol{x}) \leq g_i(\bar{\boldsymbol{x}}), \qquad i \in \mathcal{I}(\bar{\boldsymbol{x}}), \\ & h_j(\boldsymbol{x}) = h_j(\bar{\boldsymbol{x}}), \qquad j = 1, \dots, \ell. \end{aligned}$$

(b) *If* $\bar{\boldsymbol{x}}$ *is feasible in* (6.58) *and* $\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{g}(\bar{\boldsymbol{x}}) = 0$ *holds, then* $\bar{\boldsymbol{x}}$ *solves* (6.58)*, and the pair* $(\boldsymbol{\mu}, \boldsymbol{\lambda})$ *then solves the Lagrangian dual problem.*

Proof. (a) The proof proceeds by showing that the triple $(\bar{\boldsymbol{x}}, \boldsymbol{\mu}, \boldsymbol{\lambda})$ is a saddle point of the function $(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\lambda}) \mapsto f(\boldsymbol{x}) + \boldsymbol{\mu}^{\mathrm{T}}[\boldsymbol{g}(\boldsymbol{x}) - \boldsymbol{g}(\bar{\boldsymbol{x}})] + \boldsymbol{\lambda}^{\mathrm{T}}[\boldsymbol{h}(\boldsymbol{x}) - \boldsymbol{h}(\bar{\boldsymbol{x}})]$ over $X \times \mathbb{R}^m_+ \times \mathbb{R}^\ell$.

Let $\boldsymbol{x}$ satisfy the constraints of (6.60). Since we have that $\boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{h}(\bar{\boldsymbol{x}})$ and $\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{g}(\boldsymbol{x}) \leq \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{g}(\bar{\boldsymbol{x}})$, the optimality of $\bar{\boldsymbol{x}}$ in (6.59) yields

$$\begin{aligned} f(\boldsymbol{x}) + \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{g}(\bar{\boldsymbol{x}}) + \boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{h}(\bar{\boldsymbol{x}}) &\geq f(\boldsymbol{x}) + \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{g}(\boldsymbol{x}) + \boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{h}(\boldsymbol{x}) \\ &\geq f(\bar{\boldsymbol{x}}) + \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{g}(\bar{\boldsymbol{x}}) + \boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{h}(\bar{\boldsymbol{x}}), \end{aligned}$$

which shows that $f(\boldsymbol{x}) \geq f(\bar{\boldsymbol{x}})$. We are done.

(b) $\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{g}(\bar{\boldsymbol{x}}) = 0$ implies that $g_i(\bar{\boldsymbol{x}}) = 0$ for $i \in \mathcal{I}(\boldsymbol{\mu})$; from (a) $\bar{\boldsymbol{x}}$ solves the problem to

$$\begin{aligned} \underset{\boldsymbol{x}}{\text{minimize}} \quad & f(\boldsymbol{x}), && (6.61)\\ \text{subject to} \quad & \boldsymbol{x} \in X, \\ & g_i(\boldsymbol{x}) \leq 0, \qquad i \in \mathcal{I}(\bar{\boldsymbol{x}}), \\ & h_j(\boldsymbol{x}) = 0, \qquad j = 1, \dots, \ell. \end{aligned}$$

In particular, then, since the feasible set of (6.58) is contained in that of (6.61) and $\bar{\boldsymbol{x}}$ is feasible in the former, $\bar{\boldsymbol{x}}$ must also solve (6.58). That the pair $(\boldsymbol{\mu}, \boldsymbol{\lambda})$ solves the dual problem follows by the equality between the primal and dual objective functions at $(\bar{\boldsymbol{x}}, \boldsymbol{\mu}, \boldsymbol{\lambda})$, and weak duality. ∎

One important consequence of the result is that if the right-hand side perturbations $g_i(\bar{\boldsymbol{x}})$ and $h_i(\bar{\boldsymbol{x}})$ all are close to zero, the vector $\bar{\boldsymbol{x}}$ being near-feasible might mean that it is in fact acceptable as an approximate solution to the original problem. (This interpretation hinges on the dualized constraints being *soft* constraints, in the sense that a small violation is acceptable. See Section 1.8 for an introduction to the topic of soft constraints.)

6.7 *Sensitivity analysis

6.7.1 Analysis for convex problems

Consider the inequality constrained convex program (6.4), where $f : \mathbb{R}^n \to \mathbb{R}$ and g_i $(i = 1, \dots, m)$ are convex functions and $X \subseteq \mathbb{R}^n$ is a convex set. Suppose that the problem (6.4) is feasible, and that the compactness condition (6.37) and Slater condition (6.16) hold. This is the classic case where there exist multiplier vectors $\boldsymbol{\mu}^*$, according to Theorem 6.9, and strong duality holds.

For certain types of problems where the duality gap is zero and where there exist primal–dual optimal solutions, we have access to a beautiful theory of *sensitivity analysis*. The classic meaning of the term is the answer to the following question: what is the rate of change in f^* when a constraint right-hand side changes? This question answers important practical questions, like the following in manufacturing: If we buy one unit of additional resource at a given price, or if the demand of a product that we sell increases by a certain amount, then how much additional profit do we make?

We will here provide a basic result which states when this sensitivity analysis of the optimal objective value can be performed for the problem (6.4), and establish that the answer is determined precisely by the value of the Lagrange multiplier vector $\boldsymbol{\mu}^*$, provided that it is unique.

Definition 6.32 (perturbation function) *Consider the function* $p : \mathbb{R}^m \to \mathbb{R} \cup \{\pm\infty\}$ *defined by*

$$p(\boldsymbol{u}) := \underset{\boldsymbol{x}}{\text{infimum}} \quad f(\boldsymbol{x}), \tag{6.62}$$

$$\text{subject to} \quad \boldsymbol{x} \in X, \tag{6.63}$$

$$g_i(\boldsymbol{x}) \le u_i, \quad i = 1, \dots, m, \qquad \boldsymbol{u} \in \mathbb{R}^m;$$

it is called the perturbation function, *or* primal function, *associated with the problem* (6.4). *Its effective domain is the set* $P := \{\, \boldsymbol{u} \in \mathbb{R}^m \mid p(\boldsymbol{u}) < +\infty \,\}$. ∎

Under the above convexity conditions, we can establish that p is a convex function. Indeed, it holds that for any value of the Lagrange multiplier vector $\boldsymbol{\mu}^*$ for the problem (6.4) that

$$\begin{aligned}
q(\boldsymbol{\mu}^*) &= \operatorname*{infimum}_{\boldsymbol{x}\in X} \{f(\boldsymbol{x}) + (\boldsymbol{\mu}^*)^{\mathrm{T}}\boldsymbol{g}(\boldsymbol{x})\} \\
&= \operatorname*{infimum}_{\{\,(\boldsymbol{u},\boldsymbol{x})\in P\times X\,|\,\boldsymbol{g}(\boldsymbol{x})\leq \boldsymbol{u}\,\}} \{f(\boldsymbol{x}) + (\boldsymbol{\mu}^*)^{\mathrm{T}}\boldsymbol{g}(\boldsymbol{x})\} \\
&= \operatorname*{infimum}_{\{\,(\boldsymbol{u},\boldsymbol{x})\in P\times X\,|\,\boldsymbol{g}(\boldsymbol{x})\leq \boldsymbol{u}\,\}} \{f(\boldsymbol{x}) + (\boldsymbol{\mu}^*)^{\mathrm{T}}\boldsymbol{u}\} \\
&= \operatorname*{infimum}_{\boldsymbol{u}\in P} \operatorname*{infimum}_{\{\boldsymbol{x}\in X\,|\,\boldsymbol{g}(\boldsymbol{x})\leq \boldsymbol{u}\,\}} \{f(\boldsymbol{x}) + (\boldsymbol{\mu}^*)^{\mathrm{T}}\boldsymbol{u}\}.
\end{aligned}$$

Since $\boldsymbol{\mu}^*$ is assumed to be a Lagrange multiplier vector, we have that $q(\boldsymbol{\mu}^*) = f^* = p(\mathbf{0}^m)$. By the definition of infimum, then, we have that

$$p(\mathbf{0}^m) \leq p(\boldsymbol{u}) + (\boldsymbol{\mu}^*)^{\mathrm{T}}\boldsymbol{u}, \qquad \boldsymbol{u} \in \mathbb{R}^m,$$

that is, $-\boldsymbol{\mu}^*$ (notice the sign!) is a subgradient of p at $\boldsymbol{u} = \mathbf{0}^m$ (see Definition 6.16). Moreover, by the result in Proposition 6.17(c), p is differentiable at $\mathbf{0}^m$ if and only if p is finite in a neighbourhood of $\mathbf{0}^m$ and $\boldsymbol{\mu}^*$ is a *unique* Lagrange multiplier vector, that is, the Lagrangian dual problem (6.10) has a unique optimal solution. We have therefore proved the following result:

Proposition 6.33 (a sensitivity analysis result) *Suppose that in the inequality constrained problem* (6.4), $f : \mathbb{R}^n \to \mathbb{R}$ *and* $g_i : \mathbb{R}^n \to \mathbb{R}$ $(i = 1, \dots, m)$ *are convex functions and* $X \subseteq \mathbb{R}^n$ *is a convex set. Suppose that the problem* (6.4) *is feasible, and that the compactness assumption* (6.37) *and Slater condition* (6.16) *hold. Suppose further that the perturbed problem defined in* (6.62) *has an optimal solution in a neighbourhood of* $\boldsymbol{u} = \mathbf{0}^m$, *and that on the set of primal–dual optimal solutions to* (6.4)–(6.10), *the dual optimal solution* $\boldsymbol{\mu}^*$ *is unique. Then, the perturbation function* p *is differentiable at* $\boldsymbol{u} = \mathbf{0}^m$, *and*

$$\nabla p(\mathbf{0}^m) = -\boldsymbol{\mu}^*$$

holds. ∎

It is intuitive that the sign of $\nabla p(\mathbf{0}^m)$ should be non-positive; if a right-hand side of the (less-than) inequality constraints in (6.4) increases, then the feasible set becomes larger. [This means that we might be able to find feasible vectors $\boldsymbol{x}$ in the new problem with $f(\boldsymbol{x}) < f^*$, where $f^* = p(0)$ is the optimal value of the minimization problem (6.4).]

The result specializes immediately to linear programming problems, which is the problem type where this type of analysis is most often utilized. The proof of differentiability of the perturbation function at zero for that special case can however be done much more simply. (See Section 10.3.1.)

6.7.2 Analysis for differentiable problems

There exist local versions of the analysis valid also for non-convex problems, where we are interested in the effect of a problem perturbation on a KKT point. A special such analysis was recently performed by Bertsekas [Ber04], in which he shows that even when the problem is non-convex and the set of Lagrange multipliers are not unique, a sensitivity analysis is available as long as data is differentiable. Suppose then that in the problem (6.4) the functions f and g_i, $i = 1, \dots, m$ are in C^1 and that X is nonempty. We generalize the concept of a *Lagrange multiplier vector* to here mean that it is a vector $\boldsymbol{\mu}^*$ associated with a *local* minimum $\boldsymbol{x}^*$ such that

$$\left(\nabla f(\boldsymbol{x}^*) + \sum_{i=1}^{m} \mu_i^* \nabla g_i(\boldsymbol{x}^*)\right)^{\mathrm{T}} \boldsymbol{p} \geq 0, \qquad \boldsymbol{p} \in T_X(\boldsymbol{x}^*), \tag{6.64a}$$

$$\mu_i^* \geq 0, \qquad i = 1, \dots, m, \tag{6.64b}$$

$$\mu_i^* = 0, \qquad i \notin \mathcal{I}(\boldsymbol{x}^*), \tag{6.64c}$$

where $T_X(\boldsymbol{x}^*)$ is the tangent cone to X at $\boldsymbol{x}^*$ (cf. Definition 5.2). Note that under an appropriate CQ this is equivalent to the KKT conditions, in which case we are simply requiring here that $\boldsymbol{x}^*$ is a local minimum.

In the below result we utilize the notation

$$g_i^+(\boldsymbol{x}) := \text{maximum}\,\{0, g_i(\boldsymbol{x})\}, \qquad i = 1, \dots, m,$$

and let $\boldsymbol{g}^+(\boldsymbol{x})$ be the m-vector of elements $g_i^+(\boldsymbol{x})$, $i = 1, \dots, m$.

Theorem 6.34 (sensitivity from the minimum norm multiplier vector) *Suppose that $\boldsymbol{x}^*$ is a local minimum in the problem (6.4), and that the set of Lagrange multipliers is nonempty. Let $\boldsymbol{\mu}^*$ denote the Lagrange multiplier vector of minimum Euclidean norm. Then, for every sequence $\{\boldsymbol{x}_k\} \subset X$ of infeasible vectors such that $\{\boldsymbol{x}_k\} \to \boldsymbol{x}^*$ we have that*

$$f(\boldsymbol{x}^*) - f(\boldsymbol{x}_k) \leq \|\boldsymbol{\mu}^*\| \cdot \|\boldsymbol{g}^+(\boldsymbol{x}_k)\| + o(\|\boldsymbol{x}_k - \boldsymbol{x}^*\|). \tag{6.65}$$

Furthermore, if $\boldsymbol{\mu}^* \neq \mathbf{0}^m$ and $T_X(\boldsymbol{x}^*)$ is convex, the above inequality is sharp in the sense that there exists a sequence of infeasible vectors $\{\boldsymbol{x}_k\} \subset X$ such that

$$\lim_{k \to \infty} \frac{f(\boldsymbol{x}^*) - f(\boldsymbol{x}_k)}{\|\boldsymbol{g}^+(\boldsymbol{x}_k)\|} = \|\boldsymbol{\mu}^*\|,$$

and for this sequence

$$\lim_{k \to \infty} \frac{g_i^+(\boldsymbol{x}_k)}{\|\boldsymbol{g}^+(\boldsymbol{x}_k)\|} = \frac{\mu_i^*}{\|\boldsymbol{\mu}^*\|}, \qquad i = 1, \dots, m,$$

holds. ∎

Theorem 6.34 establishes the optimal rate of cost improvement with respect to infeasible constraint perturbations (in effect, those that imply an enlargement of the feasible set).

We finally remark that under stronger conditions still, even the optimal solution $\boldsymbol{x}^*$ is differentiable. Such a result is reminiscent to the Implicit Function Theorem, which however only covers equality systems. If we are to study the sensitivity of $\boldsymbol{x}^*$ to changes in the right-hand sides of inequality constraints as well, then the analysis becomes complicated due to the fact that we must be able to predict if some active constraints may become inactive in the process. In some circumstances, different directions of change in the right-hand sides may cause different subsets of the active constraints $\mathcal{I}(\boldsymbol{x}^*)$ at $\boldsymbol{x}^*$ to become inactive, and this would most probably then be a non-differentiable point. A sufficient condition (but not necessary, at least in the case of linear constraints) for this to not happen is when $\boldsymbol{x}^*$ is *strictly complementary*, that is, when there exists a multiplier vector $\boldsymbol{\mu}^*$ with $\mu_i^* > 0$ for every $i \in \mathcal{I}(\boldsymbol{x}^*)$.

6.8 Applications

We provide two example applications of Lagrangian duality. The first describes the primal–dual relationship between currents and voltages in an electrical network of devices (voltage sources, diodes, and resistors); this application illustrates that Lagrange multipliers often have direct interpretations. The second application concerns a classic combinatorial optimization problem: the traveling salesman problem. We show how to approximately solve this problem through Lagrangian relaxation and subgradient optimization.

6.8.1 Electrical networks

An *electrical network* (or, *circuit*) is an interconnection of analog electrical elements such as resistors, inductors, capacitors, diodes, and transistors. Its size varies from the smallest integrated circuit to an entire electricity distribution network. A circuit is a network that has at least one closed loop. A network is a connection of 2 or more simple circuit elements, and may not be a circuit. The goal when designing electrical networks for signal processing is to apply a predefined operation on *potential differences* (measured in *volts*) or *currents* (measured in *amperes*). Typical functions for these electrical networks are amplification, oscillation and analog linear algorithmic operations such as addition, subtraction, multiplication, and division. In the case of power distribution networks, engineers design the circuit to transport energy as efficiently as possible while at the same time taking into account economic factors, network safety and redundancy. These networks use components such as power lines, cables, circuit breakers, switches and transformers.

To design any electrical circuits, electrical engineers need to be able to predict the voltages and currents in the circuit. Linear circuits (that is, an electrical network where all elements have a linear current–voltage relation) can be quite easily analyzed through the use of complex numbers and systems of linear equations,[10] while nonlinear elements require a more sophisticated analysis. The classic electrical laws describing the equilibrium state of an electrical network are due to G. Kirchhoff [Kir1847]; referred to as *Kirchhoff's circuit laws* they express in a mathematical form the conservation of charge and energy.[11]

Formally, we let an electrical circuit be described by branches (or, links) connecting nodes. We present a simple example where the only devices are voltage sources, resistors, and diodes. The resulting equilibrium conditions will be shown to be represented as the solution to a strictly convex quadratic program. In general, devices such as resistors can be non-linear, but linearity is assumed throughout this section.

- A *voltage source* maintains a constant branch voltage v_s irrespective of the branch current c_s. The power absorbed by the device is $-v_s c_s$.
- A *diode* permits the branch current c_d to flow in one direction only, but consumes no power regardless of the current or voltage on the branch. Denoting the branch voltage by v_d, the direction condition can be stated as a complementarity condition:

$$c_d \geq 0; \qquad v_d \geq 0; \qquad v_d c_d = 0. \tag{6.66}$$

- A *resistor* consumes power in relation with its resistance, denoted by R_r. We recognize the following law describing the relationship between the branch current and voltage in a linear resistor:

$$v_r = -R_r c_r. \tag{6.67}$$

  The power consumed is given by

$$-v_r c_r = \frac{v_r^2}{R_r} = R_r c_r^2, \tag{6.68}$$

  where we have utilized (6.67) to derive two alternative relations.

[10] For such networks already Maxwell [Max1865] had stated equilibrium conditions.

[11] These laws can be derived from Maxwell's equations, but Kirchhoff preceded Maxwell and derived his equations from work done by G. Ohm.

We must be careful about the direction of flow of currents and voltages, and thus define, for each type of device, a node–branch incidence matrix of the form

$$n_{ij} := \begin{cases} -1, & \text{if branch } j \text{ has node } i \text{ as its origin,} \\ 1, & \text{if branch } j \text{ ends in node } i, \\ 0, & \text{otherwise.} \end{cases}$$

The interpretation of a current flow variable is that the direction is from the negative to the positive terminal of the device, that is, from the origin to the ending node of the branch; a negative variable value will therefore correspond to a flow in the opposite direction. Note that for the diodes, the latter is not allowed, as seen in (6.66).

For the three types of devices we hence yield incidence matrices denoted by $\boldsymbol{N}_S$, $\boldsymbol{N}_R$, and $\boldsymbol{N}_D$, creating a partitioned matrix $\boldsymbol{N} = [\boldsymbol{N}_S\, \boldsymbol{N}_D\, \boldsymbol{N}_R]$. Similarly, we let $\boldsymbol{c} = (\boldsymbol{c}_S^{\mathrm{T}}, \boldsymbol{c}_D^{\mathrm{T}}, \boldsymbol{c}_R^{\mathrm{T}})^{\mathrm{T}}$ and $\boldsymbol{v} = (\boldsymbol{v}_S^{\mathrm{T}}, \boldsymbol{v}_D^{\mathrm{T}}, \boldsymbol{v}_R^{\mathrm{T}})^{\mathrm{T}}$ represent the vectors of branch currents and voltages. We also let $\boldsymbol{p} = (\boldsymbol{p}_S^{\mathrm{T}}, \boldsymbol{p}_D^{\mathrm{T}}, \boldsymbol{p}_R^{\mathrm{T}})^{\mathrm{T}}$ denote the vector of node potentials. Before stating the optimization problem whose minimum describes the equilibrium of the system, we recall the two fundamental equilibrium laws:

Kirchhoff's current law: The sum of all currents entering a node is equal to the sum of all currents leaving the node. In other words, $\boldsymbol{N}\boldsymbol{c} = \boldsymbol{0}$, or,[12]

$$\boldsymbol{N}_S\boldsymbol{c}_S + \boldsymbol{N}_D\boldsymbol{c}_D + \boldsymbol{N}_R\boldsymbol{c}_R = \boldsymbol{0}. \tag{6.69}$$

Kirchhoff's voltage law: The difference between the node potentials at the ends of each branch is equal to the branch voltage. In other words, $\boldsymbol{N}^{\mathrm{T}}\boldsymbol{p} = \boldsymbol{v}$, or,[13]

$$\boldsymbol{N}_S^{\mathrm{T}}\boldsymbol{p} = \boldsymbol{v}_S, \tag{6.70a}$$

$$\boldsymbol{N}_D^{\mathrm{T}}\boldsymbol{p} = \boldsymbol{v}_D, \tag{6.70b}$$

$$\boldsymbol{N}_R^{\mathrm{T}}\boldsymbol{p} = \boldsymbol{v}_R. \tag{6.70c}$$

We summarize the equations representing the characteristics of the electrical devices as follows: For the diodes, (6.66) yields

$$\boldsymbol{v}_D \geq \boldsymbol{0}; \qquad \boldsymbol{c}_D \geq \boldsymbol{0}; \qquad \boldsymbol{v}_D^{\mathrm{T}}\boldsymbol{c}_D = 0. \tag{6.71}$$

For the resistors, (6.67) yields

$$\boldsymbol{v}_R = -\boldsymbol{R}\boldsymbol{c}_R, \tag{6.72}$$

[12] This law is also referred to as the first law, the point rule, the junction rule, and the node law.

[13] This law is a corollary to Ohm's law, and is also referred to as the loop law.

$\boldsymbol{R}$ being the diagonal matrix with elements equal to the values R_r.

Hence, (6.69)–(6.72) represent the equilibrium conditions of the circuit. We will now describe the optimization problem whose optimality conditions are, precisely, (6.69)–(6.72) [note that $\boldsymbol{v}_S$ is fixed]:

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\boldsymbol{c}_R^{\mathrm{T}}\boldsymbol{R}\boldsymbol{c}_R - \boldsymbol{v}_S^{\mathrm{T}}\boldsymbol{c}_S, \\ \text{subject to} \quad & \boldsymbol{N}_S\boldsymbol{c}_S + \boldsymbol{N}_D\boldsymbol{c}_D + \boldsymbol{N}_R\boldsymbol{c}_R = \boldsymbol{0}, \\ & -\boldsymbol{c}_D \leq \boldsymbol{0}. \end{aligned} \tag{6.73}$$

In the problem (6.73) we wish to determine branch currents $\boldsymbol{c}_S$, $\boldsymbol{c}_D$, and $\boldsymbol{c}_R$ so as to minimize the sum of half the energy absorbed in the resistors and the energy loss of the voltage source. Note the sign condition on the diode currents.

Note that this is a convex program with linear constraints, and thus the KKT conditions are both necessary and sufficient for the global optimality of the currents. It is instrumental to check that the KKT conditions for (6.73) are given by (6.69)–(6.72), where the Lagrange multipliers are given by $(\boldsymbol{p}^{\mathrm{T}}, \boldsymbol{v}_D^{\mathrm{T}})^{\mathrm{T}}$.

In the discussion terminating in the Strong Duality Theorem 6.13, we showed that the Lagrangian dual of a strictly convex quadratic optimization problem is yet another convex quadratic optimization problem. In our case, following that development, we can derive the following dual optimization problem in terms of the node potentials $\boldsymbol{p}$ (notice, again, that $\boldsymbol{v}_S$ is fixed):

$$\begin{aligned} \text{maximize} \quad & -\frac{1}{2}\boldsymbol{v}_R^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{v}_R, \\ \text{subject to} \quad & \boldsymbol{N}_S^{\mathrm{T}}\boldsymbol{p} = \boldsymbol{v}_S, \\ & \boldsymbol{N}_D^{\mathrm{T}}\boldsymbol{p} - \boldsymbol{v}_D = \boldsymbol{0}, \\ & \boldsymbol{N}_R^{\mathrm{T}}\boldsymbol{p} - \boldsymbol{v}_R = \boldsymbol{0}, \\ & \boldsymbol{v}_D \geq \boldsymbol{0}. \end{aligned} \tag{6.74}$$

In the dual problem (6.74) the matrix $\boldsymbol{R}^{-1}$ is the diagonal matrix of conductances. The objective function is equivalent to the minimization of the power absorbed by the resistors, and we wish to determine the branch voltages $\boldsymbol{v}_D$ and $\boldsymbol{v}_R$, and the potential vector $\boldsymbol{p}$.

Verify that the KKT conditions for this problem, again, reduce to the equilibrium conditions (6.69)–(6.72). In other words, the Lagrange multipliers for the dual problem (6.74) are the (primal) branch currents.

Finally, let us note that by Theorem 6.13(a) the two problems (6.73) and (6.74) have the same objective value at optimality. That is,

$$\frac{1}{2}\boldsymbol{c}_R^{\mathrm{T}}\boldsymbol{R}\boldsymbol{c}_R + \frac{1}{2}\boldsymbol{v}_R^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{c}_R - \boldsymbol{v}_S^{\mathrm{T}}\boldsymbol{c}_S = 0.$$

By (6.71)–(6.72), the above equation reduces to

$$\boldsymbol{v}_S^{\mathrm{T}}\boldsymbol{c}_S + \boldsymbol{v}_D^{\mathrm{T}}\boldsymbol{c}_D + \boldsymbol{v}_R^{\mathrm{T}}\boldsymbol{c}_R = 0,$$

which is precisely the principle of energy conservation.

6.8.2 A Lagrangian relaxation of the traveling salesman problem

Lagrangian relaxation has shown to be remarkably efficient for some combinatorial optimization problems. This is surprising when taking into account that such problems are integer or mixed-integer problems, which suffer from non-zero duality gaps in general. What then lies behind their popularity?

- One can show that Lagrangian relaxation of an integer program is always at least as good as that of a *continuous relaxation*[14] (in the sense that the value of f_R is higher for Lagrangian relaxation than for a continuous relaxation);
- Together with heuristics for finding primal feasible solution, good feasible solutions are often found;
- The Lagrangian relaxed problems can be made computationally much simpler than the original problem, while still keeping a lot of the structure of the original problem.

6.8.2.1 The traveling salesman problem

Let the graph $\mathcal{G} = (\mathcal{N}, \mathcal{L})$ be defined by a number of cities (or, nodes) $i \in \mathcal{N}$ and undirected links in between subsets of pairs of them: $(i, j) \in \mathcal{L} \subseteq \mathcal{N} \times \mathcal{N}$. Notice that the links (i, j) and (j, i) are identical, and are in $\mathcal{L}$ represented by one non-directed link only.

Let c_{ij} denote the distance between the cities i and j, $\{i, j\} \subset \mathcal{N}$. We introduce the following binary variables:

$$x_{ij} := \begin{cases} 1, & \text{if link } (i,j) \text{ is part of the TSP tour,} \\ 0, & \text{otherwise,} \end{cases} \qquad (i,j) \in \mathcal{L}.$$

[14]The continuous relaxation amounts to removing the integrality conditions, replacing, for example, $x_j \in \{0, 1\}$ by $x_j \in [0, 1]$.

With these definitions, the undirected traveling salesman problem (TSP) is to

$$\begin{aligned}
\underset{\boldsymbol{x}}{\text{minimize}} \quad & \sum_{(i,j)\in\mathcal{L}} c_{ij}x_{ij}, && && (6.75\text{a})\\
\text{subject to} \quad & \sum_{(i,j)\in\mathcal{L}:\{i,j\}\subseteq S} x_{ij} \leq |\mathcal{S}| - 1, && \mathcal{S}\subset\mathcal{N}, && (6.75\text{b})\\
& \sum_{(i,j)\in\mathcal{L}} x_{ij} = n, && && (6.75\text{c})\\
& \sum_{i\in\mathcal{N}:(i,j)\in\mathcal{L}} x_{ij} = 2, && j\in\mathcal{N}, && (6.75\text{d})\\
& x_{ij} \in \{0,1\}, && (i,j)\in\mathcal{L}. && (6.75\text{e})
\end{aligned}$$

The constraints have the following interpretation: (6.75b) implies that there can be no *sub-tours*, that is, a tour where fewer than n cities are visited (if $\mathcal{S} \subset \mathcal{N}$ then there can be at most $|\mathcal{S}| - 1$ links between nodes in the set $\mathcal{S}$, where $|\mathcal{S}|$ is the cardinality–number of members–of the set $\mathcal{S}$); (6.75c) implies that in total n cities must be visited; and (6.75d) implies that each city is connected to two others, such that we make sure to arrive from one city and leave for the next.

This problem is NP-hard, which implies that there is no known polynomial algorithm for solving it. We resort therefore to the use of relaxation techniques, in particular Lagrangian relaxation. We have more than one alternative relaxation to perform: If we Lagrangian relax the tree constraints (6.75b) and (6.75c) the remaining problem is a *2-matching* problem; it can be solved in polynomial time. If we instead Lagrangian relax the degree constraints (6.75d) for every node except for one node the remaining problem is a *1-MST* problem, that is, a special type of minimum spanning tree problem.

The following definition is classic: a *Hamiltonian path (respectively, cycle)* is a path (respectively, cycle) which passes every node in the graph exactly once. Every Hamiltonian cycle is a Hamiltonian path from a node s to another node, t, followed by a link (t, s); a subgraph which consists of a spanning tree plus an extra link such that all nodes have degree two. This is then a feasible solution to the TSP.

A *1-MST* problem is the problem to find an MST in the graph that excludes node s, followed by the addition of the two least expensive links from node s to that tree. If all nodes happen to get degree two, then the 1-MST solution is a traveling salesman tour (that is, a Hamiltonian cycle). The idea behind solving the Lagrangian dual problem is then to find proper multiplier values such that the Lagrangian relaxation will produce feasible solutions.

6.8.2.2 Lagrangian relaxation of the traveling salesman problem

Suppose that we Lagrangian relax the degree constraints (3), except for node 1. We assume that the starting node for the trip, node $s \in \mathcal{N}$, and all the links in $\mathcal{L}$ connected to it, have been removed temporarily (in the 1-MST, this data is re-introduced later), but without changing the notation to reflect this.

The subproblem is the following: a 1-MST defined by

$$\begin{aligned} q(\boldsymbol{\lambda}) &= \underset{\boldsymbol{x}}{\operatorname{minimum}} \sum_{(i,j)\in\mathcal{L}} c_{ij}x_{ij} + \sum_{j\in\mathcal{N}} \lambda_j \left(2 - \sum_{i\in\mathcal{N}:(i,j)\in\mathcal{L}} x_{ij}\right) \\ &= 2\sum_{j\in\mathcal{N}} \lambda_j + \underset{\boldsymbol{x}}{\operatorname{minimum}} \sum_{(i,j)\in\mathcal{L}} (c_{ij} - \lambda_i - \lambda_j)x_{ij}. \end{aligned}$$

We see immediately the role of the Lagrange multipliers: a high (low) value of the multiplier λ_j makes node j attractive (unattractive) in the above 1-MST problem, and will therefore lead to more (less) links being attached to it.

When solving the Lagrangian dual problem, we will use the class of subgradient optimization methods, an overview of which is found in Section 6.5.

What is the updating step in the subgradient method, and what is its interpretation? It is as usual an update in the direction of a subgradient, that is, the direction of

$$h_i(\boldsymbol{x}(\boldsymbol{\lambda})) := 2 - \sum_{i\in\mathcal{N}:(i,j)\in\mathcal{L}} x_{ij}(\boldsymbol{\lambda}), \qquad i \in \mathcal{N},$$

where the value of $x_{ij} \in \{0,1\}$ is the solution to the 1-MST solution with link costs $c_{ij} - \lambda_i - \lambda_j$. We see from the direction formula that

$$\lambda_j^{\text{new}} := \lambda_j + \alpha \left(2 - \sum_{i\in\mathcal{N}:(i,j)\in\mathcal{L}} x_{ij}(\boldsymbol{\lambda})\right), \qquad j \in \mathcal{N},$$

where $\alpha > 0$ is a step length. It is interesting to investigate what the update means:

$$\text{current degree at node } j : \begin{cases} > 2 \Longrightarrow \lambda_j \downarrow \text{(link cost} \uparrow) \\ = 2 \Longrightarrow \lambda_j - \text{(link cost constant)} \\ < 2 \Longrightarrow \lambda_j \uparrow \text{(link cost} \downarrow) \end{cases}$$

In other words, the updating formula in a subgradient method is such that the link cost in the 1-MST subproblem is shifted upwards (downwards) if there are too many (too few) links connected to node j in the 1-MST. We are hence adjusting the *node prices* of the nodes in such a way as to try to influence the 1-MST problem to always choose 2 links per node to connect to.

6.8.2.3 A feasibility heuristic

A feasibility heuristic takes the optimal solution from the Lagrangian minimization problem over $\boldsymbol{x}$ and adjusts it such that a feasible solution to the original problem is constructed. As one cannot predict if, or when, a primal feasible solution will be found directly from the subproblem, the heuristic will provide a solution that can be used in place of an optimal one, should one not be found. Moreover, as we know from Lagrangian duality theory, we then have access to both lower and upper bounds on the optimal value f^* of the original problem, and so we have a quality measure of the feasible solutions found.

A feasibility heuristic which can be used together with our Lagrangian heuristic is as follows.

Identify a path in the 1-MST with many links. Then form a subgraph with the remaining nodes and find a path that passes all of them. Put the two paths together in the best way. The resulting path is a Hamiltonian cycle, that is, a feasible solution.

6.8.2.4 The Philips example

In 1987–1988 an M.Sc. project was performed at the department of mathematics at Linköping University, in cooperation with the company Philips, Norrköping. The project was initiated with the goal to improve the current practice of solving a production planning problem.

The problem was as follows: Philips produce circuit boards, perhaps several hundreds or thousands of the same type. There is a new batch of patterns (holes) to be drilled every day, and perhaps even several such batches per day.

In order to speed up the production process the drilling machine is connected to a microcomputer that selects the ordering of the holes to be drilled automatically, given their coordinates. The algorithm for performing the sorting used to be a simple sorting operation that found, for every fixed x-coordinate, the corresponding y-coordinates and sorted them in increasing order. The movement of the drill was therefore from left to right, and for each fixed x-coordinate the movement was vertical. The time it took to drill the holes on one circuit board was, however, far too long, simply because the drill traveled around a lot without performing any tasks, following a path that was too long. (On the other hand, the actual ordering was very fast to produce!) All in all, the complete batch production took too long because of the poorly planned drill movement.

It was observed that the production planning problem is a traveling salesman problem, where the cities are the holes to be drilled, and the distances between them correspond to the Euclidean distances between them. Therefore, an efficient TSP heuristic was devised and implemented, for use in conjunction with the microcomputer. In fact, it was based on precisely the above Lagrangian relaxation, a subgradient optimization method, and a graph-search type heuristic of the form discussed above.

A typical run with the algorithm took a few minutes, and was always stopped after a fixed number of subgradient iterations; the generation of feasible solutions with the above-mentioned graph search technique was performed at every Kth iteration, where K was set to a value strictly larger than one. (Moreover, feasible solutions were not generated during the first iterations of the dual procedure, because of the poor quality of $\boldsymbol{\lambda}_k$ for low values of k; it is often the case that the traveling salesman tour resulting from the heuristic is better when the multipliers are near-optimal in the Lagrangian dual problem.)

In one of the examples implemented it was found that the optimal path length was in the order to 2 meters, and that the upper and lower bounds on f^* produced lead to the conclusion that the relative error of the path length of the best feasible solution found was *less than 7 %*, a quite good result, also showing that the duality gap for the problem at hand (together with the Lagrangian relaxation chosen) is quite small.

After implementing the new procedure, Philips could report an increase in production by some 70 %. Hence, the slightly longer time it took to provide a better production plan, that is, the traveling salesman tour for the drill to follow, was more than well compensated by the fact that the drilling could be done much faster.

Here is hence an interesting case where Lagrangian relaxation helped to solve a large-scale, complex and difficult problem by utilizing problem structure.

6.9 Notes and further reading

Lagrangian duality has been developed in many sources, including early developments by Arrow, Hurwicz, and Uzawa [AHU58], Everett [Eve63], and Falk [Fal67], and later on by Rockafellar [Roc70]. Our development follows to a large extent that of portions of the text books by Bertsekas [Ber99], Bazaraa et al. [BSS93], and Rockafellar [Roc70].

The Relaxation Theorem 6.1 can almost be considered to be folklore, and can be found in a slightly different form in [Wol98, Proposition 2.3].

The differentiability properties of convex functions were developed largely by Rockafellar [Roc70], whose text we mostly follow.

Subgradient methods were developed in the Soviet Union in the 1960s, predominantly by Ermol'ev, Polyak, and Shor. Text book treatments of subgradient methods are found, for example, in [Sho85, HiL93, Ber99]. Theorem 6.23 is essentially due to Ermol'ev [Erm66]; the proof stems from [LPS96]. Theorem 6.24 is due to Shepilov [She76]; finally, Theorem 6.25 is due to Polyak [Pol69].

Everett's Theorem is due to Everett [Eve63].

Theorem 6.34 stems from [Ber04, Proposition 1.1].

That the equilibrium conditions of an electrical or hydraulic network are attained as the minimum of the total energy loss were known more than a century ago. Mathematical programming models for the electrical network equilibrium problems described in Section 6.8.1 date at least as far back as to Duffin [Duf46, Duf47] and d'Auriac [dAu47]. Duffin constructs his objective function as a sum of integrals of resistance functions. The possibility of viewing the equilibrium problem in at least two related, dual, ways as that of either finding the optimal flows of currents or the optimal potentials was also known early in the analysis of electrical networks; these two principles are written out in [Cro36] in work on pipe networks, and explicitly stated as a pair of primal–dual quadratic programming problems in [Den59]; we followed his development, as represented in [BSS93, Section 1.2.D].

The traveling salesman problem is an essential model problem in combinatorial optimization. Excellent introductions to the field can be found in [Law76, PaS82, NeW88, Wol98, Sch03]. It was the work in [HWC74, Geo74, Fis81, Fis85], among others, in the 1970s and 1980s on the traveling salesman problem and its relatives that made Lagrangian relaxation and subgradient optimization popular, and it remains most popular within the combinatorial optimization field.

6.10 Exercises

Exercise 6.1 (numerical example of Lagrangian relaxation) Consider the convex problem to

$$\begin{aligned} \text{minimize} \quad & \frac{1}{x_1} + \frac{4}{x_2}, \\ \text{subject to} \quad & x_1 + x_2 \leq 4, \\ & x_1, x_2 \geq 0. \end{aligned}$$

(a) Lagrangian relax the first constraint, and write down the resulting implicit dual objective function and the dual problem. Motivate why the relaxed problem always has a unique optimum, whence the dual objective function is everywhere differentiable.

(b) Solve the implicit Lagrangian dual problem by utilizing that the gradient to a differentiable dual objective function can be expressed by using the functions that are involved in the relaxed constraints and the unique solution to the relaxed problem.

(c) Write down an explicit Lagrangian dual problem, that is, a dual problem only in terms of the Lagrange multipliers. Solve it, and confirm the results in (b).

(d) Find the original problem's optimal solution.

(e) Show that strong duality holds. ■

Exercise 6.2 (global optimality conditions) Consider the problem to

$$\begin{aligned} \text{minimize } & f(\boldsymbol{x}) := x_1 + 2x_2^2 + 3x_3^3, \\ \text{subject to } & x_1 + 2x_2 + x_3 \leq 3, \\ & 2x_1^2 + x_2 \geq 2, \\ & 2x_1 + x_3 = 2, \\ & x_j \geq 0, \qquad j = 1, 2, 3. \end{aligned}$$

(a) Formulate the Lagrangian dual problem that results from Lagrangian relaxing all but the sign constraints.

(b) State the global primal–dual optimality conditions. ■

Exercise 6.3 (Lagrangian relaxation) Consider the problem to

$$\begin{aligned} \text{minimize } & f(\boldsymbol{x}) := x_1^2 + 2x_2^2, \\ \text{subject to } & x_1 + x_2 \geq 2, \\ & x_1^2 + x_2^2 \leq 5. \end{aligned}$$

Find an optimal solution through Lagrangian duality. ■

Exercise 6.4 (Lagrangian relaxation) In many circumstances it is of interest to calculate the Euclidean projection of a vector onto a subspace. Especially, consider the problem to find the Euclidean projection of the vector $\boldsymbol{y} \in \mathbb{R}^n$ onto the null space of the matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, that is, to find an $\boldsymbol{x} \in \mathbb{R}^n$ that solves the problem to

$$\begin{aligned} \text{minimize } & f(\boldsymbol{x}) := \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^2, \\ \text{subject to } & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}^m, \end{aligned}$$

where $\boldsymbol{A}$ is such that rank $\boldsymbol{A} = m$.

The solution to this problem is classic: the projection is given by

$$\boldsymbol{x}^* = \boldsymbol{y} - \boldsymbol{A}^{\mathrm{T}}(\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}})^{-1}\boldsymbol{A}\boldsymbol{y}.$$

If we let $\boldsymbol{P} := \boldsymbol{I}^n - \boldsymbol{A}^{\mathrm{T}}(\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}})^{-1}\boldsymbol{A}$, where $\boldsymbol{I}^n \in \mathbb{R}^{n\times n}$ is the unit matrix, be the *projection matrix*, the formula is simply $\boldsymbol{x}^* = \boldsymbol{P}\boldsymbol{y}$.

Your task is to derive this formula by utilizing Lagrangian duality. Motivate every step made by showing that the necessary properties are fulfilled.

[*Note:* This exercise is similar to that in Example 5.51, but utilizes Lagrangian duality rather than the KKT conditions to derive the projection formula.] ■

Exercise 6.5 (Lagrangian relaxation, exam 040823) Consider the following linear optimization problem:

$$\begin{aligned} \text{minimize } & f(x,y) := x - 0.5y, \\ \text{subject to } & -x + y \leq -1, \\ & -2x + y \leq -2, \\ & (x,y) \in \mathbb{R}^2_+. \end{aligned}$$

(a) Show that the problem satisfies Slater's constraint qualification. Derive the Lagrangian dual problem corresponding to the Lagrangian relaxation of the two linear inequality constraints, and show that its set of optimal solutions is convex and bounded.

(b) Calculate the set of subgradients of the Lagrangian dual function at the dual points $(1/4, 1/3)^{\mathrm{T}}$ and $(1,0)^{\mathrm{T}}$. ■

Exercise 6.6 (Lagrangian relaxation) Provide an explicit form of the Lagrangian dual problem for the problem to

$$\begin{aligned} \text{minimize } \quad & \sum_{i=1}^{m}\sum_{j=1}^{n} x_{ij}\ln x_{ij} \\ \text{subject to } \quad & \sum_{i=1}^{m} x_{ij} = b_j, \qquad j = 1,\dots,n, \\ & \sum_{j=1}^{n} x_{ij} = a_i, \qquad i = 1,\dots,m, \\ & x_{ij} \geq 0, \qquad i = 1,\dots,m, \quad j = 1,\dots,n, \end{aligned}$$

where $a_i > 0$, $b_j > 0$ for all i, j, and where the linear equalities are Lagrangian relaxed. ■

Exercise 6.7 (Lagrangian relaxation) Given is the problem to

$$
\begin{aligned}
\underset{\boldsymbol{x}}{\text{minimize}} \quad & f(\boldsymbol{x}) = 2x_1^2 + x_2^2 + x_1 - 3x_2, & (6.76a)\\
\text{subject to} \quad & x_1^2 + x_2 \geq 8, & (6.76b)\\
& x_1 \in [1, 3], & (6.76c)\\
& x_2 \in [2, 5]. & (6.76d)
\end{aligned}
$$

Lagrangian relax the constraint (6.76b) with a multiplier μ. Formulate the Lagrangian dual problem and calculate the dual function's value at $\mu = 1$, $\mu = 2$, and $\mu = 3$. Within which interval lies the optimal value f^*? Also, draw the dual function. ■

Exercise 6.8 (Lagrangian duality for integer problems) Consider the primal problem to

$$
\begin{aligned}
\text{minimize} \quad & f(\boldsymbol{x}),\\
\text{subject to} \quad & \boldsymbol{g}(\boldsymbol{x}) \leq \boldsymbol{0}^m,\\
& \boldsymbol{x} \in X,
\end{aligned}
$$

where $X \subseteq \mathbb{R}^n$, $f : \mathbb{R}^n \to \mathbb{R}$, and $\boldsymbol{g} : \mathbb{R}^n \to \mathbb{R}^m$. If the restrictions $\boldsymbol{g}(\boldsymbol{x}) \leq \boldsymbol{0}^m$ are complicating side constraints which are Lagrangian relaxed, we obtain the Lagrangian dual problem to

$$
\underset{\boldsymbol{\mu} \geq \boldsymbol{0}^m}{\text{maximize}} \quad q(\boldsymbol{\mu}),
$$

where

$$
q(\boldsymbol{\mu}) := \underset{\boldsymbol{x} \in X}{\text{minimum}} \, \{ f(\boldsymbol{x}) + \boldsymbol{\mu}^{\mathrm{T}} \boldsymbol{g}(\boldsymbol{x}) \}, \qquad \boldsymbol{\mu} \in \mathbb{R}^m.
$$

(a) Suppose that the set X is finite (for example, consisting of a finite number of integer vectors). Denote the elements of X by $\boldsymbol{x}^p$, $p = 1, \dots, P$. Show that the dual objective function is piece-wise linear. How many linear segments can it have, at most? Why is it *not always* built up by that many segments?

[*Note:* This property holds regardless of any properties of f and $\boldsymbol{g}$.]

(b) Illustrate the result in (a) on the linear 0/1 problem to find

$$
\begin{array}{rcrcrcrcrcl}
z^* = \text{maximum } z = & 5x_1 & + & 8x_2 & + & 7x_3 & + & 9x_4, & & \\
\text{subject to} \quad & 3x_1 & + & 2x_2 & + & 2x_3 & + & 4x_4 & \leq & 5,\\
& 2x_1 & + & x_2 & + & 2x_3 & + & x_4 & = & 3,\\
& x_1 & , & x_2 & , & x_3 & , & x_4 & = & 0/1,
\end{array}
$$

where the first constraint is considered complicating.

(c) Suppose that the function f and all components of $\boldsymbol{g}$ are linear, and that the set X is a polytope (that is, a bounded polyhedron). Show that the dual objective function is also in this case piece-wise linear. How many linear pieces can it be built from, at most? ■

Exercise 6.9 (Lagrangian relaxation) Consider the problem to

$$\begin{array}{llllll} \text{minimize } z = & 2x_1 & + & x_2, & & \\ \text{subject to} & x_1 & + & x_2 & \geq & 5, \\ & x_1 & & & \leq & 4, \\ & & & x_2 & \leq & 4, \\ & x_1 & , & x_2 & \geq & 0, \quad \text{integer}. \end{array}$$

Lagrangian relax the first constraint. Describe the Lagrangian function and the dual problem. Calculate the Lagrangian dual function at these four points: $\mu = 0, 1, 2, 3$. Give the best lower and upper bounds on the optimal value of the original problem that you have found. ■

Exercise 6.10 (surrogate relaxation) Consider an optimization problem of the form

$$\begin{aligned} &\text{minimize } f(\boldsymbol{x}), \\ &\text{subject to } g_i(\boldsymbol{x}) \leq 0, \qquad i = 1, \dots, m, \qquad (P) \\ &\qquad\qquad \boldsymbol{x} \in X, \end{aligned}$$

where the functions $f, g_i : \mathbb{R}^n \to \mathbb{R}$ are continuous and the set $X \subset \mathbb{R}^n$ is closed and bounded. The problem is assumed to have an optimal solution, $\boldsymbol{x}^*$. Introduce parameters $\mu_i \geq 0$, $i = 1, \dots, m$, and define

$$\begin{aligned} s(\boldsymbol{\mu}) := \text{minimum } & f(\boldsymbol{x}), \\ \text{subject to } & \boldsymbol{\mu}^{\mathrm{T}} \boldsymbol{g}(\boldsymbol{x}) \leq 0, \qquad (S) \\ & \boldsymbol{x} \in X. \end{aligned}$$

This problem therefore has exactly one explicit constraint.

(a) [weak duality] Show that $\boldsymbol{x}^*$ is a feasible solution to the problem (S) and that $s(\boldsymbol{\mu}) \leq f^*$ therefore always holds, that is, the problem (S) is a *relaxation* of the original one. Motivate also why $\text{maximum}_{\boldsymbol{\mu} \geq \mathbf{0}^m} s(\boldsymbol{\mu}) \leq f^*$ must hold. Explain the potential usefulness of this result!

(b) [example] Consider the linear 0/1 problem

$$\begin{array}{lllr} z^* = \text{maximum } z = & 5x_1 + 8x_2 + 7x_3 + 9x_4, & & \\ \text{subject to} & 3x_1 + 2x_2 + 3x_3 + 3x_4 \leq 6, & & (1) \\ & 2x_1 + 3x_2 + 3x_3 + 4x_4 \leq 5, & & (2) \\ & 2x_1 + x_2 + 2x_3 + x_4 = 3, & & \\ & x_1, \; x_2, \; x_3, \; x_4 \in 0/1. & & \end{array}$$

Surrogate relax the constraints (1) and (2) with multipliers $\mu_1, \mu_2 \geq 0$ and formulate the problem (S). Let $\bar{\boldsymbol{\mu}} = (1, 2)^{\mathrm{T}}$. Calculate $s(\bar{\boldsymbol{\mu}})$.

Consider again the original problem and Lagrangian relax the constraints (1) and (2) with multipliers $\mu_1, \mu_2 \geq 0$. Calculate the Lagrangian dual objective value at $\boldsymbol{\mu} = \bar{\boldsymbol{\mu}}$.

Compare the two results!

(c) [comparison with Lagrangian duality] Let $\boldsymbol{\mu} \geq \mathbf{0}^m$ and

$$q(\boldsymbol{\mu}) := \underset{\boldsymbol{x} \in X}{\operatorname{minimum}} \, \{ f(\boldsymbol{x}) + \boldsymbol{\mu}^{\mathrm{T}} \boldsymbol{g}(\boldsymbol{x}) \}.$$

Show that $q(\boldsymbol{\mu}) \leq s(\boldsymbol{\mu})$, and that

$$\underset{\boldsymbol{\mu} \geq \mathbf{0}^m}{\operatorname{maximum}} \, q(\boldsymbol{\mu}) \leq \underset{\boldsymbol{\mu} \geq \mathbf{0}^m}{\operatorname{maximum}} \, s(\boldsymbol{\mu}) \leq f^*$$

holds. ■