Chapter 11
The KKT Conditions

In this chapter we will further develop the KKT conditions discussed in Chapter 10, but we will consider general constraints and not restrict ourselves to linear constraints. The Karush–Kuhn–Tucker (KKT) conditions were originally named after Harold Kuhn and Albert Tucker, who first published the conditions in 1951. Later on it was discovered that William Karush developed the necessary conditions in his master's thesis back in 1939, and the conditions were thus named after the three researchers.

11.1 ▪ Inequality Constrained Problems

We will begin our exploration into the KKT conditions by analyzing the inequality constrained problem

$$\text{(P)} \quad \begin{array}{ll} \min & f(\mathbf{x}) \\ \text{s.t.} & g_i(\mathbf{x}) \leq 0, \quad i = 1, 2, \ldots, m, \end{array} \tag{11.1}$$

where $f, g_1, \ldots, g_m$ are continuously differentiable functions over $\mathbb{R}^n$. Our first task is to develop necessary optimality conditions. For that, we will define the concept of a feasible descent direction.

Definition 11.1 (feasible descent directions). *Consider the problem*

$$\begin{array}{ll} \min & h(\mathbf{x}) \\ \textit{s.t.} & \mathbf{x} \in C, \end{array}$$

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

Obviously, a necessary local optimality condition of a point $\mathbf{x}$ is that it does not have any feasible descent directions.

Lemma 11.2. *Consider the problem*

$$\text{(G)} \quad \begin{array}{ll} \min & f(\mathbf{x}) \\ \textit{s.t.} & \mathbf{x} \in C, \end{array}$$

where f is a continuously differentiable function over the set $C \subseteq \mathbb{R}^n$. If $\mathbf{x}^*$ is a local optimal solution of (G), *then there are no feasible descent directions at $\mathbf{x}^*$.*

Proof. The proof is by contradiction. If there is a feasible descent direction, that is, a vector $\mathbf{d}$ and $\varepsilon_1 > 0$ such that $\mathbf{x}^* + t\mathbf{d} \in C$ for all $t \in [0, \varepsilon_1]$ and $\nabla f(\mathbf{x}^*)^T \mathbf{d} < 0$, then by the definition of the directional derivative (see also Lemma 4.2) there is an $\varepsilon_2 < \varepsilon_1$ such that $f(\mathbf{x}^* + t\mathbf{d}) < f(\mathbf{x}^*)$ for all $t \in [0, \varepsilon_2]$, which is a contradiction to the local optimality of $\mathbf{x}^*$. $\square$

We can now write a necessary condition for local optimality in the form of an infeasibility of a set of strict linear inequalities. Before doing that, we mention the following terminology: Given a set of inequalities

$$g_i(\mathbf{x}) \leq 0, \qquad i = 1, 2, \ldots, m,$$

where $g_i : \mathbb{R}^n \to \mathbb{R}$ are functions, and a vector $\tilde{\mathbf{x}} \in \mathbb{R}^n$, the *active constraints* at $\tilde{\mathbf{x}}$ are the constraints satisfied as equalities at $\tilde{\mathbf{x}}$. The set of active constraints is denoted by

$$I(\tilde{\mathbf{x}}) = \{i : g_i(\tilde{\mathbf{x}}) = 0\}.$$

Lemma 11.3. *Let $\mathbf{x}^*$ be a local minimum of the problem*

$$\begin{array}{ll} \min & f(\mathbf{x}) \\ s.t. & g_i(\mathbf{x}) \leq 0, \quad i = 1, 2, \ldots, m, \end{array}$$

where $f, g_1, \ldots, g_m$ are continuously differentiable functions over $\mathbb{R}^n$. Let $I(\mathbf{x}^*)$ be the set of active constraints at $\mathbf{x}^*$:

$$I(\mathbf{x}^*) = \{i : g_i(\mathbf{x}^*) = 0\}.$$

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

$$\begin{array}{ll} \nabla f(\mathbf{x}^*)^T \mathbf{d} < 0, & \\ \nabla g_i(\mathbf{x}^*)^T \mathbf{d} < 0, & i \in I(\mathbf{x}^*). \end{array} \tag{11.2}$$

Proof. Suppose by contradiction that $\mathbf{d}$ satisfies the system of inequalities (11.2). Then by Lemma 4.2, it follows that there exists $\varepsilon_1 > 0$ such that $f(\mathbf{x}^* + t\mathbf{d}) < f(\mathbf{x}^*)$ and $g_i(\mathbf{x}^* + t\mathbf{d}) < g_i(\mathbf{x}^*) = 0$ for any $t \in (0, \varepsilon_1)$ and $i \in I(\mathbf{x}^*)$. For any $i \notin I(\mathbf{x}^*)$ we have that $g_i(\mathbf{x}^*) < 0$, and hence, by the continuity of g_i for all i, it follows that there exists $\varepsilon_2 > 0$ such that $g_i(\mathbf{x}^* + t\mathbf{d}) < 0$ for any $t \in (0, \varepsilon_2)$ and $i \notin I(\mathbf{x}^*)$. We can thus conclude that

$$\begin{array}{l} f(\mathbf{x}^* + t\mathbf{d}) < f(\mathbf{x}^*), \\ g_i(\mathbf{x}^* + t\mathbf{d}) < 0, \quad i = 1, 2, \ldots, m, \end{array}$$

for all $t \in (0, \min\{\varepsilon_1, \varepsilon_2\})$, which is a contradiction to the local optimality of $\mathbf{x}^*$. $\square$

We have thus shown that a necessary optimality condition for local optimality is the infeasibility of a certain system of strict inequalities. We can now invoke Gordan's theorem of the alternative (Theorem 10.4) in order to obtain the so-called Fritz-John conditions.

Theorem 11.4 (Fritz-John conditions for inequality constrained problems). *Let $\mathbf{x}^*$ be a local minimum of the problem*

$$\begin{array}{ll} \min & f(\mathbf{x}) \\ s.t. & g_i(\mathbf{x}) \leq 0, \quad i = 1, 2, \ldots, m, \end{array}$$

where $f, g_1, \ldots, g_m$ are continuously differentiable functions over $\mathbb{R}^n$. Then there exist multipliers $\lambda_0, \lambda_1, \ldots, \lambda_m \geq 0$, which are not all zeros, such that

$$\lambda_0 \nabla f(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i \nabla g_i(\mathbf{x}^*) = 0, \tag{11.3}$$
$$\lambda_i g_i(\mathbf{x}^*) = 0, \quad i = 1, 2, \ldots, m.$$

Proof. By Lemma 11.3 it follows that the following system of inequalities does not have a solution:

$$\text{(S)} \quad \begin{array}{ll} \nabla f(\mathbf{x}^*)^T \mathbf{d} < 0, & \\ \nabla g_i(\mathbf{x}^*)^T \mathbf{d} < 0, & i \in I(\mathbf{x}^*), \end{array} \tag{11.4}$$

where $I(\mathbf{x}^*) = \{i : g_i(\mathbf{x}^*) = 0\} = \{i_1, i_2, \ldots, i_k\}$. System (S) can be rewritten as

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

where

$$\mathbf{A} = \begin{pmatrix} \nabla f(\mathbf{x}^*)^T \\ \nabla g_{i_1}(\mathbf{x}^*)^T \\ \vdots \\ \nabla g_{i_k}(\mathbf{x}^*)^T \end{pmatrix}.$$

By Gordan's theorem of the alternative (Theorem 10.4), system (S) is infeasible if and only if there exists a vector $\boldsymbol{\eta} = (\lambda_0, \lambda_{i_1}, \ldots, \lambda_{i_k})^T \neq 0$ such that

$$\mathbf{A}^T \boldsymbol{\eta} = 0, \qquad \boldsymbol{\eta} \geq 0,$$

which is the same as

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

Define $\lambda_i = 0$ for any $i \notin I(\mathbf{x}^*)$, and we obtain that

$$\lambda_0 \nabla f(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i \nabla g_i(\mathbf{x}^*) = 0$$

and that $\lambda_i g_i(\mathbf{x}^*) = 0$ for any $i \in \{1, 2, \ldots, m\}$ as required. $\square$

A major drawback of the Fritz-John conditions is in the fact that they allow λ_0 to be zero. The case $\lambda_0 = 0$ is not particularly informative since condition (11.3) then becomes

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

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

linearly independent at $\mathbf{x}^*$, then we can establish the KKT conditions, which are the same as the Fritz-John conditions with $\lambda_0 = 1$.

Theorem 11.5 (KKT conditions for inequality constrained problems). *Let $\mathbf{x}^*$ be a local minimum of the problem*

$$\begin{array}{ll} \min & f(\mathbf{x}) \\ s.t. & g_i(\mathbf{x}) \leq 0, \quad i = 1, 2, \ldots, m, \end{array}$$

where $f, g_1, \ldots, g_m$ are continuously differentiable functions over $\mathbb{R}^n$. Let

$$I(\mathbf{x}^*) = \{i : g_i(\mathbf{x}^*) = 0\}$$

be the set of active constraints. Suppose that the gradients of the active constraints $\{\nabla g_i(\mathbf{x}^*)\}_{i \in I(\mathbf{x}^*)}$ are linearly independent. Then there exist multipliers $\lambda_1, \lambda_2, \ldots, \lambda_m \geq 0$ such that

$$\nabla f(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i \nabla g_i(\mathbf{x}^*) = 0, \tag{11.5}$$

$$\lambda_i g_i(\mathbf{x}^*) = 0, \quad i = 1, 2, \ldots, m. \tag{11.6}$$

Proof. By the Fritz-John conditions it follows that there exist $\tilde{\lambda}_0, \tilde{\lambda}_1, \ldots, \tilde{\lambda}_m \geq 0$, not all zeros, such that

$$\tilde{\lambda}_0 \nabla f(\mathbf{x}^*) + \sum_{i=1}^{m} \tilde{\lambda}_i \nabla g_i(\mathbf{x}^*) = 0, \tag{11.7}$$

$$\tilde{\lambda}_i g_i(\mathbf{x}^*) = 0, \quad i = 1, 2, \ldots, m. \tag{11.8}$$

We have that $\tilde{\lambda}_0 \neq 0$ since otherwise, if $\tilde{\lambda}_0 = 0$, by (11.7) and (11.8) it follows that

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

where not all the scalars $\tilde{\lambda}_i, i \in I(\mathbf{x}^*)$ are zeros, leading to a contradiction to the basic assumption that $\{\nabla g_i(\mathbf{x}^*)\}_{i \in I(\mathbf{x}^*)}$ are linearly independent, and hence $\tilde{\lambda}_0 > 0$. Defining $\lambda_i = \frac{\tilde{\lambda}_i}{\tilde{\lambda}_0}$, the result directly follows from (11.7) and (11.8). $\square$

The condition that the gradients of the active constraints are linearly independent is one of many types of assumptions that are referred to in the literature as "constraint qualifications."

11.2 ▪ Inequality and Equality Constrained Problems

By using the implicit function theorem, it is possible to generalize the KKT conditions for problems involving also equality constraints. We will state this generalization without a proof.

Theorem 11.6 (KKT conditions for inequality/equality constrained problems). *Let $\mathbf{x}^*$ be a local minimum of the problem*

$$\begin{array}{ll} \min & f(\mathbf{x}) \\ s.t. & g_i(\mathbf{x}) \leq 0, \quad i = 1, 2, \ldots, m, \\ & h_j(\mathbf{x}) = 0, \quad j = 1, 2, \ldots, p. \end{array} \tag{11.9}$$

where $f, g_1, \ldots, g_m, h_1, h_2, \ldots, h_p$ are continuously differentiable functions over $\mathbb{R}^n$. Suppose that the gradients of the active constraints and the equality constraints

$$\{\nabla g_i(\mathbf{x}^*) : i \in I(\mathbf{x}^*)\} \cup \{\nabla h_j(\mathbf{x}^*) : j = 1, 2, \ldots, p\}$$

are linearly independent (where as before $I(\mathbf{x}^*) = \{i : g_i(\mathbf{x}^*) = 0\}$). Then there exist multipliers $\lambda_1, \lambda_2 \ldots, \lambda_m \geq 0$ and $\mu_1, \mu_2, \ldots, \mu_p \in \mathbb{R}$ such that

$$\nabla f(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i \nabla g_i(\mathbf{x}^*) + \sum_{j=1}^{p} \mu_j \nabla h_j(\mathbf{x}^*) = 0,$$
$$\lambda_i g_i(\mathbf{x}^*) = 0, \quad i = 1, 2, \ldots, m.$$

We will now add to our terminology two concepts: KKT points and regularity. The first was already discussed in the previous chapter in the context of linearly constrained problems and is now extended.

Definition 11.7 (KKT points). *Consider the minimization problem* (11.9), *where $f, g_1, \ldots, g_m, h_1, h_2, \ldots, h_p$ are continuously differentiable functions over $\mathbb{R}^n$. A feasible point $\mathbf{x}^*$ is called a* **KKT point** *if there exist $\lambda_1, \lambda_2 \ldots, \lambda_m \geq 0$ and $\mu_1, \mu_2, \ldots, \mu_p \in \mathbb{R}$ such that*

$$\nabla f(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i \nabla g_i(\mathbf{x}^*) + \sum_{j=1}^{p} \mu_j \nabla h_j(\mathbf{x}^*) = 0,$$
$$\lambda_i g_i(\mathbf{x}^*) = 0, \quad i = 1, 2, \ldots, m.$$

Definition 11.8 (regularity). *Consider the minimization problem* (11.9), *where $f, g_1, \ldots, g_m, h_1, h_2, \ldots, h_p$ are continuously differentiable functions over $\mathbb{R}^n$. A feasible point $\mathbf{x}^*$ is called* **regular** *if the gradients of the active constraints among the inequality constraints and of the equality constraints*

$$\{\nabla g_i(\mathbf{x}^*) : i \in I(\mathbf{x}^*)\} \cup \{\nabla h_j(\mathbf{x}^*) : j = 1, 2, \ldots, p\}$$

are linearly independent.

In the terminology of the above definitions, Theorem 11.6 states that a necessary optimality condition for local optimality of a regular point is that it is a KKT point. The additional requirement of regularity is not required in the linearly constrained case in which no such assumption is needed; see Theorem 10.7.

Example 11.9. Consider the problem

$$\begin{array}{ll} \min & x_1 + x_2 \\ \text{s.t.} & x_1^2 + x_2^2 = 1. \end{array}$$

Note that this is not a convex optimization problem due to the fact that the equality constraint is nonlinear. In addition, since the problem consists of minimizing a continuous function over a nonempty compact set, it follows that the minimizer exists (by the Weierstrass theorem, Theorem 2.30).

Let us first address the issue of whether the KKT conditions are necessary for this problem. Since by Theorem 11.6 we know that the KKT conditions are necessary optimality conditions for regular points, we will find the irregular points of the problem. These are

exactly the points $\mathbf{x}^*$ for which the set of gradients of active constraints, which is given here by $\{2\binom{x_1^*}{x_2^*}\}$, is a dependent set of vectors; this can happen only when $x_1^* = x_2^* = 0$, that is, at a nonfeasible point. The conclusion is that the problem does not have irregular points, and hence the KKT conditions are necessary optimality conditions.

In order to write the KKT conditions, we will form the Lagrangian:

$$L(x_1, x_2, \lambda) = x_1 + x_2 + \lambda(x_1^2 + x_2^2 - 1).$$

The KKT conditions are

$$\begin{aligned} \frac{\partial L}{\partial x_1} &= 1 + 2\lambda x_1 = 0, \\ \frac{\partial L}{\partial x_2} &= 1 + 2\lambda x_2 = 0, \\ x_1^2 + x_2^2 &= 1. \end{aligned}$$

By the first two conditions, $\lambda \neq 0$, and hence $x_1 = x_2 = -\frac{1}{2\lambda}$. Plugging this expression of x_1, x_2 into the last equation yields

$$\left(-\frac{1}{2\lambda}\right)^2 + \left(-\frac{1}{2\lambda}\right)^2 = 1,$$

so that $\lambda = \pm\frac{1}{\sqrt{2}}$. The problem thus has two KKT points: $(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$ and $(-\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}})$. Since an optimal solution does exist, and the KKT conditions are necessary for this problem, it follows that at least one of these two points is optimal, and obviously the point $(-\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}})$ which has the smaller objective value $(-\sqrt{2})$ is the optimal solution. ■

Example 11.10. Consider a problem which is equivalent to the one given in the previous example:

$$\begin{array}{ll} \min & x_1 + x_2 \\ \text{s.t.} & (x_1^2 + x_2^2 - 1)^2 = 0. \end{array}$$

Writing the KKT conditions yields

$$\begin{aligned} 1 + 4\lambda x_1(x_1^2 + x_2^2 - 1) &= 0, \\ 1 + 4\lambda x_2(x_1^2 + x_2^2 - 1) &= 0, \\ (x_1^2 + x_2^2 - 1)^2 &= 0. \end{aligned}$$

This system is of course infeasible since the combination of the first and third equations gives the impossible equation $1 = 0$. It is not a surprise that the KKT conditions are not satisfied here, since all the feasible points are in fact irregular. Indeed, the gradient of the constraint is $\binom{4x_1(x_1^2+x_2^2-1)}{4x_2(x_1^2+x_2^2-1)}$, which is the zeros vector for any feasible point. ■

Example 11.11. Consider the optimization problem

$$\begin{array}{ll} \min & 2x_1 + 3x_2 - x_3 \\ \text{s.t.} & x_1^2 + x_2^2 + x_3^2 = 1, \\ & x_1^2 + 2x_2^2 + 2x_3^2 = 2. \end{array}$$

The problem does have an optimal solution since it consists of minimizing a continuous function over a nonempty and compact set. The KKT system is

$$\begin{aligned} 2+2(\lambda+\mu)x_1 &= 0, \\ 3+2(\lambda+2\mu)x_2 &= 0, \\ -1+2(\lambda+2\mu)x_3 &= 0, \\ x_1^2+x_2^2+x_3^2 &= 1, \\ x_1^2+2x_2^2+2x_3^2 &= 2. \end{aligned}$$

Obviously, by the first three equations $\lambda+\mu \neq 0, \lambda+2\mu \neq 0$, and in addition

$$x_1 = -\frac{1}{\lambda+\mu}, \quad x_2 = -\frac{3}{2(\lambda+2\mu)}, \quad x_3 = \frac{1}{2(\lambda+2\mu)}.$$

Denoting $t_1 = \frac{1}{\lambda+\mu}, t_2 = \frac{1}{2(\lambda+2\mu)}$, we obtain that $x_1 = -t_1, x_2 = -3t_2, x_3 = t_2$. Substituting these expressions into the constraints we obtain that

$$\begin{aligned} t_1^2+10t_2^2 &= 1, \\ t_1^2+20t_2^2 &= 2. \end{aligned}$$

implying that $t_1^2 = 0$, which is impossible. We obtain that there are no KKT points. Thus, since as was already observed, an optimal solution does exist, it follows that it is an irregular point. To find the irregular points, note that the gradients of the constraints are given by

$$\begin{pmatrix} 2x_1 \\ 2x_2 \\ 2x_3 \end{pmatrix}, \qquad \begin{pmatrix} 2x_1 \\ 4x_2 \\ 4x_3 \end{pmatrix}.$$

The two gradients are linearly dependent in the following two cases. (A) $x_1 = 0$. In this case, the problem becomes

$$\begin{array}{ll} \min & 3x_2 - x_3 \\ \text{s.t.} & x_2^2 + x_3^2 = 1, \end{array}$$

whose optimal solution is $(x_2, x_3) = (-\frac{3}{\sqrt{10}}, \frac{1}{\sqrt{10}})$, and the obtained objective function value is $-\sqrt{10}$.

(B) $x_2 = x_3 = 0$. However, the constraints in this case reduce to $x_1^2 = 1, x_1^2 = 2$, which is impossible.

The conclusion is that the optimal solution of the problem is

$$(x_1, x_2, x_3) = \left(0, -\frac{3}{\sqrt{10}}, \frac{1}{\sqrt{10}}\right)$$

with optimal value $-\sqrt{10}$. ∎

11.3 ▪ The Convex Case

The KKT conditions are necessary optimality condition under the regularity condition. When the problem is convex, the KKT conditions are *always* sufficient and no further conditions are required.

Theorem 11.12 (sufficiency of the KKT conditions for convex optimization problems). *Let* $\mathbf{x}^*$ *be a feasible solution of*

$$\begin{array}{lll} \min & f(\mathbf{x}) & \\ s.t. & g_i(\mathbf{x}) \leq 0, & i = 1,2,\ldots,m, \\ & h_j(\mathbf{x}) = 0, & j = 1,2,\ldots,p, \end{array} \tag{11.10}$$

where $f, g_1, \ldots, g_m$ *are continuously differentiable convex functions over* $\mathbb{R}^n$ *and* $h_1, h_2, \ldots, h_p$ *are affine functions. Suppose that there exist multipliers* $\lambda_1, \lambda_2 \ldots, \lambda_m \geq 0$ *and* $\mu_1, \mu_2, \ldots, \mu_p \in \mathbb{R}$ *such that*

$$\nabla f(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i \nabla g_i(\mathbf{x}^*) + \sum_{j=1}^{p} \mu_j \nabla h_j(\mathbf{x}^*) = 0,$$
$$\lambda_i g_i(\mathbf{x}^*) = 0, \quad i = 1,2,\ldots,m.$$

Then $\mathbf{x}^*$ *is an optimal solution of* (11.10).

Proof. Let $\mathbf{x}$ be a feasible solution of (11.10). We will show that $f(\mathbf{x}) \geq f(\mathbf{x}^*)$. Note that the function

$$s(\mathbf{x}) = f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i g_i(\mathbf{x}) + \sum_{j=1}^{p} \mu_j h_j(\mathbf{x})$$

is convex, and since $\nabla s(\mathbf{x}^*) = \nabla f(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i \nabla g_i(\mathbf{x}^*) + \sum_{j=1}^{p} \mu_j \nabla h_j(\mathbf{x}^*) = 0$, it follows by Proposition 7.8 that $\mathbf{x}^*$ is a minimizer of $s(\cdot)$ over $\mathbb{R}^n$, and in particular $s(\mathbf{x}^*) \leq s(\mathbf{x})$. We can thus conclude that

$$\begin{array}{ll} f(\mathbf{x}^*) & = f(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i g_i(\mathbf{x}^*) + \sum_{j=1}^{p} \mu_j h_j(\mathbf{x}^*) \quad (\lambda_i g_i(\mathbf{x}^*) = 0, h_j(\mathbf{x}^*) = 0) \\ & = s(\mathbf{x}^*) \\ & \leq s(\mathbf{x}) \\ & = f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i g_i(\mathbf{x}) + \sum_{j=1}^{p} \mu_j h_j(\mathbf{x}) \\ & \leq f(\mathbf{x}) \qquad\qquad (\lambda_i \geq 0, g_i(\mathbf{x}) \leq 0, h_j(\mathbf{x}) = 0), \end{array}$$

showing that $\mathbf{x}^*$ is the optimal solution of (11.10). □

In the convex case we can find a different condition than regularity that guarantees the necessity of the KKT condition. This condition is called *Slater's condition*. Slater's condition, like regularity, is a condition on the constraints of the problem. We will say that Slater's condition is satisfied for a set of convex inequalities

$$g_i(\mathbf{x}) \leq 0, \quad i = 1,2,\ldots,m,$$

where $g_1, g_2, \ldots, g_m$ are given convex functions if there exists $\hat{\mathbf{x}} \in \mathbb{R}^n$ such that

$$g_i(\hat{\mathbf{x}}) < 0, \quad i = 1,2,\ldots,m.$$

Note that Slater's condition requires that there exists a point that strictly satisfies the constraints, and does not require, like in the regularity condition, an a priori knowledge on the point that is a candidate to be an optimal solution. This is the reason why checking the validity of Slater's condition is usually a much easier task than checking regularity.

Next, the necessity of the KKT conditions for problems with convex inequalities under Slater's condition is stated and proved.

Theorem 11.13 (necessity of the KKT conditions under Slater's condition). *Let* $\mathbf{x}^*$ *be an optimal solution of the problem*

$$\begin{array}{ll} \min & f(\mathbf{x}) \\ s.t. & g_i(\mathbf{x}) \leq 0, \quad i = 1,2,\ldots,m, \end{array} \tag{11.11}$$

where $f, g_1, \ldots, g_m$ *are continuously differentiable functions over* $\mathbb{R}^n$. *In addition,* $g_1, g_2, \ldots, g_m$ *are convex functions over* $\mathbb{R}^n$. *Suppose that there exists* $\hat{\mathbf{x}} \in \mathbb{R}^n$ *such that*

$$g_i(\hat{\mathbf{x}}) < 0, \quad i = 1,2,\ldots,m.$$

Then there exist multipliers $\lambda_1, \lambda_2 \ldots, \lambda_m \geq 0$ *such that*

$$\nabla f(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i \nabla g_i(\mathbf{x}^*) = 0, \tag{11.12}$$

$$\lambda_i g_i(\mathbf{x}^*) = 0, \quad i = 1,2,\ldots,m. \tag{11.13}$$

Proof. By Theorem 11.4, since $\mathbf{x}^*$ is an optimal solution of (11.11), then the Fritz-John conditions are satisfied. That is, there exist $\tilde{\lambda}_0, \tilde{\lambda}_1, \ldots, \tilde{\lambda}_m \geq 0$, which are not all zeros, such that

$$\begin{aligned} \tilde{\lambda}_0 \nabla f(\mathbf{x}^*) + \sum_{i=1}^{m} \tilde{\lambda}_i \nabla g_i(\mathbf{x}^*) = 0, \\ \tilde{\lambda}_i g_i(\mathbf{x}^*) = 0, \quad i = 1,2,\ldots,m. \end{aligned} \tag{11.14}$$

All that we need to show is that $\tilde{\lambda}_0 > 0$, and then the conditions (11.12) and (11.13) will be satisfied with $\lambda_i = \frac{\tilde{\lambda}_i}{\tilde{\lambda}_0}, i = 1,2,\ldots,m$. To prove that $\tilde{\lambda}_0 > 0$, assume in contradiction that it is zero; then

$$\sum_{i=1}^{m} \tilde{\lambda}_i \nabla g_i(\mathbf{x}^*) = 0. \tag{11.15}$$

By the gradient inequality we have that for all $i = 1,2,\ldots,m$

$$0 > g_i(\hat{\mathbf{x}}) \geq g_i(\mathbf{x}^*) + \nabla g_i(\mathbf{x}^*)^T (\hat{\mathbf{x}} - \mathbf{x}^*).$$

Multiplying the ith inequality by $\tilde{\lambda}_i$ and summing over $i = 1,2,\ldots,m$, we obtain

$$0 > \sum_{i=1}^{m} \tilde{\lambda}_i g_i(\mathbf{x}^*) + \left[\sum_{i=1}^{m} \tilde{\lambda}_i \nabla g_i(\mathbf{x}^*) \right]^T (\hat{\mathbf{x}} - \mathbf{x}^*), \tag{11.16}$$

where the inequality is strict since not all the $\tilde{\lambda}_i$ are zero. Plugging the identities (11.14) and (11.15) into (11.16), we obtain the impossible statement that $0 > 0$, thus establishing the result. □

Example 11.14. Consider the convex optimization problem

$$\begin{array}{ll} \min & x_1^2 - x_2 \\ \text{s.t.} & x_2 = 0. \end{array}$$

The Lagrangian is

$$L(x_1, x_2, \lambda) = x_1^2 - x_2 + \lambda x_2.$$

Since the problem is a linearly constrained convex problem, the KKT conditions are necessary and sufficient (Theorem 10.7). The conditions are

$$\begin{aligned} 2x_1 &= 0, \\ -1 + \lambda &= 0, \\ x_2 &= 0, \end{aligned}$$

and they are satisfied for $(x_1, x_2) = (0, 0)$, which is the optimal solution.

Now, consider a different reformulation of the problem:

$$\begin{array}{ll} \min & x_1^2 - x_2 \\ \text{s.t.} & x_2^2 \leq 0. \end{array}$$

Slater's condition is not satisfied since the constraint cannot be satisfied strictly, and therefore the KKT conditions are not guaranteed to hold at the optimal solution. The KKT conditions in this case are

$$\begin{aligned} 2x_1 &= 0, \\ -1 + 2\lambda x_2 &= 0, \\ \lambda x_2^2 &= 0, \\ x_2^2 &\leq 0, \\ \lambda &\geq 0. \end{aligned}$$

The above system is infeasible since $x_2 = 0$, and hence the equality $-1 + 2\lambda x_2 = 0$ is impossible. ■

A slightly more refined analysis can show that in the presence of affine constraints, one can prove the necessity of the KKT conditions under a generalized Slater's condition which states that there exists a point that strictly satisfies all the nonlinear inequality constraints as well as satisfies the affine equality and inequality constraints.

Definition 11.15 (generalized Slater's condition). *Consider the system*

$$\begin{aligned} g_i(\mathbf{x}) &\leq 0, \quad i = 1, 2, \ldots, m, \\ h_j(\mathbf{x}) &\leq 0, \quad j = 1, 2, \ldots, p, \\ s_k(\mathbf{x}) &= 0, \quad k = 1, 2, \ldots, q, \end{aligned}$$

where $g_i, i = 1, 2, \ldots, m$, are convex functions and $h_j, s_k, j = 1, 2, \ldots, p, k = 1, 2, \ldots, q$, are affine functions. Then we say that the **generalized Slater's condition** *is satisfied if there exists $\hat{\mathbf{x}} \in \mathbb{R}^n$ for which*

$$\begin{aligned} g_i(\hat{\mathbf{x}}) &< 0, \quad i = 1, 2, \ldots, m, \\ h_j(\hat{\mathbf{x}}) &\leq 0, \quad j = 1, 2, \ldots, p, \\ s_k(\hat{\mathbf{x}}) &= 0, \quad k = 1, 2, \ldots, q. \end{aligned}$$

The necessity of the KKT conditions under the generalized Slater's condition is now stated.

Theorem 11.16 (necessity of the KKT conditions under the generalized Slater's condition). *Let* $\mathbf{x}^*$ *be an optimal solution of the problem*

$$\begin{array}{lll} \min & f(\mathbf{x}) & \\ s.t. & g_i(\mathbf{x}) \le 0, & i = 1,2,\dots,m, \\ & h_j(\mathbf{x}) \le 0, & j = 1,2,\dots,p, \\ & s_k(\mathbf{x}) = 0, & k = 1,2,\dots,q, \end{array}$$

where $f, g_1, \dots, g_m$ *are continuously differentiable convex functions over* $\mathbb{R}^n$*, and* $h_j, s_k, j = 1,2,\dots,p, k = 1,2,\dots,q$*, are affine. Suppose that there exists* $\hat{\mathbf{x}} \in \mathbb{R}^n$ *such that*

$$\begin{array}{ll} g_i(\hat{\mathbf{x}}) < 0, & i = 1,2,\dots,m, \\ h_j(\hat{\mathbf{x}}) \le 0, & j = 1,2,\dots,p, \\ s_k(\hat{\mathbf{x}}) = 0, & k = 1,2,\dots,q. \end{array}$$

Then there exist multipliers $\lambda_1, \lambda_2 \dots, \lambda_m, \eta_1, \eta_2, \dots, \eta_p \ge 0, \mu_1, \mu_2, \dots, \mu_q \in \mathbb{R}$ *such that*

$$\nabla f(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i \nabla g_i(\mathbf{x}^*) + \sum_{j=1}^{p} \eta_j \nabla h_j(\mathbf{x}^*) + \sum_{k=1}^{q} \mu_k \nabla s_k(\mathbf{x}^*) = 0,$$

$$\begin{array}{ll} \lambda_i g_i(\mathbf{x}^*) = 0, & i = 1,2,\dots,m, \\ \eta_j h_j(\mathbf{x}^*) = 0, & j = 1,2,\dots,p. \end{array}$$

Example 11.17. Consider the convex optimization problem

$$\begin{array}{ll} \min & 4x_1^2 + x_2^2 - x_1 - 2x_2 \\ \text{s.t.} & 2x_1 + x_2 \le 1, \\ & x_1^2 \le 1. \end{array}$$

Slater's condition is satisfied with $(\hat{x}_1, \hat{x}_2) = (0,0)$ $(2 \cdot 0 + 0 < 1, 0^2 - 1 < 0)$, so the KKT conditions are necessary and sufficient. The Lagrangian function is

$$L(x_1, x_2, \lambda_1, \lambda_2) = 4x_1^2 + x_2^2 - x_1 - 2x_2 + \lambda_1(2x_1 + x_2 - 1) + \lambda_2(x_1^2 - 1),$$

and the KKT system is

$$\frac{\partial L}{\partial x_1} = 8x_1 - 1 + 2\lambda_1 + 2\lambda_2 x_1 = 0, \tag{11.17}$$

$$\frac{\partial L}{\partial x_2} = 2x_2 - 2 + \lambda_1 = 0, \tag{11.18}$$

$$\begin{aligned} \lambda_1(2x_1 + x_2 - 1) &= 0, \\ \lambda_2(x_1^2 - 1) &= 0, \\ 2x_1 + x_2 &\le 1, \\ x_1^2 &\le 1, \\ \lambda_1, \lambda_2 &\ge 0. \end{aligned}$$

We will consider four cases.

Case I: If $\lambda_1 = \lambda_2 = 0$, then by the first two equations, $x_1 = \frac{1}{8}, x_2 = 1$, which is not a feasible solution.

Case II: If $\lambda_1 > 0, \lambda_2 > 0$, then by the complementary slackness conditions

$$\begin{aligned} 2x_1 + x_2 &= 1, \\ x_1^2 &= 1. \end{aligned}$$

The two solutions of this system are $(1,-1),(-1,3)$. Plugging the first solution into the first equation of the KKT system (equation (11.17)) yields

$$7 + 2\lambda_1 + 2\lambda_2 = 0,$$

which is impossible since $\lambda_1, \lambda_2 > 0$. Plugging the second solution into the second equation of the KKT system (equation (11.18)) results in the equation

$$4 + \lambda_1 = 0,$$

which has no solution since $\lambda_1 > 0$.

Case III: If $\lambda_1 > 0, \lambda_2 = 0$, then by the complementary slackness conditions we have that

$$2x_1 + x_2 = 1, \tag{11.19}$$

which combined with (11.17) and (11.18) yields the set of equations (recalling that $\lambda_2 = 0$)

$$\begin{aligned} 8x_1 - 1 + 2\lambda_1 &= 0, \\ 2x_2 - 2 + \lambda_1 &= 0, \\ 2x_1 + x_2 &= 1, \end{aligned}$$

whose unique solution is $(x_1, x_2, \lambda_1) = (\frac{1}{16}, \frac{7}{8}, \frac{1}{4})$, and we obtain that $(x_1, x_2, \lambda_1, \lambda_2) = (\frac{1}{16}, \frac{7}{8}, \frac{1}{4}, 0)$ satisfies the KKT system. Hence, $(x_1, x_2) = (\frac{1}{16}, \frac{7}{8})$ is a KKT point, and since the problem is convex, it is an optimal solution. In principle, we do not have to check the fourth case since we already found an optimal solution, but it might be that there exist additional optimal solutions, and if our objective is to find *all* the optimal solutions, then all the cases should be covered.

Case IV: If $\lambda_1 = 0, \lambda_2 > 0$, then by the complementary slackness conditions we have $x_1^2 = 1$. By (11.18) we have that $x_2 = 1$. The two candidate solutions in this case are therefore $(1,1)$ and $(-1,1)$. The point $(1,1)$ does not satisfy the first constraint of the problem and is therefore infeasible. Plugging $(x_1, x_2) = (-1, 1)$ and $\lambda_1 = 0$ into (11.17) yields

$$-9 - 2\lambda_2 = 0,$$

which is a contradiction to the positivity of λ_2.

To conclude, the unique optimal solution of the problem is $(x_1, x_2) = (\frac{1}{16}, \frac{7}{8})$. ∎

11.4 ▪ Constrained Least Squares

Consider the problem

$$\text{(CLS)} \quad \begin{array}{ll} \min & \|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2 \\ \text{s.t.} & \|\mathbf{x}\|^2 \le \alpha, \end{array}$$

where $\mathbf{A} \in \mathbb{R}^{m \times n}$ is assumed to be of full column rank, $\mathbf{b} \in \mathbb{R}^m$, and $\alpha > 0$. We will refer to this problem as a *constrained least squares (CLS)* problem. In Section 3.3 we considered

the regularized least squares (RLS) problem which has the form[5] $\min\{\|\mathbf{A}\mathbf{x}-\mathbf{b}\|^2+\mu\|\mathbf{x}\|^2\}$. The two problems are related in the sense that they both regularize the least squares solution by a quadratic regularization term. In the RLS problem, the regularization is done by a penalty function, while in the CLS problem the regularization is performed by incorporating it as a constraint.

Problem (CLS) is a convex problem and satisfies Slater's condition since $\hat{\mathbf{x}}=0$ strictly satisfies the constraint of the problem. To solve the problem, we begin by forming the Lagrangian:

$$L(\mathbf{x},\lambda)=\|\mathbf{A}\mathbf{x}-\mathbf{b}\|^2+\lambda(\|\mathbf{x}\|^2-\alpha) \quad (\lambda\geq 0).$$

The KKT conditions are

$$\begin{aligned}\nabla_{\mathbf{x}}L=2\mathbf{A}^T(\mathbf{A}\mathbf{x}-\mathbf{b})+2\lambda\mathbf{x}&=0,\\ \lambda(\|\mathbf{x}\|^2-\alpha)&=0,\\ \|\mathbf{x}\|^2&\leq\alpha,\\ \lambda&\geq 0.\end{aligned}$$

There are two options. In the first, $\lambda=0$, and then by the first equation we have that

$$\mathbf{x}=\mathbf{x}_{\mathrm{LS}}\equiv(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{b}.$$

This is a KKT point and hence the optimal solution if and only if $\mathbf{x}_{\mathrm{LS}}$ is feasible, that is, if $\|\mathbf{x}_{\mathrm{LS}}\|^2\leq\alpha$. This is not a surprising result since it is clear that when the unconstrained minimizer ($\mathbf{x}_{\mathrm{LS}}$) satisfies the constraint, it is also the optimal solution of the constrained problem.

On the other hand, if $\|\mathbf{x}_{\mathrm{LS}}\|^2>\alpha$, then $\lambda>0$. By the complementary slackness condition we have that $\|\mathbf{x}\|^2=\alpha$, and the first equation implies that

$$\mathbf{x}=\mathbf{x}_\lambda\equiv(\mathbf{A}^T\mathbf{A}+\lambda\mathbf{I})^{-1}\mathbf{A}^T\mathbf{b}.$$

The multiplier $\lambda>0$ should be thus chosen to satisfy $\|\mathbf{x}_\lambda\|^2=\alpha$; that is, λ is the solution of

$$f(\lambda)\equiv\|(\mathbf{A}^T\mathbf{A}+\lambda\mathbf{I})^{-1}\mathbf{A}^T\mathbf{b}\|^2-\alpha=0. \tag{11.20}$$

We have $f(0)=\|(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{b}\|^2-\alpha=\|\mathbf{x}_{\mathrm{LS}}\|^2-\alpha>0$, and it is not difficult to show that f is a strictly decreasing function satisfying $f(\lambda)\to-\alpha$ as $\lambda\to\infty$. Thus, there exists a unique λ for which $f(\lambda)=0$, and this λ can be found for example by a simple bisection procedure. To conclude, the optimal solution of the CLS problem is given by

$$\mathbf{x}=\begin{cases}\mathbf{x}_{\mathrm{LS}}, & \|\mathbf{x}_{\mathrm{LS}}\|^2\leq\alpha,\\ (\mathbf{A}^T\mathbf{A}+\lambda\mathbf{I})^{-1}\mathbf{A}^T\mathbf{b} & \|\mathbf{x}_{\mathrm{LS}}\|^2>\alpha,\end{cases}$$

where λ is the unique root of f over $[0,\infty)$. We will now construct a MATLAB function for solving the CLS problem. For that, we need to write a MATLAB function that performs a bisection algorithm. The bisection method for finding the root of a scalar equation $f(x)=0$ is described below.

[5] In Section 3.3 we actually considered a more general model in which the regularizer had the form $\mu\|\mathbf{D}\mathbf{x}\|^2$, where $\mathbf{D}$ is a given matrix.

Bisection

Input: $\varepsilon > 0$ - tolerance parameter. $a < b$ - two numbers satisfying $f(a)f(b) < 0$.

Initialization: Take $l_0 = a, u_0 = b$.
General step: For any $k = 0, 1, 2, \ldots$ execute the following steps:

(a) Take $x_k = \frac{u_k + l_k}{2}$.

(b) If $f(l_k) \cdot f(x_k) > 0$, define $l_{k+1} = x_k, u_{k+1} = u_k$. Otherwise, define $l_{k+1} = l_k, u_{k+1} = x_k$.

(c) if $u_{k+1} - l_{k+1} \leq \varepsilon$, then STOP, and x_k is the output.

A MATLAB function implementing the bisection method is given below.

```
function z=bisection(f,lb,ub,eps)
%INPUT
%================
%f ................... a scalar function
%lb .................. the initial lower bound
%ub .................. the initial upper bound
%eps ................. tolerance parameter
%OUTPUT
%================
% z .................. a root of the equation f(x)=0

if (f(lb)*f(ub)>0)
    error('f(lb)*f(ub)>0')
end

iter=0;
while (ub-lb>eps)
    z=(lb+ub)/2;
    iter=iter+1;
    if(f(lb)*f(z)>0)
        lb=z;
    else
        ub=z;
    end
    fprintf('iter_numer = %3d current_sol = %2.6f \n',iter,z);
end
```

Therefore, for example, if we wish to find the square root of 2 with an accuracy of 10^{-4}, then we can solve the equation $x^2 - 2 = 0$ by the following MATLAB command:

```
>> bisection(@(x)x^2-2,1,2,1e-4);
iter_numer =   1 current_sol = 1.500000
iter_numer =   2 current_sol = 1.250000
iter_numer =   3 current_sol = 1.375000
iter_numer =   4 current_sol = 1.437500
```

```
iter_numer =   5 current_sol = 1.406250
iter_numer =   6 current_sol = 1.421875
iter_numer =   7 current_sol = 1.414063
iter_numer =   8 current_sol = 1.417969
iter_numer =   9 current_sol = 1.416016
iter_numer =  10 current_sol = 1.415039
iter_numer =  11 current_sol = 1.414551
iter_numer =  12 current_sol = 1.414307
iter_numer =  13 current_sol = 1.414185
iter_numer =  14 current_sol = 1.414246
```

As for the CLS problem, note that the scalar function f given in (11.20) satisfies $f(0) > 0$. Therefore, all that is left is to find a point $u > 0$ satisfying $f(u) < 0$. For that, we will start with guessing $u = 1$ and then make the update $u \leftarrow 2u$ until $f(u) < 0$. The MATLAB function implementing these ideas is given below.

```
function x_cls=cls(A,b,alpha)
%INPUT
%================
%A ................... an mxn matrix
%b ................... an m-length vector
%alpha ............... positive scalar
%OUTPUT
%================
% x_cls ............ an optimal solution of
%                    min{||A*x-b||:||x||^2<=alpha}
d=size(A);
n=d(2);
x_ls=A\b;
if (norm(x_ls)^2<=alpha)
    x_cls=x_ls;
else
    f=@(lam) norm((A'*A+lam*eye(n))\(A'*b))^2-alpha;
    u=1;
    while (f(u)>0)
        u=2*u;
    end
    lam=bisection(f,0,u,1e-7);
    x_cls=(A'*A+lam*eye(n))\(A'*b);
end
```

For example, assume that we pick **A** and **b** as

```
A=[1,2;3,1;2,3];
b=[2;3;4];
```

The least squares solution and its squared norm can be easily found:

```
>> x_ls=A\b
x_ls =

    0.7600
    0.7600
```

```
>> norm(x_ls)^2

ans =

    1.1552
```

If we use the `cls` function with an α which is greater than 1.1552, then we will obviously get back the least squares solution:

```
>> cls(A,b,1.5)
ans =

    0.7600
    0.7600
```

On the other hand, taking an α with a smaller value than 1.552, will result in a different solution (the bisection output was suppressed):

```
>> cls(A,b,0.5)
ans =

    0.5000
    0.5000
```

To double check the result, we can run CVX,

```
cvx_begin
variable x_cvx(2)
minimize(norm(A*x_cvx-b))
norm(x_cvx)<=sqrt(0.5)
cvx_end
```

and get the same result:

```
>> x_cvx
x_cvx =

    0.5000
    0.5000
```

11.5 ▪ Second Order Optimality Conditions

11.5.1 ▪ Necessary Conditions for Inequality Constrained Problems

We can also establish necessary second order optimality conditions in the general non-convex case. We will begin by stating and proving the result for inequality constrained problem.

Theorem 11.18 (second order necessary conditions for inequality constrained problems). *Consider the problem*

$$\begin{array}{ll} \min & f_0(\mathbf{x}) \\ s.t. & f_i(\mathbf{x}) \leq 0, \quad i = 1,2,\ldots,m, \end{array} \tag{11.21}$$

where $f_0, f_1, \ldots, f_m$ are twice continuously differentiable over $\mathbb{R}^n$. Let $\mathbf{x}^*$ be a local minimum of problem (11.21)*, and suppose that $\mathbf{x}^*$ is regular, meaning that the set $\{\nabla f_i(\mathbf{x}^*)\}_{i \in I(\mathbf{x}^*)}$ is linearly independent, where*

$$I(\mathbf{x}^*) = \{i \in \{1,2,\ldots,m\} : f_i(\mathbf{x}^*) = 0\}.$$

Denote the Lagrangian by

$$L(\mathbf{x}, \lambda) = f_0(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i f_i(\mathbf{x}).$$

Then there exist $\lambda_1, \lambda_2, \ldots, \lambda_m \geq 0$ such that

$$\begin{aligned} \nabla_{\mathbf{x}} L(\mathbf{x}^*, \lambda) &= 0, \\ \lambda_i f_i(\mathbf{x}^*) &= 0, \quad i = 1,2,\ldots,m, \end{aligned}$$

and

$$\mathbf{y}^T \nabla^2_{\mathbf{xx}} L(\mathbf{x}^*, \lambda)\mathbf{y} = \mathbf{y}^T \left[\nabla^2 f_0(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i \nabla^2 f_i(\mathbf{x}^*) \right] \mathbf{y} \geq 0$$

for all $\mathbf{y} \in \Lambda(\mathbf{x}^*)$, where

$$\Lambda(\mathbf{x}^*) \equiv \{\mathbf{d} \in \mathbb{R}^n : \nabla f_i(\mathbf{x}^*)^T \mathbf{d} = 0, i \in I(\mathbf{x}^*)\}.$$

Proof. Let $\mathbf{d} \in D(\mathbf{x}^*)$, where

$$D(\mathbf{x}^*) = \{\mathbf{d} \in \mathbb{R}^n : \nabla f_i(\mathbf{x}^*)^T \mathbf{d} \leq 0, i \in I(\mathbf{x}^*) \cup \{0\}\}.$$

From this point until further notice we will assume that $\mathbf{d}$ is fixed. Let $\mathbf{z} \in \mathbb{R}^n$, and $i \in \{0,1,2,\ldots,m\}$. Define

$$\mathbf{x}(t) \equiv \mathbf{x}^* + t\mathbf{d} + \frac{t^2}{2}\mathbf{z}$$

and the one-dimensional functions

$$g_i(t) \equiv f_i(\mathbf{x}(t)), \quad i \in I(\mathbf{x}^*) \cup \{0\}.$$

Then

$$\begin{aligned} g_i'(t) &= (\mathbf{d} + t\mathbf{z})^T \nabla f_i(\mathbf{x}(t)), \\ g_i''(t) &= (\mathbf{d} + t\mathbf{z})^T \nabla^2 f_i(\mathbf{x}(t))(\mathbf{d} + t\mathbf{z}) + \mathbf{z}^T \nabla f_i(\mathbf{x}(t)), \end{aligned}$$

which in particular implies that

$$\begin{aligned} g_i'(0) &= \nabla f_i(\mathbf{x}^*)^T \mathbf{d}, \\ g_i''(0) &= \mathbf{d}^T \nabla^2 f_i(\mathbf{x}^*)\mathbf{d} + \nabla f_i(\mathbf{x}^*)^T \mathbf{z}. \end{aligned}$$

By the quadratic approximation theorem (Theorem 1.25) we have

$$g_i(t) = f_i(\mathbf{x}^*) + (\nabla f_i(\mathbf{x}^*)^T \mathbf{d})t + \left(\mathbf{d}^T \nabla^2 f_i(\mathbf{x}^*)\mathbf{d} + \nabla f_i(\mathbf{x}^*)^T \mathbf{z}\right) \frac{t^2}{2} + o(t^2). \tag{11.22}$$

Therefore, for any $i \in I(\mathbf{x}^*) \cup \{0\}$ there are two cases:

1. $\nabla f_i(\mathbf{x}^*)^T \mathbf{d} < 0$, and in this case $f_i(\mathbf{x}(t)) < f_i(\mathbf{x}^*)$ for small enough $t > 0$.
2. $\nabla f_i(\mathbf{x}^*)^T \mathbf{d} = 0$. In this case, by (11.22), if $\nabla f_i(\mathbf{x}^*)^T \mathbf{z} + \mathbf{d}^T \nabla^2 f_i(\mathbf{x}^*)\mathbf{d} < 0$, then $f_i(\mathbf{x}(t)) < f_i(\mathbf{x}^*)$ for small enough t.

As a conclusion, since $\mathbf{x}^*$ is a local minimum of problem (11.21), the following system of strict inequalities in $\mathbf{z}$ (recall that $\mathbf{d}$ is fixed) does not have a solution:

$$\nabla f_i(\mathbf{x}^*)^T\mathbf{z}+\mathbf{d}^T\nabla^2 f_i(\mathbf{x}^*)\mathbf{d}<0, \quad i\in J(\mathbf{x}^*)\cup\{0\}, \tag{11.23}$$

where

$$J(\mathbf{x}^*)=\{i\in I(\mathbf{x}^*):\nabla f_i(\mathbf{x}^*)^T\mathbf{d}=0\}.$$

Indeed, if there was a solution to system (11.23), then for small enough t, the vector $\mathbf{x}(t)$ would be a feasible solution satisfying $f_0(\mathbf{x}(t))<f_0(\mathbf{x}^*)$, contradicting the local optimality of $\mathbf{x}^*$. System (11.23) can be written as

$$\mathbf{A}\mathbf{z}<\mathbf{b},$$

where $\mathbf{A}$ is the matrix whose components are $\nabla f_i(\mathbf{x}^*)^T, i\in J(\mathbf{x}^*)\cup\{0\}$, and $\mathbf{b}$ is the vector whose components are $-\mathbf{d}^T\nabla^2 f_i(\mathbf{x}^*)\mathbf{d},\ i\in J(\mathbf{x}^*)\cup\{0\}$. By the nonhomogenous Gordan's theorem (see Exercise 10.5), we have that there exists $\mathbf{y}$ such that $\mathbf{A}^T\mathbf{y}=0, \mathbf{b}^T\mathbf{y}\le 0, \mathbf{y}\ge 0, \mathbf{y}\ne 0$, meaning that there exist $0\le y_i, i\in J(\mathbf{x}^*)\cup\{0\}$, not all zeros, such that

$$\sum_{i\in J(\mathbf{x}^*)\cup\{0\}} y_i\nabla f_i(\mathbf{x}^*)=0 \tag{11.24}$$

and

$$\sum_{i\in J(\mathbf{x}^*)\cup\{0\}} y_i(-\mathbf{d}^T\nabla^2 f_i(\mathbf{x}^*)\mathbf{d})\le 0,$$

that is,

$$\mathbf{d}^T\left[\sum_{i\in J(\mathbf{x}^*)\cup\{0\}} y_i\nabla^2 f_i(\mathbf{x}^*)\right]\mathbf{d}\ge 0.$$

By the regularity of $\mathbf{x}^*$ and (11.24), we have that $y_0>0$, and hence by defining $\lambda_i=\frac{y_i}{y_0}$ for $i\in J(\mathbf{x}^*)$ and $\lambda_i=0$ for $i\in I(\mathbf{x}^*)\backslash J(\mathbf{x}^*)$, we obtain that

$$\nabla f_0(\mathbf{x}^*)+\sum_{i\in I(\mathbf{x}^*)}\lambda_i\nabla f_i(\mathbf{x}^*)=0, \tag{11.25}$$

$$\mathbf{d}^T\left[\nabla^2 f_0(\mathbf{x}^*)+\sum_{i\in I(\mathbf{x}^*)}\lambda_i\nabla^2 f_i(\mathbf{x}^*)\right]\mathbf{d}\ge 0.$$

Since $\{\nabla f_i(\mathbf{x}^*)\}_{i\in I(\mathbf{x}^*)}$ are linearly independent, equation (11.25) implies that the multipliers $\lambda_i, i\in I(\mathbf{x}^*)$, do not depend on the initial choice of $\mathbf{d}$. Therefore, by defining $\lambda_i=0$ for any $i\notin I(\mathbf{x}^*)$, we obtain that

$$\begin{aligned}\nabla f_0(\mathbf{x}^*)+\sum_{i=1}^m\lambda_i\nabla f_i(\mathbf{x}^*)&=0,\\ \lambda_i f_i(\mathbf{x}^*)&=0, \quad i=1,2,\ldots,m,\\ \mathbf{d}^T\left[\nabla^2 f_0(\mathbf{x}^*)+\sum_{i=1}^m\lambda_i\nabla^2 f_i(\mathbf{x}^*)\right]\mathbf{d}&\ge 0\end{aligned} \tag{11.26}$$

for any $\mathbf{d}\in D(\mathbf{x}^*)$. All that is left to prove is that (11.26) is satisfied for all $\mathbf{d}\in\Lambda(\mathbf{x}^*)$. Indeed, if $\mathbf{d}\in\Lambda(\mathbf{x}^*)$, then either $\mathbf{d}$ or $-\mathbf{d}$ is in $D(\mathbf{x}^*)$. Thus, $c\mathbf{d}\in D(\mathbf{x}^*)$ for some $c\in\{-1,1\}$, and as a result

$$(c\mathbf{d})^T\left[\nabla^2 f_0(\mathbf{x}^*)+\sum_{i=1}^m\lambda_i\nabla^2 f_i(\mathbf{x}^*)\right](c\mathbf{d})\ge 0,$$

which is the same as

$$\mathbf{d}^T\left[\nabla^2 f_0(\mathbf{x}^*)+\sum_{i=1}^m \lambda_i \nabla^2 f_i(\mathbf{x}^*)\right]\mathbf{d}\geq 0,$$

and the result is established. □

11.5.2 ▪ Necessary Second Order Conditions for Equality and Inequality Constrained Problems

When the problem involves both equality and inequality constraints, a similar result can be proved, and it is stated without a proof below.

Theorem 11.19 (second order necessary conditions for equality and inequality constrained problems). *Consider the problem*

$$\begin{array}{lll} \min & f(\mathbf{x}) & \\ s.t. & g_i(\mathbf{x})\leq 0, & i=1,2,\ldots,m, \\ & h_j(\mathbf{x})=0, & j=1,2,\ldots,p, \end{array} \tag{11.27}$$

where $f, g_1,\ldots,g_m,h_1,\ldots,h_p$ are twice continuously differentiable over $\mathbb{R}^n$. Let $\mathbf{x}^*$ be a local minimum of problem (11.27)*, and suppose that $\mathbf{x}^*$ is regular, meaning that $\{\nabla g_i(\mathbf{x}^*), \nabla h_j(\mathbf{x}^*), i\in I(\mathbf{x}^*), j=1,2,\ldots,p\}$ are linearly independent, where*

$$I(\mathbf{x}^*)=\{i\in\{1,2,\ldots,m\}: g_i(\mathbf{x}^*)=0\}.$$

Then there exist $\lambda_1,\lambda_2,\ldots,\lambda_m\geq 0$ and $\mu_1,\mu_2,\ldots,\mu_p\in\mathbb{R}$ such that

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

$$\lambda_i g_i(\mathbf{x}^*)=0,\quad i=1,2,\ldots,m,$$

and

$$\mathbf{d}^T\left[\nabla^2 f_0(\mathbf{x}^*)+\sum_{i=1}^m \lambda_i\nabla^2 g_i(\mathbf{x}^*)+\sum_{j=1}^p \mu_j\nabla^2 h_j(\mathbf{x}^*)\right]\mathbf{d}\geq 0$$

for all $\mathbf{d}\in\Lambda(\mathbf{x}^*)$ where

$$\Lambda(\mathbf{x}^*)\equiv\{\mathbf{d}\in\mathbb{R}^n: \nabla g_i(\mathbf{x}^*)^T\mathbf{d}=0, \nabla h_j(\mathbf{x}^*)^T\mathbf{d}=0, i\in I(\mathbf{x}^*), j=1,2,\ldots,p\}.$$

Example 11.20. Consider the problem

$$\begin{array}{ll} \min & (2x_1-1)^2+x_2^2 \\ \text{s.t.} & h(x_1,x_2)\equiv -2x_1+x_2^2=0. \end{array}$$

We first note that since the problem consists of minimizing a coercive objective function over a closed set, it follows that an optimal solution does exist. In addition, there are no irregular points to the problem since the gradient of the constraint function h is always different from the zeros vector. Therefore, the optimal solution is one of the KKT points of the problem. The Lagrangian of the problem is

$$L(x_1,x_2,\mu)=(2x_1-1)^2+x_2^2+\mu(-2x_1+x_2^2),$$

and the KKT system is

$$\frac{\partial L}{\partial x_1} = 4(2x_1 - 1) - 2\mu = 0,$$
$$\frac{\partial L}{\partial x_2} = 2x_2 + 2\mu x_2 = 0,$$
$$-2x_1 + x_2^2 = 0.$$

By the second equation

$$2(1+\mu)x_2 = 0,$$

and hence there are two cases. In one case, $x_2 = 0$; then by the third equation, $x_1 = 0$, and by the first equation $\mu = -2$. The second case is when $\mu = -1$, and then by the first equation, $4(2x_1 - 1) = -2$, that is, $x_1 = \frac{1}{4}$, and then $x_2 = \pm\frac{1}{\sqrt{2}}$. We thus obtain that there are three KKT points: $(x_1, x_2, \mu) = (0, 0, -2), (\frac{1}{4}, \frac{1}{\sqrt{2}}, -1), (\frac{1}{4}, -\frac{1}{\sqrt{2}}, -1)$.

Note that

$$\nabla^2_{\mathbf{xx}} L(x_1, x_2, \mu) = \begin{pmatrix} 8 & 0 \\ 0 & 2(1+\mu) \end{pmatrix}.$$

Note that for the points $(x_1, x_2, \mu) = (\frac{1}{4}, \frac{1}{\sqrt{2}}, -1), (\frac{1}{4}, -\frac{1}{\sqrt{2}}, -1)$ the Hessian of the Lagrangian is positive semidefinite:

$$\nabla^2_{\mathbf{xx}} L(x_1, x_2, \mu) = \begin{pmatrix} 8 & 0 \\ 0 & 0 \end{pmatrix} \succeq 0.$$

Therefore, these points satisfy the second order necessary conditions. On the other hand, for the first point $(0, 0)$ where $\mu = -2$, the Hessian is given by

$$\nabla^2_{\mathbf{xx}} L(x_1, x_2, \mu) = \begin{pmatrix} 8 & 0 \\ 0 & -2 \end{pmatrix},$$

which is not a positive semidefinite matrix. To check the validity of the second order conditions at (0,0), note that

$$\nabla h(x_1, x_2) = \begin{pmatrix} -2 \\ 2x_2 \end{pmatrix},$$

and thus $\nabla h(0, 0) = \binom{-2}{0}$. We therefore need to check whether

$$\mathbf{d}^T \nabla^2_{\mathbf{xx}} L(x_1, x_2, \mu)\mathbf{d} \geq 0 \text{ for all } \mathbf{d} \text{ satisfying } \nabla h(0,0)^T \mathbf{d} = 0.$$

Since the condition $\nabla h(0,0)^T \mathbf{d} = 0$ translates to $d_1 = 0$, we need to check whether

$$\begin{pmatrix} 0 & d_2 \end{pmatrix} \nabla^2_{\mathbf{xx}} L(x_1, x_2, \mu) \begin{pmatrix} 0 \\ d_2 \end{pmatrix} \geq 0$$

for any d_2. However, the latter inequality is equivalent to saying that $-2d_2^2 \geq 0$ for any d_2, which is of course not correct.

The conclusion is that $(0, 0)$ does not satisfy the second order necessary conditions and hence cannot be an optimal solution. The optimal solution must be either $(\frac{1}{4}, \frac{1}{\sqrt{2}})$ or $(\frac{1}{4}, -\frac{1}{\sqrt{2}})$ (or both). Since the points have the same objective function value, it follows that they are the optimal solutions of the problem. ∎

11.6 ▪ Optimality Conditions for the Trust Region Subproblem

In Section 8.2.7 we considered the trust region subproblem (TRS) in which one minimizes a (possibly) nonconvex quadratic function subject to a norm constraint:

$$\text{(TRS):} \quad \min\{f(\mathbf{x}) \equiv \mathbf{x}^T\mathbf{A}\mathbf{x} + 2\mathbf{b}^T\mathbf{x} + c : \|\mathbf{x}\|^2 \le \alpha\},$$

where $\mathbf{A} = \mathbf{A}^T \in \mathbb{R}^{n\times n}, \mathbf{b} \in \mathbb{R}^n, c \in \mathbb{R}$, and $\alpha \in \mathbb{R}_{++}$. Note that here we consider a slight extension of the model given in Section 8.2.7 since the norm constraint has a general upper bound and not 1. We have seen that the problem can be recast as a convex optimization problem. In this section we will look at another aspect of the "easiness" of the problem: the problem possesses necessary and sufficient optimality conditions. We will show that these optimality conditions can be used to develop an algorithm for solving the problem. We begin by stating the necessary and sufficient conditions.

Theorem 11.21 (necessary and sufficient conditions for problem (TRS)). *A vector* $\mathbf{x}^*$ *is an optimal solution of problem* (TRS) *if and only if there exists* $\lambda^* \ge 0$ *such that*

$$(\mathbf{A} + \lambda^*\mathbf{I})\mathbf{x}^* = -\mathbf{b}, \tag{11.28}$$

$$\|\mathbf{x}^*\|^2 \le \alpha, \tag{11.29}$$

$$\lambda^*(\|\mathbf{x}^*\|^2 - \alpha) = 0, \tag{11.30}$$

$$\mathbf{A} + \lambda^*\mathbf{I} \succeq 0. \tag{11.31}$$

Proof. **Sufficiency:** To prove the sufficiency, let us assume that $\mathbf{x}^*$ satisfies (11.28)–(11.31) for some $\lambda^* \ge 0$. Define the function

$$h(\mathbf{x}) = f(\mathbf{x}) + \lambda^*(\|\mathbf{x}\|^2 - \alpha) = \mathbf{x}^T(\mathbf{A} + \lambda^*\mathbf{I})\mathbf{x} + 2\mathbf{b}^T\mathbf{x} + c - \alpha\lambda^*. \tag{11.32}$$

Then by (11.31) h is a convex quadratic function. By (11.28) it follows that $\nabla h(\mathbf{x}^*) = 0$, which combined with the convexity of h implies that $\mathbf{x}^*$ is an unconstrained minimizer of h over $\mathbb{R}^n$ (see Proposition 7.8). Let $\mathbf{x}$ be a feasible point, that is, $\|\mathbf{x}\|^2 \le \alpha$. Then

$$\begin{aligned}
f(\mathbf{x}) &\ge f(\mathbf{x}) + \lambda^*(\|\mathbf{x}\|^2 - \alpha) && (\lambda^* \ge 0, \|\mathbf{x}\|^2 - \alpha \le 0)\\
&= h(\mathbf{x}) && \text{(by (11.32))}\\
&\ge h(\mathbf{x}^*) && (\mathbf{x}^* \text{ is a minimizer of } h)\\
&= f(\mathbf{x}^*) + \lambda^*(\|\mathbf{x}^*\|^2 - \alpha) &&\\
&= f(\mathbf{x}^*) && \text{(by (11.30))}
\end{aligned}$$

and we have established that $\mathbf{x}^*$ is a global optimal solution of the problem.

Necessity: To prove the necessity, note that the second order necessary optimality conditions are satisfied since all the feasible points of (TRS) are regular. Indeed, the regularity condition states that when the constraint is active, that is, when $\|\mathbf{x}^*\|^2 = \alpha$, the gradient of the constraint is not the zeros vector, and indeed, since the gradient in this case is $2\mathbf{x}^*$, it is not equal to the zeros vector (since $\|\mathbf{x}^*\|^2 = \alpha$).

If $\|\mathbf{x}^*\| < \alpha$, then the second order necessary optimality conditions (see Theorem 11.18) are exactly (11.28)–(11.31). If $\|\mathbf{x}^*\|^2 = \alpha$, then by the second order necessary optimality conditions there exists $\lambda^* \ge 0$ such that

$$(\mathbf{A} + \lambda^*\mathbf{I})\mathbf{x}^* = -\mathbf{b} \tag{11.33}$$

$$\|\mathbf{x}^*\|^2 \le \alpha, \tag{11.34}$$

$$\lambda^*(\|\mathbf{x}^*\|^2 - \alpha) = 0, \tag{11.35}$$

$$\mathbf{d}^T(\mathbf{A} + \lambda^*\mathbf{I})\mathbf{d} \ge 0 \quad \text{for all } \mathbf{d} \text{ satisfying } \mathbf{d}^T\mathbf{x}^* = 0. \tag{11.36}$$

All that is left to show is that the inequality (11.36) is true for any $\mathbf{d}$ and not only for those which are orthogonal to $\mathbf{x}^*$. Suppose on the contrary that there exists a $\mathbf{d}$ such that $\mathbf{d}^T\mathbf{x}^* \neq 0$ and $\mathbf{d}^T(\mathbf{A}+\lambda^*\mathbf{I})\mathbf{d} < 0$. Consider the point $\bar{\mathbf{x}} = \mathbf{x}^* + t\mathbf{d}$, where $t = -2\frac{\mathbf{d}^T\mathbf{x}^*}{\|\mathbf{d}\|^2}$. The vector $\bar{\mathbf{x}}$ is a feasible point since

$$\begin{aligned}\|\bar{\mathbf{x}}\|^2 &= \|\mathbf{x}^* + t\mathbf{d}\|^2 = \|\mathbf{x}^*\|^2 + 2t\mathbf{d}^T\mathbf{x}^* + t^2\|\mathbf{d}\|^2 \\ &= \|\mathbf{x}^*\|^2 - 4\frac{(\mathbf{d}^T\mathbf{x}^*)^2}{\|\mathbf{d}\|^2} + 4\frac{(\mathbf{d}^T\mathbf{x}^*)^2}{\|\mathbf{d}\|^2} \\ &= \|\mathbf{x}^*\|^2 \leq \alpha.\end{aligned}$$

In addition,

$$\begin{aligned}f(\bar{\mathbf{x}}) &= \bar{\mathbf{x}}^T\mathbf{A}\bar{\mathbf{x}} + 2\mathbf{b}^T\bar{\mathbf{x}} + c \\ &= (\mathbf{x}^* + t\mathbf{d})^T\mathbf{A}(\mathbf{x}^* + t\mathbf{d}) + 2\mathbf{b}^T(\mathbf{x}^* + t\mathbf{d}) + c \\ &= \underbrace{(\mathbf{x}^*)^T\mathbf{A}\mathbf{x}^* + 2\mathbf{b}^T\mathbf{x}^* + c}_{f(\mathbf{x}^*)} + t^2\mathbf{d}^T\mathbf{A}\mathbf{d} + 2t\mathbf{d}^T(\mathbf{A}\mathbf{x}^* + \mathbf{b}) \\ &= f(\mathbf{x}^*) + t^2\mathbf{d}^T(\mathbf{A}+\lambda^*\mathbf{I})\mathbf{d} + 2t\mathbf{d}^T(\underbrace{(\mathbf{A}+\lambda^*\mathbf{I})\mathbf{x}^* + \mathbf{b}}_{=0 \text{ by } (11.33)}) - \lambda^* t\underbrace{\left[t\|\mathbf{d}\|^2 + 2\mathbf{d}^T\mathbf{x}^*\right]}_{=0 \text{ by def. of } t} \\ &= f(\mathbf{x}^*) + t^2\mathbf{d}^T(\mathbf{A}+\lambda^*\mathbf{I})\mathbf{d} \\ &< f(\mathbf{x}^*),\end{aligned}$$

which is a contradiction to the optimality of $\mathbf{x}^*$. $\square$

Theorem 11.21 can be used in order to construct an algorithm for solving the trust region subproblem. We will make the following assumption, which is rather conventional in the literature:

$$-\mathbf{b} \notin \text{Range}(\mathbf{A} - \lambda_{\min}(\mathbf{A})\mathbf{I}). \tag{11.37}$$

Under this condition there cannot be a vector $\mathbf{x}$ for which $(\mathbf{A} - \lambda_{\min}(\mathbf{A})\mathbf{I})\mathbf{x} = -\mathbf{b}$. This means that the multiplier λ^* from the optimality conditions must be different than $-\lambda_{\min}(\mathbf{A})$. The condition (11.37) is considered to be rather mild in the sense that the range space of the matrix $\mathbf{A} - \lambda_{\min}(\mathbf{A})\mathbf{I}$ is of rank which is at most $n-1$. Therefore, at least when $\mathbf{A}$ and $\mathbf{b}$ are generated from a continuous random distribution, the probability that $-\mathbf{b}$ will *not* be in this space is 1.

We will consider two cases.

Case I: $\mathbf{A} \succ 0$. Since in this case the problem is convex, $\mathbf{x}^*$ is an optimal solution of (TRS) if and only if there exists $\lambda^* \geq 0$ such that

$$(\mathbf{A}+\lambda^*\mathbf{I})\mathbf{x}^* = -\mathbf{b}, \quad \lambda^*(\|\mathbf{x}^*\|^2 - \alpha) = 0, \quad \|\mathbf{x}^*\|^2 \leq \alpha.$$

If $\lambda^* = 0$, then $\mathbf{A}\mathbf{x}^* = -\mathbf{b}$, and hence $\mathbf{x}^* = -\mathbf{A}^{-1}\mathbf{b}$. This will be the optimal solution if and only if $\|\mathbf{A}^{-1}\mathbf{b}\|^2 \leq \alpha$. If $\lambda^* > 0$, then $\|\mathbf{x}^*\|^2 = \alpha$, and thus the optimal solution is given by $\mathbf{x}^* = -(\mathbf{A}+\lambda^*\mathbf{I})^{-1}\mathbf{b}$, where λ^* is the unique root of the strictly decreasing function

$$f(\lambda) = \|(\mathbf{A}+\lambda\mathbf{I})^{-1}\mathbf{b}\|^2 - \alpha$$

over $(0, \infty)$.

Case II: $\mathbf{A} \not\succ 0$ In this case, condition (11.31) combined with the nonnegativity of λ^* is the same as $\lambda^* \geq -\lambda_{\min}(\mathbf{A})(\geq 0)$. Under Assumption (11.37), λ^* cannot be equal to

$-\lambda_{\min}(\mathbf{A})$, and we can thus assume that $\lambda^* > -\lambda_{\min}(\mathbf{A})$. In particular, $\lambda^* > 0$ and hence we have $||\mathbf{x}^*||^2 = \alpha$ as well as $\mathbf{A} + \lambda^*\mathbf{I} \succ 0$. Therefore, (11.28) yields

$$\mathbf{x}^* = -(\mathbf{A} + \lambda^*\mathbf{I})^{-1}\mathbf{b}, \tag{11.38}$$

so that $||\mathbf{x}^*||^2 = ||(\mathbf{A} + \lambda^*\mathbf{I})^{-1}\mathbf{b}||^2 = \alpha$. The optimal solution is therefore given by (11.38), where λ^* is chosen as the unique root of the strictly decreasing function $f(\lambda) = ||(\mathbf{A} + \lambda\mathbf{I})^{-1}\mathbf{b}||^2 - \alpha$ over $(-\lambda_{\min}(\mathbf{A}), \infty)$.

An implementation of the algorithm for solving the trust region subproblem in MATLAB is given below by the function `trs`. The function uses the bisection method in order to find λ^*. Note that the function uses the fact that $f(\lambda) \to \infty$ as $\lambda \to -\lambda_{\min}(\mathbf{A})^+$ by taking the initial lower bound as $-\lambda_{\min}(\mathbf{A}) + \varepsilon$ for some small $\varepsilon > 0$.

```
function x_trs=trs(A,b,alpha)
%INPUT
%================
%A ................... an nxn matrix
%b ................... an n-length vector
%alpha ............... positive scalar
%OUTPUT
%================
% x_trs ............... an optimal solution of
%                       min{x'*A*x+2b'*x:||x||^2<=alpha}

n=length(b);
f=@(lam) norm((A+lam*eye(n))\b)^2-alpha;
[L,p]=chol(A,'lower');
% the case when A is positive definite
if (p==0)
    x_naive=-L'\(L\b);
    if(norm(x_naive)^2<=alpha)
        x_trs=x_naive;
    else
        u=1;
        while (f(u)>0)
            u=2*u;
        end
        lam=bisection(f,0,u,1e-7);
        x_trs=-(A+lam*eye(n))\b;
    end
else
    %when A is not positive definite
    u=max(1,-min(eig(A))+1e-7);
    while (f(u)>0)
        u=2*u;
    end
    lam=bisection(f,-min(eig(A))+1e-7,u,1e-7);
    x_trs=-(A+lam*eye(n))\b;
end
```

We can for example use the MATLAB function `trs` in order to solve Example 8.16.

```
>> A=[1,2,3;2,1,4;3,4,3];
>> b=[0.5;1;-0.5];
>> x_trs = trs(A,b,1)
x_trs =

   -0.2300
   -0.7259
    0.6482
```

This is of course the same as the solution obtained in Example 8.16.

11.7 ▪ Total Least Squares

Given an approximate linear system $\mathbf{Ax} \approx \mathbf{b}$ ($\mathbf{A} \in \mathbb{R}^{m\times n}, \mathbf{b} \in \mathbb{R}^m$), the least squares problem discussed in Chapter 3 can be seen as the problem of finding the minimum norm perturbation of the right-hand side of the linear system such that the resulting system is consistent:

$$\begin{array}{ll} \min_{\mathbf{w},\mathbf{x}} & ||\mathbf{w}||^2 \\ \text{s.t.} & \mathbf{Ax} = \mathbf{b} + \mathbf{w}, \\ & \mathbf{w} \in \mathbb{R}^m. \end{array}$$

This is a different presentation of the least squares problem from the one given in Chapter 3, but it is totally equivalent since plugging the expression for $\mathbf{w}$ ($\mathbf{w} = \mathbf{Ax} - \mathbf{b}$) into the objective function gives the well-known formulation of the problem as one consisting of minimizing the function $||\mathbf{Ax} - \mathbf{b}||^2$ over $\mathbb{R}^n$. The least squares problem essentially assumes that the right-hand side is unknown and is subjected to noise and that the matrix $\mathbf{A}$ is known and fixed. However, in many applications the matrix $\mathbf{A}$ is not exactly known and is also subjected to noise. In these cases it is more logical to consider a different problem, which is called *the total least squares problem*, in which one seeks to find a minimal norm perturbation to both the right-hand-side vector and the matrix so that the resulting perturbed system is consistent:

$$\text{(TLS)} \quad \begin{array}{ll} \min_{\mathbf{E},\mathbf{w},\mathbf{x}} & ||\mathbf{E}||_F^2 + ||\mathbf{w}||^2 \\ \text{s.t.} & (\mathbf{A} + \mathbf{E})\mathbf{x} = \mathbf{b} + \mathbf{w}, \\ & \mathbf{E} \in \mathbb{R}^{m\times n}, \mathbf{w} \in \mathbb{R}^m. \end{array}$$

Note that we use here the Frobenius norm as a matrix norm. Problem (TLS) is not a convex problem since the constraints are quadratic *equality* constraints. However, despite the nonconvexity of the problem we can use the KKT conditions in order to simplify it considerably and eventually even solve it. The trick is to fix $\mathbf{x}$ and solve the problem with respect to the variables $\mathbf{E}$ and $\mathbf{w}$:

$$(P_{\mathbf{x}}) \quad \begin{array}{ll} \min_{\mathbf{E},\mathbf{w}} & ||\mathbf{E}||_F^2 + ||\mathbf{w}||^2 \\ \text{s.t.} & (\mathbf{A} + \mathbf{E})\mathbf{x} = \mathbf{b} + \mathbf{w}. \end{array}$$

Problem $(P_{\mathbf{x}})$ is a linearly constrained convex problem and hence the KKT conditions are necessary and sufficient (Theorem 10.7). The Lagrangian of problem $(P_{\mathbf{x}})$ is given by

$$L(\mathbf{E}, \mathbf{w}, \lambda) = ||\mathbf{E}||_F^2 + ||\mathbf{w}||^2 + 2\lambda^T[(\mathbf{A} + \mathbf{E})\mathbf{x} - \mathbf{b} - \mathbf{w}].$$

By the KKT conditions, $(\mathbf{E},\mathbf{w})$ is an optimal solution of $(P_{\mathbf{x}})$ if and only if there exists $\lambda \in \mathbb{R}^m$ such that

$$2\mathbf{E} + 2\lambda\mathbf{x}^T = 0 \qquad (\nabla_{\mathbf{E}} L = 0), \tag{11.39}$$

$$2\mathbf{w} - 2\lambda = 0 \qquad (\nabla_{\mathbf{w}} L = 0), \tag{11.40}$$

$$(\mathbf{A}+\mathbf{E})\mathbf{x} = \mathbf{b}+\mathbf{w} \quad \text{(feasibility)}. \tag{11.41}$$

From (11.40) we have $\lambda = \mathbf{w}$. Substituting this in (11.39) we obtain

$$\mathbf{E} = -\mathbf{w}\mathbf{x}^T. \tag{11.42}$$

Combining (11.42) with (11.41) we have $(\mathbf{A} - \mathbf{w}\mathbf{x}^T)\mathbf{x} = \mathbf{b}+\mathbf{w}$, so that

$$\mathbf{w} = \frac{\mathbf{A}\mathbf{x}-\mathbf{b}}{\|\mathbf{x}\|^2+1}, \tag{11.43}$$

and consequently, by plugging the above into (11.42) we have

$$\mathbf{E} = -\frac{(\mathbf{A}\mathbf{x}-\mathbf{b})\mathbf{x}^T}{\|\mathbf{x}\|^2+1}. \tag{11.44}$$

Finally, by substituting (11.43) and (11.44) into the objective function of problem $(P_{\mathbf{x}})$ we obtain that the value of problem $(P_{\mathbf{x}})$ is equal to $\frac{\|\mathbf{A}\mathbf{x}-\mathbf{b}\|^2}{\|\mathbf{x}\|^2+1}$. Consequently, the TLS problem reduces to

$$(\text{TLS}') \quad \min_{\mathbf{x}\in\mathbb{R}^n} \frac{\|\mathbf{A}\mathbf{x}-\mathbf{b}\|^2}{\|\mathbf{x}\|^2+1}.$$

We have thus proven the following result.

Theorem 11.22. $\mathbf{x}$ *is an optimal solution of* (TLS′) *if and only if* $(\mathbf{x},\mathbf{E},\mathbf{w})$ *is an optimal solution of* (TLS) *where* $\mathbf{E} = -\frac{(\mathbf{A}\mathbf{x}-\mathbf{b})\mathbf{x}^T}{\|\mathbf{x}\|^2+1}$ *and* $\mathbf{w} = \frac{\mathbf{A}\mathbf{x}-\mathbf{b}}{\|\mathbf{x}\|^2+1}$.

The new formulation (TLS′) is much simpler than the original one. However, the objective function is still nonconvex, and the question that still remains is whether we can efficiently find an optimal solution of this simplified formulation. Using the special structure of the problem, we will show that the problem can actually be solved efficiently by using a *homogenization* argument. Indeed, problem (TLS′) is equivalent to

$$\min_{\mathbf{x}\in\mathbb{R}^n, t\in\mathbb{R}} \left\{ \frac{\|\mathbf{A}\mathbf{x}-t\mathbf{b}\|^2}{\|\mathbf{x}\|^2+t^2} : t = 1 \right\},$$

which is the same as (denoting $\mathbf{y} = \binom{\mathbf{x}}{t}$)

$$f^* = \min_{\mathbf{y}\in\mathbb{R}^{n+1}} \left\{ \frac{\mathbf{y}^T\mathbf{B}\mathbf{y}}{\|\mathbf{y}\|^2} : y_{n+1} = 1 \right\}, \tag{11.45}$$

where

$$\mathbf{B} = \begin{pmatrix} \mathbf{A}^T\mathbf{A} & -\mathbf{A}^T\mathbf{b} \\ -\mathbf{b}^T\mathbf{A} & \|\mathbf{b}\|^2 \end{pmatrix}.$$

Now, let us remove the constraint from problem (11.45) and consider the following problem:

$$g^* = \min_{\mathbf{y}\in\mathbb{R}^{n+1}} \left\{ \frac{\mathbf{y}^T\mathbf{B}\mathbf{y}}{\|\mathbf{y}\|^2} : \mathbf{y} \neq 0 \right\}. \tag{11.46}$$

This is a problem consisting of minimizing the so-called Rayleigh quotient (see Section 1.4) associated with the matrix $\mathbf{B}$, and hence an optimal solution is the vector corresponding to the minimum eigenvalue of $\mathbf{B}$; see Lemma 1.12. Of course, the obtained solution is not guaranteed to satisfy the constraint $y_{n+1} = 1$, but under a rather mild condition, the optimal solution of problem (11.45) can be extracted.

Lemma 11.23. *Let* $\mathbf{y}^*$ *be an optimal solution of* (11.46) *and assume that* $y^*_{n+1} \neq 0$. *Then* $\tilde{\mathbf{y}} = \frac{1}{y^*_{n+1}}\mathbf{y}^*$ *is an optimal solution of* (11.45).

Proof. Note that since (11.46) is formed from (11.45) by replacing the constraint $y_{n+1} = 1$ with $\mathbf{y} \neq 0$, we have

$$f^* \geq g^*.$$

However, $\tilde{\mathbf{y}}$ is a feasible point of problem (11.45) ($\tilde{y}_{n+1} = 1$), and we have

$$\frac{\tilde{\mathbf{y}}^T\mathbf{B}\tilde{\mathbf{y}}}{||\tilde{\mathbf{y}}||^2} = \frac{\frac{1}{(y^*_{n+1})^2}(\mathbf{y}^*)^T\mathbf{B}\mathbf{y}^*}{\frac{1}{(y^*_{n+1})^2}||\mathbf{y}^*||^2} = \frac{(\mathbf{y}^*)^T\mathbf{B}\mathbf{y}^*}{||\mathbf{y}^*||^2} = g^*.$$

Therefore, $\tilde{\mathbf{y}}$ attains the lower bound g^* on the optimal value of problem (11.45), and consequently, it is an optimal solution of (11.45) and the optimal values of the two problems (11.45) and (11.46) are consequently the same. □

All that is left is to find a computable condition under which $y^*_{n+1} \neq 0$. The following theorem presents such a condition and summarizes the solution method of the TLS problem.

Theorem 11.24. *Assume that the following condition holds:*

$$\lambda_{\min}(\mathbf{B}) < \lambda_{\min}(\mathbf{A}^T\mathbf{A}), \tag{11.47}$$

where

$$\mathbf{B} = \begin{pmatrix} \mathbf{A}^T\mathbf{A} & -\mathbf{A}^T\mathbf{b} \\ -\mathbf{b}^T\mathbf{A} & ||\mathbf{b}||^2 \end{pmatrix}.$$

Then the optimal solution of problem (TLS′) *is given by* $\frac{1}{y_{n+1}}\mathbf{v}$, *where* $\mathbf{y} = \binom{\mathbf{v}}{y_{n+1}}$ *is an eigenvector corresponding to the minimum eigenvalue of* $\mathbf{B}$.

Proof. By Lemma 11.23 all that we need to prove is that under condition (11.47), an optimal solution $\mathbf{y}^*$ of (11.46) must satisfy $y^*_{n+1} \neq 0$. Assume on the contrary that $y^*_{n+1} = 0$. Then

$$\lambda_{\min}(\mathbf{B}) = \frac{(\mathbf{y}^*)^T\mathbf{B}\mathbf{y}^*}{||\mathbf{y}^*||^2} = \frac{\mathbf{v}^T\mathbf{A}^T\mathbf{A}\mathbf{v}}{||\mathbf{v}||^2} \geq \lambda_{\min}(\mathbf{A}^T\mathbf{A}),$$

which is a contradiction to (11.47). □

Exercises

11.1. Consider the optimization problem

$$\text{(P)} \quad \begin{array}{ll} \min & x_1 - 4x_2 + x_3 \\ \text{s.t.} & x_1 + 2x_2 + 2x_3 = -2, \\ & x_1^2 + x_2^2 + x_3^2 \leq 1. \end{array}$$

(i) Given a KKT point of problem (P), must it be an optimal solution?

(ii) Find the optimal solution of the problem using the KKT conditions.

11.2. Consider the optimization problem

$$\text{(P)} \quad \min\{\mathbf{a}^T\mathbf{x} : \mathbf{x}^T\mathbf{Q}\mathbf{x} + 2\mathbf{b}^T\mathbf{x} + c \leq 0\},$$

where $\mathbf{Q} \in \mathbb{R}^{n\times n}$ is positive definite, $\mathbf{a}(\neq 0), \mathbf{b} \in \mathbb{R}^n$, and $c \in \mathbb{R}$.

(i) For which values of $\mathbf{Q}, \mathbf{b}, \mathbf{c}$ is the problem feasible?

(ii) For which values of $\mathbf{Q}, \mathbf{b}, \mathbf{c}$ are the KKT conditions necessary?

(iii) For which values of $\mathbf{Q}, \mathbf{b}, \mathbf{c}$ are the KKT conditions sufficient?

(iv) Under the condition of part (ii), find the optimal solution of (P) using the KKT conditions.

11.3. Consider the optimization problem

$$\begin{array}{ll} \min & x_1^4 - 2x_2^2 - x_2 \\ \text{s.t.} & x_1^2 + x_2^2 + x_2 \leq 0. \end{array}$$

(i) Is the problem convex?

(ii) Prove that there exists an optimal solution to the problem.

(iii) Find all the KKT points. For each of the points, determine whether it satisfies the second order necessary conditions.

(iv) Find the optimal solution of the problem.

11.4. Consider the optimization problem

$$\begin{array}{ll} \min & x_1^2 - x_2^2 - x_3^2 \\ \text{s.t.} & x_1^4 + x_2^4 + x_3^4 \leq 1. \end{array}$$

(i) Is the problem convex?

(ii) Find all the KKT points of the problem.

(iii) Find the optimal solution of the problem.

11.5. Consider the optimization problem

$$\begin{array}{ll} \min & -2x_1^2+2x_2^2+4x_1 \\ \text{s.t.} & x_1^2+x_2^2-4\le 0, \\ & x_1^2+x_2^2-4x_1+3\le 0. \end{array}$$

(i) Prove that there exists an optimal solution to the problem.

(ii) Find all the KKT points.

(iii) Find the optimal solution of the problem.

11.6. Use the KKT conditions in order to find an optimal solution of the each of the following problems:

(i)

$$\begin{array}{ll} \min & 3x_1^2+x_2^2 \\ \text{s.t.} & x_1-x_2+8,\le 0, \\ & x_2\ge 0. \end{array}$$

(ii)

$$\begin{array}{ll} \min & 3x_1^2+x_2^2 \\ \text{s.t.} & 3x_1^2+x_2^2+x_1+x_2+0.1\le 0, \\ & x_2+10\ge 0. \end{array}$$

(iii)

$$\begin{array}{ll} \min & 2x_1+x_2 \\ \text{s.t.} & 4x_1^2+x_2^2-2\le 0, \\ & 4x_1+x_2+3\le 0. \end{array}$$

(iv)

$$\begin{array}{ll} \min & x_1^3+x_2^3 \\ \text{s.t.} & x_1^2+x_2^2\le 1. \end{array}$$

(v)

$$\begin{array}{ll} \min & x_1^4-x_2^2 \\ \text{s.t.} & x_1^2+x_2^2\le 1, \\ & 2x_2+1\le 0. \end{array}$$

11.7. Let $a>0$. Find all the optimal solutions of

$$\max\{x_1x_2x_3 : a^2x_1^2+x_2^2+x_3^2\le 1\}.$$

11.8. (i) Find a formula for the orthogonal projection of a vector $\mathbf{y}\in\mathbb{R}^3$ onto the set

$$C=\{\mathbf{x}\in\mathbb{R}^3 : x_1^2+2x_2^2+3x_3^2\le 1\}.$$

The formula should depend on a single parameter that is a root of a strictly decreasing one-dimensional function.

(ii) Write a MATLAB function whose input is a three-dimensional vector and its output is the orthogonal projection of the input onto C.

11.9. Consider the optimization problem

$$\text{(P)}\quad \begin{array}{ll} \min & 2x_1x_2+\frac{1}{2}x_3^2 \\ \text{s.t.} & 2x_1x_3+\frac{1}{2}x_2^2\le 0, \\ & 2x_2x_3+\frac{1}{2}x_1^2\le 0. \end{array}$$

(i) Show that the optimal solution of problem (P) is $\mathbf{x}^* = (0,0,0)$.

(ii) Show that $\mathbf{x}^*$ does not satisfy the second order necessary optimality conditions.

11.10. Consider the convex optimization problem

$$(\mathrm{P}) \quad \begin{array}{ll} \min & f_0(\mathbf{x}) \\ \text{s.t.} & f_i(\mathbf{x}) \le 0, \quad i = 1,2,\ldots,m, \end{array}$$

where f_0 is a continuously differentiable convex function and $f_1, f_2, \ldots, f_m$ are continuously differentiable *strictly* convex functions. Let $\mathbf{x}^*$ be a feasible solution of (P). Suppose that the following condition is satisfied: there exist $y_i \ge 0, i \in \{0\} \cup I(\mathbf{x}^*)$, which are not all zeros such that

$$y_0 \nabla f_0(\mathbf{x}^*) + \sum_{i \in I(\mathbf{x}^*)} y_i \nabla f_i(\mathbf{x}^*) = 0.$$

Prove that $\mathbf{x}^*$ is an optimal solution of (P).

11.11. Consider the optimization problem

$$\begin{array}{ll} \min & \mathbf{c}^T \mathbf{x} \\ \text{s.t.} & f_i(\mathbf{x}) \le 0, \quad i = 1,2,\ldots,m, \end{array}$$

where $\mathbf{c} \neq 0$ and $f_1, f_2, \ldots, f_m$ are continuous over $\mathbb{R}^n$. Prove that if $\mathbf{x}^*$ is a local minimum of the problem, then $I(\mathbf{x}^*) \neq \emptyset$.

11.12. Consider the QCQP problem

$$(\mathrm{QCQP}) \quad \begin{array}{ll} \min & \mathbf{x}^T \mathbf{A}_0 \mathbf{x} + 2\mathbf{b}_0^T \mathbf{x} \\ \text{s.t.} & \mathbf{x}^T \mathbf{A}_i \mathbf{x} + 2\mathbf{b}_i^T \mathbf{x} + c_i \le 0, \quad i = 1,2,\ldots,m, \end{array}$$

where $\mathbf{A}_0, \mathbf{A}_1, \ldots, \mathbf{A}_m \in \mathbb{R}^{n \times n}$ are symmetric matrices, $\mathbf{b}_0, \mathbf{b}_1, \ldots, \mathbf{b}_m \in \mathbb{R}^n$, and $c_1, c_2, \ldots, c_m \in \mathbb{R}$. Suppose that $\mathbf{x}^*$ satisfies the following condition: there exist $\lambda_1, \lambda_2, \ldots, \lambda_m \ge 0$ such that

$$\begin{aligned} \left(\mathbf{A}_0 + \sum_{i=1}^m \lambda_i \mathbf{A}_i\right)\mathbf{x}^* + \left(\mathbf{b}_0 + \sum_{i=1}^m \lambda_i \mathbf{b}_i\right) &= 0, \\ \lambda_i \left[(\mathbf{x}^*)^T \mathbf{A}_i (\mathbf{x}^*) + 2\mathbf{b}_i^T \mathbf{x}^* + c_i\right] &= 0, \quad i = 1,2,\ldots,m, \\ (\mathbf{x}^*)^T \mathbf{A}_i (\mathbf{x}^*) + 2\mathbf{b}_i^T \mathbf{x}^* + c_i &\le 0, \quad i = 1,2,\ldots,m, \\ \mathbf{A}_0 + \sum_{i=1}^m \lambda_i \mathbf{A}_i &\succeq 0. \end{aligned}$$

Prove that $\mathbf{x}^*$ is an optimal solution of (QCQP).