Chapter 8

Introduction to inequality-constrained minimization

8.1 Introduction to inequality-constrained optimization: The logarithmic barrier method

I now turn to the inequality-constrained nonlinear program

\[
\begin{align*}
\min f(x) \\
\text{s.t. } h(x) &\geq 0,
\end{align*}
\]

(8.1)

(8.2)

where \( f : \mathbb{R}^n \to \mathbb{R} \) and \( h : \mathbb{R}^n \to \mathbb{R}^p \). The reader should notice that \( h(x) \geq 0 \) is to be interpreted component-wise; that is, \( h(x) \geq 0 \) if and only if \( h_i(x) \geq 0 \) for each \( i = 1, 2, \ldots, p \). Inequality constraints are common in optimization problems; indeed, almost every optimization problems does or could have inequality constraints, although it is sometimes safe to ignore them.

The most common type of inequality constraints are simple bounds that the variables must satisfy. For example, if every variable must satisfy both lower and upper bound, then the constraints could be written as

\[ a \leq x \leq b, \]

where \( a, b \in \mathbb{R}^n \) with \( a \leq b \). To incorporate these constraints into the standard form (8.2), one would write them as

\[ x - a \geq 0, \]

\[ b - x \geq 0. \]

Therefore, if a problem contained only these simple bounds, the constraint function
$h : \mathbb{R}^n \to \mathbb{R}^p \ (p = 2n)$ would be defined as

$$h(x) = \begin{bmatrix}
    x - a \\
    b - x
\end{bmatrix} =
\begin{bmatrix}
    x_1 - a_1 \\
    x_2 - a_2 \\
    \vdots \\
    x_n - a_n \\
    b_1 - x_1 \\
    b_2 - x_2 \\
    \vdots \\
    b_n - x_n
\end{bmatrix}.$$

Simple bounds are common because variables typically have a physical interpretation and some real numbers do not make physical sense for a given variable. For example, physical parameters (density, elasticity, thermal conductivity, etc.) typically cannot take negative values, so nonnegativity constraints of the form

$$x \geq 0$$

are common. The same constraint appears when the variables represents quantities (for example, the number of barrels of petroleum) that cannot be negative. Upper bounds often represent limited resources.

In some cases, it may seem that the appropriate constraint is a strict inequality, as when the variables represent physical parameters that must be strictly positive. However, a strict inequality constraint may lead to an optimization problem that is ill-posed in that a minimizer is infeasible but on the boundary of the feasible set. In such a case, there may not be a solution to the optimization problem. A simple example of this is

$$\begin{align*}
    \min & \quad x^2 \\
    \text{s.t.} & \quad x > 0.
\end{align*}$$

For this reason, strict inequality constraints are not used in nonlinear programming.

When the appropriate constraint seems to be a strict inequality, one of the following is usually true:

1. The problem is expected to have a solution that easily satisfies the strict inequality. In this case, as will be shown below, the constraint plays little part in the theory or algorithm other than as a “sanity check” on the variable. Therefore, the nonstrict inequality constraint is just as useful.

2. Due to noise in the data or other complications, the solution to the optimization problem may lie on the boundary of the feasible set, even though such a solution is not physically meaningful. In this case, the inequality must be perturbed slightly and written as a nonstrict inequality. For example, a constraint of the form $x_i > 0$ should be replaced with $x_i \geq a_i$, where $a_i > 0$ is the smallest value of $x_i$ that is physically meaningful.
It may not be clear how to distinguish the first case from the second; if it is not, the second approach is always valid. The first case can usually be identified by the fact that the solution is not expected to be close to satisfying the inequality as an equation.

Inequality constraints that are not simple bounds are usually bounds on derived quantities. For example, if \( x \) represents design parameters for a certain object, the mass \( m \) of the object may be represented as a function of \( x \): \( m = q(x) \). In this case, constraints of the form \( q(x) \geq M_1 \) or \( q(x) \leq M_2 \) (or both) may be appropriate.

Here is a simple example of an inequality-constrained nonlinear program.

**Example 8.1.** Define \( f : \mathbb{R}^2 \to \mathbb{R} \) and \( h : \mathbb{R}^2 \to \mathbb{R}^2 \) by

\[
f(x) = (x_1 - 1)^2 + 2(x_2 - 2)^2, \quad h(x) = \begin{bmatrix} 1 - x_1^2 - x_2^2 \\ x_1 + x_2 \end{bmatrix}
\]

The feasible set for

\[
\begin{align*}
\min f(x) \\
\text{s.t. } h(x) &\geq 0
\end{align*}
\]

is shown in Figure 8.1, together with the contours of the objective function \( f \).

![Figure 8.1](image_url)

**Figure 8.1.** The contours of \( f \) and the feasible set determined by \( h(x) \geq 0 \) (see Example 8.1). The feasible set is half of the unit disk (the shaded region). The minimizer \( x^* \approx (0.3116, 0.9502) \) is indicated by an asterisk.

An important aspect of this example is that the second constraint, \( x_1 + x_2 \geq 0 \), does not affect the solution. If the second constraint were changed to \( x_1 + x_2 \geq u \) for
some value of $u$ that is not too large, or if the second constraint were simply omitted, the optimization problem would have the same solution. Both the theory of and algorithms for inequality-constrained problems must address the issue of “inactive” constraints.

The first step in analyzing NLP (8.1–8.2) should be to derive the optimality conditions. However, since the optimality conditions are somewhat more complicated than in the case of equality constraints, I will begin by presenting an algorithm for solving (8.1–8.2), namely, the logarithmic barrier method. From this algorithm, the optimality conditions for the inequality-constrained NLP will be deduced. A rigorous derivation of these optimality conditions will be given later.

### 8.1.1 The logarithmic barrier method

Assuming it is possible to find a strictly feasible point $x^{(0)}$, that is, a point satisfying $h(x^{(0)}) > 0$, a natural strategy for solving (8.1–8.2) is to decrease $f$ as much as possible while ensuring that the boundary of the feasible set is never crossed. One way to prevent an optimization algorithm from crossing the boundary is to assign a penalty to approaching it. The most popular way of doing this is to augment the objective function by a logarithmic barrier term:

$$
B(x; \mu) = f(x) - \mu \sum_{i=1}^{p} \log (h_i(x)).
$$

(8.3)

Here $\log$ denotes the natural logarithm. Since

$$
- \log(t) \to \infty \text{ as } t \to 0,
$$

$B(\cdot; \mu)$ “blows up” at the boundary and therefore presents an optimization algorithm with a “barrier” to crossing the boundary. Of course, the solution to an inequality-constrained is likely to lie on the boundary of the feasible set, so the barrier must be gradually removed by reducing $\mu$ toward zero. This suggests the following strategy:

Choose $\mu_0 > 0$ and a strictly feasible point $x^{(0)}$.

For $k = 1, 2, 3, \ldots$

Choose $\mu_k \in (0, \mu_{k-1})$ (perhaps $\mu_k = \beta \mu_{k-1}$ for some constant $\beta \in (0, 1)$).

Using $x^{(k-1)}$ as the starting point, solve

$$
\min B(x; \mu_k)
$$

to get $x^{(k)}$.

Later I will prove that, under certain conditions, $B(\cdot; \mu)$ has a unique minimizer $x^*_\mu$ in a neighborhood of $x^*$ and that $x^*_\mu \to x^*$ as $\mu \to 0$. 

**Figure**
8.1 Introduction to inequality-constrained optimization: The logarithmic barrier method

<table>
<thead>
<tr>
<th>$\mu$</th>
<th>$| x^* - x^*_\mu |$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.5912 \cdot 10^{-1}</td>
</tr>
<tr>
<td>$2^{-1}$</td>
<td>9.4064 \cdot 10^{-2}</td>
</tr>
<tr>
<td>$2^{-2}$</td>
<td>5.1947 \cdot 10^{-2}</td>
</tr>
<tr>
<td>$2^{-3}$</td>
<td>2.7449 \cdot 10^{-2}</td>
</tr>
<tr>
<td>$2^{-4}$</td>
<td>1.4134 \cdot 10^{-2}</td>
</tr>
<tr>
<td>$2^{-5}$</td>
<td>7.1748 \cdot 10^{-3}</td>
</tr>
<tr>
<td>$2^{-6}$</td>
<td>3.6152 \cdot 10^{-3}</td>
</tr>
<tr>
<td>$2^{-7}$</td>
<td>1.8146 \cdot 10^{-3}</td>
</tr>
</tbody>
</table>

**Table 8.1.** Errors in the solutions computed by the logarithmic barrier method (see Example 8.2).

For now, however, I will assume that the above method works, first applying it to Example 8.1 and then using it to deduce the optimality conditions for NLP (8.1–8.2).

**Example 8.2 (Example 8.1, continued).** Let $f$ and $h$ be defined as in Example 8.1, and consider the NLP

$$\begin{align*}
\min & \; f(x) \\
\text{s.t.} & \; h(x) \geq 0.
\end{align*}$$

Minimizing $B(\cdot, \mu)$ by Newton’s method for $\mu = 2^0, 2^{-1}, \ldots, 2^{-7}$, beginning with $x^{[0]} = (0.5, 0.5)$, yield the results shown in Figure 8.2. The errors in the computed solutions are displayed in Table 8.1. The reader will notice that, at least for this example, $\| x^* - x^*_\mu \| = O(\mu)$ appears to hold.

### 8.1.2 Optimality conditions for inequality-constrained problems: Introduction

I now assume that $x^*$ is a local minimizer of (8.1–8.2) and that the logarithmic barrier method works as intended:

1. For all $\mu$ sufficiently small, $B(\cdot; \mu)$ has a unique local minimizer $x^*_\mu$ in a neighborhood of $x^*$; and
2. $x^*_\mu \to x^*$ as $\mu \to 0$.

Then, for all $\mu$ sufficiently small,

$$\nabla B(x^*_\mu; \mu) = 0.$$ 

The following notation will be used: For any vector $v$, $v^{-1}$ is the vector of the same size defined by $v_i^{-1} = 1/v_i$. (Of course, this is only defined if every component of $v$
Figure 8.2. The approximate solutions to Example 8.1 computed by the logarithmic barrier method (see Example 8.2).

is nonzero.) Then it is easy to show that

\[
\nabla B(x; \mu) = \nabla f(x) - \mu \sum_{i=1}^{p} \frac{1}{h_i(x)} \nabla h_i(x) \\
= \nabla f(x) - \nabla h(x)(\mu h(x)^{-1}).
\]

In particular, \(\nabla B(x^*_\mu; \mu) = 0\) implies that

\[
\nabla f(x^*_\mu) = \nabla h(x^*_\mu)(\mu h(x^*_\mu)^{-1}).
\]

Since \(x^*_\mu \to x^*\) as \(\mu \to 0\), this suggests that there exists \(\lambda^* \in \mathbb{R}^p\) with

\[
\mu h(x^*_\mu)^{-1} \to \lambda^* \text{ as } \mu \to 0
\]

and

\[
\nabla f(x^*) = \nabla h(x^*)\lambda^*.
\]

I will now assume that such a \(\lambda^*\) exists (rigorous proofs will be given later). Then, since \(h(x^*_\mu) > 0\) for all \(\mu > 0\), it follows that

\[
\lambda^* \geq 0.
\]

Moreover, for any \(i\) such that \(h_i(x^*) > 0\),

\[
\lambda^*_i = \lim_{\mu \to 0} \frac{\mu}{h_i(x^*_\mu)} = \frac{0}{h_i(x^*)} = 0.
\]
This analysis suggests the following: If \( x^* \) is a local minimizer of (8.1–8.2), then there exists a Lagrange multiplier \( \lambda^* \in \mathbb{R}^p \) such that

\[
\nabla f(x^*) = \nabla h(x^*) \lambda^*.
\]

In addition, \( \lambda^* \) satisfies the following properties:

\[
\lambda^*_i \geq 0 \text{ for all } i = 1, 2, \ldots, p,
\]

\[
h_i(x^*) > 0 \Rightarrow \lambda^*_i = 0.
\]

The second condition means that, for each \( i \), either \( h_i(x^*) = 0 \) or \( \lambda^*_i = 0 \) (or both). This is referred to as the complementarity condition and can be written as

\[
\lambda^*_i h_i(x^*) = 0, \text{ for all } i = 1, 2, \ldots, p.
\]

If \( h_i(x^*) = 0 \), then constraint \( i \) is said to be active (or binding) at \( x^* \). On the other hand, if \( h_i(x^*) > 0 \), then constraint \( i \) is inactive. The equation

\[
\nabla f(x^*) = \nabla h(x^*) \lambda^* = \sum_{i=1}^{p} \lambda^*_i \nabla h_i(x^*),
\]

together with the fact that \( \lambda^*_i = 0 \) if constraint \( i \) is inactive, means that \( \nabla f(x^*) \) can be expressed as a linear combination of the gradients of the active constraints. Therefore, the inactive constraints do not play a role in the first-order necessary conditions. This is not surprising, as I pointed out above that a constraint that is inactive at the solution is locally unimportant.

My analysis suggests that the first-order necessary conditions for \( x^* \) to be a local minimizer of (8.1–8.2) are:

There exists \( \lambda^* \in \mathbb{R}^p \) that, together with \( x^* \), satisfies

\[
\nabla f(x^*) = \nabla h(x^*) \lambda^*;
\]

\[
h(x^*) \geq 0;
\]

\[
\lambda^* \geq 0;
\]

\[
\lambda^*_i h_i(x^*) = 0 \text{ for all } i = 1, 2, \ldots, p.
\]

Based on the theory for equality-constrained NLPs, the reader should not be surprised to learn that a constraint qualification is required to ensure that the Lagrange multiplier \( \lambda^* \) exists.

### 8.1.3 Another barrier function

Although the logarithmic barrier function is the most popular barrier function, it is not the only one. An alternative is

\[
\tilde{B}(x; \mu) = f(x) + \mu \sum_{i=1}^{p} \frac{1}{h_i(x)}.
\]
Assuming that \( x \) is strictly feasible, it is clear that \( \tilde{B}(x;\mu) \to \infty \) as \( x \) approaches the boundary of the feasible set.

Much the same theory can be developed for the barrier function \( B \) as for the logarithmic barrier function \( B \). However, \( B \) is usually preferred in practice because the logarithm increases so slowly as the boundary is approached.

### 8.2 Optimality conditions for inequality-constrained nonlinear programs

I now turn to a rigorous development of the optimality conditions for

\[
\begin{align*}
\min f(x) \\
\text{s.t. } h(x) \geq 0,
\end{align*}
\]  

(8.4) (8.5)

where \( f : \mathbb{R}^n \to \mathbb{R} \) and \( h : \mathbb{R}^n \to \mathbb{R}^p \). I will assume that \( x^* \) is local minimizer of (8.4–8.5) and consider a feasible path \( x : [0,a] \to \mathbb{R}^n \) such that \( x(0) = x^* \). In contrast to the case of an equality-constrained NLP, it is natural to consider “one-sided” paths since \( x^* \) may lie on the boundary of the feasible region. In such a case, \( \dot{x}(0) \) may be a feasible direction while \( -\dot{x}(0) \) is not.

If \( x \) is any feasible path satisfying \( x(0) = x^* \) and \( \phi : [0,a] \to \mathbb{R} \) is defined by

\[
\phi(t) = f(x(t)), \quad t \in [0,a],
\]

then \( t = 0 \) is a local minimizer of \( \phi \). Since 0 is an endpoint of \([0,a]\), it need not be the case that \( \phi'(0) = 0 \). For example, \( \phi : [0,a] \to \mathbb{R} \) defined by \( \phi(t) = t \) has a local minimizer at \( t = 0 \) and yet \( \phi'(0) = 1 \). As this example suggests, the relevant condition on \( \phi'(0) \) is \( \phi'(0) \geq 0 \). Since

\[
\phi'(0) = \nabla f(x^*) \cdot \dot{x}(0),
\]

the first form of the necessary condition is:

If \( x : [0,a] \to \mathbb{R}^n \) is a feasible path satisfying \( x(0) = x^* \), then

\[
\nabla f(x^*) \cdot \dot{x}(0) \geq 0.
\]

As in the case of equality constraints, it is not convenient to work directly with feasible paths. The next step, therefore, is to express the optimality condition in terms of feasible directions. Characterizing feasible directions is now considerably more complicated because it is necessary to distinguish between active and inactive constraints. The notation

\[
\mathcal{A}(x) = \{i : h_i(x) = 0\}
\]

will be used to denote the (indices of the) active constraints at a feasible point \( x \).

If \( h_i(x^*) > 0 \) (that is, \( i \not\in \mathcal{A}(x^*) \)), then no sufficiently small change in \( x^* \) will cause it to violate the \( i \)th constraint. In other words, every direction is feasible as far as the \( i \)th constraint is concerned. On the other hand, if \( h_i(x^*) = 0 \) (that is,
8.2. Optimality conditions for inequality-constrained nonlinear programs

If \( i \in \mathcal{A}(x^*) \), then a direction \( z \) is feasible only if it points into the feasible set or along (that is, tangent to) its boundary. In other words, \( z \) is a feasible direction, with respect to the \( i \)th constraint \( (i \in \mathcal{A}(x^*)) \), if

\[
\nabla h_i(x^*) \cdot z \geq 0. \tag{8.6}
\]

This inequality says that either \( z \) is tangent to the set \( h_i(x) = 0 \) at \( x = x^* \) \((\nabla h_i(x^*) \cdot z = 0)\) or else it points in a direction in which \( h_i \) increases \((\nabla h_i(x^*) \cdot z > 0)\). To demonstrate the necessity of (8.6) directly, suppose \( x : [0,a] \to \mathbb{R}^n \) is a feasible path with \( x(0) = x^* \) and \( \dot{x}(0) = z \). Then

\[
h_i(x(0)) = 0, \nabla h_i(x(0)) \cdot \dot{x}(0) < 0 \Rightarrow \frac{d}{dt}[h_i(x(t))]_{t=0} < 0 \Rightarrow \dot{h}(x(t)) < 0 \text{ for all } t > 0 \text{ sufficiently small.}
\]

But, by assumption, \( h(x(t)) \geq 0 \) for all \( t \in [0,a] \). This contradiction shows that \( \nabla h_i(x^*) \cdot z < 0 \) is impossible for a direction \( z \) that is feasible at \( x^* \) when \( h_i(x^*) = 0 \); in other words, any feasible direction \( z \) at \( x^* \) must satisfy \( \nabla h_i(x^*) \cdot z \geq 0 \) if \( h_i(x^*) = 0 \).

To summarize, if \( z \) is a feasible direction at \( x^* \), then

\[
\nabla h_i(x^*) \cdot z \geq 0 \text{ for all } i \in \mathcal{A}(x^*). \tag{8.7}
\]

The set of all feasible directions \( z \) at \( x^* \) is called the tangent cone (to the feasible set) at \( x^* \). It would be convenient if (8.7) characterized the tangent cone at \( x^* \). However, this is not necessarily true unless the constraint function \( h \) satisfies a constraint qualification. The first constraint qualification is simply a statement of the desired property:

**Constraint qualification 8.3.** There exists a feasible path \( x : [0,a] \to \mathbb{R}^n \) such that \( x(0) = x^* \) and \( \dot{x}(0) = z \) if and only if \nabla h_i(x^*) \cdot z \geq 0 \text{ for all } i \in \mathcal{A}(x^*). \tag{8.8}

Moreover, if (8.8) holds, then the feasible path \( x \) can be chosen to satisfy \( h_i(x(t)) = 0 \) for all \( t \in [0,a] \) for each \( i \) such that \( \nabla h_i(x^*) \cdot z = 0 \).

The following definition gives a notion of regular point suitable for inequality constraints.

**Definition 8.4.** Suppose \( h : \mathbb{R}^n \to \mathbb{R}^p \) is continuously differentiable and \( x^* \in \mathbb{R}^n \) satisfies \( h(x^*) \geq 0 \). Then \( x^* \) is called a regular point of the constraint \( h(x) \geq 0 \) if \nabla h_i(x^*) : i \in \mathcal{A}(x^*) \}

is a linearly independent set.

The proof of the following theorem is very similar to the analogous proof for equality constraints and will be omitted.
Theorem 8.5. Suppose $h : \mathbb{R}^n \to \mathbb{R}^p$ is continuously differentiable and $x^*$ is a regular point of the constraint $h(x) \geq 0$. Then Constraint Qualification 8.3 holds at $x^*$.

The reader should notice that, once again, it is the constraints active at $x^*$ that must be taken into account.

Assuming that $x^*$ is a regular point (or, more generally, that Constraint Qualification 8.3 holds), a necessary condition for $x^*$ to be a local minimizer of (8.4–8.5) is

$$\nabla h_i(x^*) \cdot z \geq 0 \quad \text{for all } i \in A(x^*) \Rightarrow \nabla f(x^*) \cdot z \geq 0. \quad (8.9)$$

To understand the next step in the development, it may help to remember the analogous reasoning for equality constraints. I showed that, if $x^*$ was a local minimizer under the constraint $g(x) = 0$ and the constraint qualification held at $x = x^*$, then

$$z \in \mathcal{N}(\nabla g(x^*)^T) \Rightarrow \nabla f(x^*) \cdot z = 0.$$

The Fundamental Theorem of Linear Algebra was then used to write $\nabla f(x^*)$ as a linear combination of the columns of $\nabla g(x^*)$: $\nabla f(x^*) = \nabla g(x^*)^T \lambda^*$. A result analogous to the Fundamental Theorem of Algebra is needed to put (8.9) into a more usable form. This result is Farkas’s lemma:

Theorem 8.6 (Farkas’s lemma). For any $A \in \mathbb{R}^{n \times p}, c \in \mathbb{R}^n$, exactly one of the two problems has a solution:

1. $A^T x \geq 0, c \cdot x < 0$;
2. $c = A\lambda, \lambda \geq 0$.

Farkas’s lemma is rather difficult to prove; a number of different proofs have been discovered, but none is very elementary. So as not to distract the reader from the argument here, I will give a proof of Farkas’s lemma after completing this discussion.

The following is a corollary of Farkas’s lemma:

Corollary 8.7. A matrix $A \in \mathbb{R}^{n \times p}$ and a vector $c \in \mathbb{R}^n$ satisfy

$$A^T x \geq 0 \Rightarrow c \cdot x \geq 0, \quad (8.10)$$

if and only if there exists $\lambda \in \mathbb{R}^p$ such that $\lambda \geq 0$ and

$$c = A\lambda. \quad (8.11)$$

The proof is immediate from (8.6).

To apply this corollary, I define $A$ to be the matrix whose columns are the gradients of the active constraints at $x^*$ and take $c$ to be $\nabla f(x^*)$. Then (8.9) is
8.2. Optimality conditions for inequality-constrained nonlinear programs

equivalent to (8.10) and hence implies (8.11); that is, (8.9) implies

There exist \( \lambda^*_i \geq 0, \ i \in \mathcal{A}(x^*) \), such that \( \nabla f(x^*) = \sum_{i \in \mathcal{A}(x^*)} \lambda^*_i \nabla h_i(x^*) \). (8.12)

Defining \( \lambda^*_i = 0 \) for \( i \notin \mathcal{A}(x^*) \), the necessary condition can be written as follows:

**Theorem 8.8.** Suppose \( f : \mathbb{R}^n \rightarrow \mathbb{R} \) and \( h : \mathbb{R}^n \rightarrow \mathbb{R} \) are continuously differentiable, and \( x^* \) is a local minimizer of

\[
\begin{align*}
\min f(x) \\
\text{s.t. } h(x) \geq 0.
\end{align*}
\]

If Constraint Qualification 8.3 holds at \( x^* \), then there exists \( \lambda^* \in \mathbb{R}^p \) such that \( x^* \) and \( \lambda^* \) satisfy

\[
\begin{align*}
\nabla f(x^*) &= \nabla h(x^*) \lambda^*, \\
h(x^*) &\geq 0, \\
\lambda^* &\geq 0, \\
\lambda^*_i h_i(x^*) &= 0 \text{ for all } i = 1, 2, \ldots, p.
\end{align*}
\]

The last condition is referred to as complementarity.

Of course, Theorem 8.12 is more useful than (8.9), since it expresses the necessary condition in terms of a system of equations. Defining the Lagrangian \( \ell \) to be the function

\( \ell(x; \lambda) = f(x) - \lambda \cdot h(x) \),

the first-order necessary condition can be written as

\( \nabla \ell(x^*; \lambda^*) = 0 \),

subject to the nonnegativity constraints on \( \lambda^* \) and the complementarity condition.

The reader should notice that Theorem 8.12 specifies the conditions that were already deduced by assuming that the logarithmic barrier algorithm converges.

### 8.2.1 Second-order necessary conditions

Second-order optimality conditions are complicated by the fact that it is necessary to distinguish not only between active and inactive constraints, but also between active constraints corresponding to positive Lagrange multipliers and active constraints corresponding to zero Lagrange multipliers. The basic principle determining the second-order necessary condition is this: If \( z \) is a feasible direction at \( x^* \) and \( f \) has zero slope in the direction of \( z \) (that is, if \( \nabla f(x^*) \cdot z = 0 \)), then the curvature of the Lagrangian\(^{22}\) must be nonnegative in that direction:

\[
z \cdot \nabla^2 \ell(x^*; \lambda^*) z \geq 0.
\]

\(^{22}\) The reader will recall that it is the curvature of the Lagrangian \( \ell \), not just the objective function \( f \), that is important. This is because if \( x \) is a feasible path, then the curvature of \( \phi(t) = f(x(t)) \) is determined both by the curvature of \( f \) and by the curvature of the constraints. The Hessian of \( \ell \) accounts for both sources of curvature.
The directions in which \( f \) has zero slope can be determined from the formula
\[
\nabla f(x^*) = \nabla h(x^*)\lambda^*.
\]
This shows that
\[
\nabla f(x^*) \cdot z = (\nabla h(x^*)\lambda^*) \cdot z
= \lambda^* \cdot (\nabla h(x^*)^T z)
= \sum_{i=1}^{p} \lambda_i^* \nabla h_i(x^*) \cdot z
= \sum_{\lambda_i^* \neq 0} \lambda_i^* \nabla h_i(x^*) \cdot z.
\]
For any \( i \notin \mathcal{A}(x^*) \), \( \lambda_i^* = 0 \) must hold because of the complementarity condition:
\[
\lambda_i^* h_i(x^*) = 0 \text{ for all } i = 1, 2, \ldots, p.
\]
However, it may be the case that \( \lambda_i^* = 0 \) for certain \( i \in \mathcal{A}(x^*) \) (if both \( \lambda_i^* \) and \( h_i(x^*) \) are zero). In this regard, it is helpful to define the strict complementarity condition:
\[
\text{Exactly one of } \lambda_i^* \text{ and } h_i(x^*) \text{ is zero.} \tag{8.13}
\]
The subset of \( \mathcal{A}(x^*) \) consisting of those indices \( i \) such that strict complementarity holds for constraint \( i \) will be denoted by \( \bar{\mathcal{A}}(x^*) \). If \( \bar{\mathcal{A}}(x^*) = \mathcal{A}(x^*) \), then it is said simply that \( x^* \) satisfies the strict complementarity condition.

Using this notation,
\[
\nabla f(x^*) \cdot z = \sum_{i \in \bar{\mathcal{A}}(x^*)} \lambda_i^* \nabla h_i(x^*) \cdot z.
\]
Since \( \lambda_i^* > 0 \) for all \( i \in \bar{\mathcal{A}}(x^*) \) and \( \nabla h_i(x^*) \cdot z \geq 0 \) for all \( i \in \mathcal{A}(x^*) \), it follows that
\[
\nabla f(x^*) \cdot z = 0
\]
for a feasible direction \( z \) if and only if
\[
\nabla h_i(x^*) \cdot z = 0 \text{ for all } i \in \bar{\mathcal{A}}(x^*),
\]
\[
\nabla h_i(x^*) \cdot z \geq 0 \text{ for all } i \in \mathcal{A}(x^*) \setminus \bar{\mathcal{A}}(x^*),
\]
that is, if and only if
\[
\nabla h_i(x^*) \cdot z = 0 \text{ for all } i \in \mathcal{A}(x^*) \text{ such that } \lambda_i^* > 0,
\]
\[
\nabla h_i(x^*) \cdot z \geq 0 \text{ for all } i \in \mathcal{A}(x^*) \text{ such that } \lambda_i^* = 0.
\]
The set of all vectors \( z \) satisfying these conditions will be denoted by \( D(x^*, \lambda^*) \).

This discussion suggests the following result:
8.2. Optimality conditions for inequality-constrained nonlinear programs

**Theorem 8.9.** Suppose \( f : \mathbb{R}^n \to \mathbb{R} \) and \( h : \mathbb{R}^n \to \mathbb{R}^p \) are continuously differentiable, \( x^* \) is a local minimizer of

\[
\begin{align*}
\min f(x) \\
\text{s.t. } h(x) &\geq 0,
\end{align*}
\]

and \( \lambda^* \) is a corresponding Lagrange multiplier. If Constraint Qualification (8.3) is satisfied at \( x^* \), then

\[
z \cdot \nabla^2 \ell(x^*; \lambda^*) z \geq 0 \text{ for all } z \in D(x^*, \lambda^*).
\]

**Proof:** Suppose there exists \( z \in D(x^*, \lambda^*) \) such that

\[
z \cdot \nabla^2 \ell(x^*; \lambda^*) z < 0.
\]

Let \( x : [0,a] \to \mathbb{R}^n \) be a feasible path satisfying \( x(0) = x^* \), \( x'(0) = z \) and define \( \phi : [0,a] \to \mathbb{R} \) by

\[
\phi(t) = f(x(t)).
\]

Then

\[
\phi'(0) = \nabla f(x^*) \cdot z = 0
\]

since \( z \in D(x^*, \lambda^*) \) and (8.14) implies that \( \phi''(0) < 0 \). Taylor’s theorem applied to \( \phi \) then shows that

\[
\phi(t) < \phi(0) \text{ for all } t \text{ sufficiently small},
\]

that is,

\[
f(x(t)) < f(x^*) \text{ for all } t \text{ sufficiently small}.
\]

This contradicts the local optimality of \( x^* \) and shows that (8.14) cannot hold. QED

### 8.2.2 Second-order sufficient conditions

The following theorem gives sufficient conditions for \( x^* \) to be a strict local minimizer of (8.4–8.5).

**Theorem 8.10.** Suppose \( f : \mathbb{R}^n \to \mathbb{R} \) and \( h : \mathbb{R}^n \to \mathbb{R}^p \) are continuously differentiable, and \( x^* \in \mathbb{R}^n \), \( \lambda^* \in \mathbb{R}^p \) satisfy

\[
\begin{align*}
\nabla f(x^*) &= \nabla h(x^*) \lambda^*, \\
h(x^*) &\geq 0, \\
\lambda^* &\geq 0, \\
\lambda_i^* h_i(x^*) &= 0 \text{ for all } i = 1,2,\ldots,p, \\
z \cdot \nabla^2 \ell(x^*; \lambda^*) z &> 0 \text{ for all } z \in D(x^*, \lambda^*), z \neq 0.
\end{align*}
\]
Then \( x^* \) is a strict local minimizer of
\[
\min f(x) \\
\text{s.t. } h(x) \geq 0,
\]
and \( \lambda^* \) is a corresponding Lagrange multiplier.

The proof will be omitted.

### 8.2.3 Convex programs

Unlike the case of equality constraints, many nonlinear inequality constraints give rise to a convex feasible set. This is because of the following theorem:

**Theorem 8.11.** Suppose \( h : \mathbb{R}^n \to \mathbb{R} \) is a convex function. Then
\[
C = \{ x \in \mathbb{R}^n : h(x) \leq 0 \}
\]
is a convex set.

**Proof:** Suppose \( x, y \in C \) and \( \alpha, \beta \) are positive real numbers such that \( \alpha + \beta = 1 \). Then, by definition,
\[
h(x) \leq 0, \quad h(y) \leq 0
\]
and, by the convexity of \( h \),
\[
h(\alpha x + \beta y) \leq \alpha h(x) + \beta h(y).
\]
Since \( \alpha, \beta \geq 0 \), (8.15) and (8.16) imply that
\[
h(\alpha x + \beta y) \leq 0,
\]
that is, \( \alpha x + \beta y \in C \). Therefore \( C \) is convex. QED

The following theorem is also needed:

**Theorem 8.12.** Let \( A \) be any set and suppose that \( C_a \) is a convex set for each \( a \in A \). Then
\[
C = \bigcap_{a \in A} C_a
\]
is convex.

**Proof:** Suppose \( x, y \in C \) and \( \alpha, \beta \) are positive real numbers such that \( \alpha + \beta = 1 \). Then, by definition of \( C \), \( x, y \in C_a \) for all \( a \in A \). By the convexity of each \( C_a \), it follows that \( \alpha x + \beta y \in C_a \) for all \( a \in A \). But then
\[
\alpha x + \beta y \in \bigcap_{a \in A} C_a = C.
\]
8.2. Optimality conditions for inequality-constrained nonlinear programs

QED

Putting the previous two results together yields the following corollary:

**Corollary 8.13.** Suppose \( h : \mathbb{R}^n \to \mathbb{R}^p \) has the property that each component \( h_i \) is a convex function. Then \[ \{ x \in \mathbb{R}^n : h(x) \leq 0 \} \]
is a convex set.

For convenience, when treating inequality-constrained NLPs, I have written the inequality constraints as \( h(x) \geq 0 \). (This is convenient because it causes the Lagrange multipliers to be nonnegative.) Therefore, the NLP

\[
\begin{align*}
\min f(x) \\
\text{s.t. } h(x) \geq 0
\end{align*}
\]

is a convex program if \( f \) is convex and each component of \(-h\) is convex, that is, if each component of \( h \) is concave. In the interests of conciseness, a vector-valued function \( h \) will be called convex (or concave) if each of its components is convex (or concave).

Assuming, then, that \( f \) is convex and \( h \) is concave, Theorem 6.14 shows that every local minimizer of (8.17–8.18) is in fact a global minimizer. I now prove the stronger result that the first-order necessary conditions are actually sufficient for an inequality-constrained convex program. Suppose that \( x^* \), \( \lambda^* \) satisfy the first-order necessary conditions for (8.17–8.18). Then \( x^* \) is a stationary point of the Lagrangian \( \ell(\cdot; \lambda^*) \). Since

\[
\ell(x; \lambda^*) = f(x) - \lambda^* \cdot h(x) = f(x) + \sum_{i=1}^{p} \lambda^*_i (-h_i(x)),
\]

the following theorem shows that \( \ell(\cdot, \lambda^*) \) is convex:

**Theorem 8.14.** If \( f_i : \mathbb{R}^n \to \mathbb{R} \) is convex for \( i = 1, 2, \ldots, k \) and \( c_i \) is nonnegative for \( i = 1, 2, \ldots, k \), then \( f : \mathbb{R}^n \to \mathbb{R} \) defined by

\[
f = \sum_{i=1}^{k} c_i f_i
\]

is also convex.

The proof is immediate from the definition of convexity and will be omitted. Since \( \ell(\cdot; \lambda^*) \) is convex and \( x^* \) is a stationary point, it follows that

\[
\ell(x^*; \lambda^*) \leq \ell(x; \lambda^*) \text{ for all } x \in \mathbb{R}^n.
\]

But \( f(x^*) = \ell(x^*) \) by the complementarity condition and, since \( \lambda^* \geq 0 \) and \( h(x) \geq 0 \) for all feasible \( x \),

\[
\ell(x; \lambda^*) \leq f(x) \text{ for all feasible } x.
\]
Then, for any feasible $x$,
\[ f(x^*) = \ell(x^*; \lambda^*) \leq \ell(x; \lambda^*) \leq f(x), \]
which shows that $x^*$ is a global minimizer of the convex program.

### 8.2.4 Proof of Farkas’s lemma

There are many proofs of Farkas’s lemma in the literature; the reader may consult Dax [4] or Broyden [2] for two recent proofs based on elementary ideas. If $A$ is any $n \times p$ matrix and $c \in \mathbb{R}^n$, then Farkas’s lemma states that exactly one of the following two problems has a solution:

\[ A^T x \geq 0, \quad c \cdot x < 0, \]
\[ c = A\lambda, \quad \lambda \geq 0. \tag{8.19} \tag{8.20} \]

If $x \in \mathbb{R}^n$ satisfies (8.19) and $\lambda \in \mathbb{R}^p$ satisfies (8.20), then
\[ c \cdot x = (A\lambda) \cdot x = \lambda \cdot (A^T x) \geq 0 \]
(since $\lambda \geq 0$, $A^T x \geq 0$). But this contradicts that $c \cdot x < 0$. Thus it is not possible to solve both (8.19) and (8.20).

The constrained least-squares problem

\[ \min ||A\lambda - c||^2 \]
\[ \text{s.t. } \lambda \geq 0 \]

will be used to show that one of (8.19) or (8.20) must have a solution.

Obviously (8.20) has a solution if and only if (8.21–8.22) has a zero-residual solution, that is, a solution $\lambda$ with $A\lambda = c$.

The following lemma gives the optimality conditions for (8.21–8.22). These conditions are precisely the first-order conditions described above, and they are necessary and sufficient because (8.21–8.22) is a convex problem. However, so as not to engage in circular reasoning (after all, Farkas’s lemma was used above to derive the optimality conditions), the necessity and sufficiency of these conditions are proved directly.

**Theorem 8.15.** Suppose $A \in \mathbb{R}^{n \times p}$ and $c \in \mathbb{R}^n$. Then $\lambda$ solves (8.21–8.22) if and only if

\[ \lambda \geq 0, \]
\[ A^T r \geq 0, \text{ where } r = A\lambda - c, \]
\[ \lambda_i (A^T r)_i = 0, \quad i = 1, 2, \ldots, p. \]
8.2. Optimality conditions for inequality-constrained nonlinear programs

**Proof:** Suppose first that \( \lambda \) solves (8.21–8.22), let \( a^{(i)} \) denote the \( i \)th column of \( A \), and let \( e^{(i)} \) denote the \( i \)th standard basis vector (that is, the \( i \)th column of the \( p \times p \) identity matrix). Define \( f_i : \mathbb{R} \to \mathbb{R} \) by

\[
f_i(\theta) = \|A(\lambda + \theta e^{(i)}) - c\|^2 = \|\theta a^{(i)} - (A\lambda - c)\|^2 = \|\theta a^{(i)} - r\|^2.
\]

Clearly \( \theta = 0 \) is the solution of

\[
\min f_i(\theta) \\
\text{s.t. } \lambda_i + \theta \geq 0.
\]

If \( \lambda_i > 0 \), then \( f_i'(0) = 0 \) must hold; otherwise, \( f_i'(0) \geq 0 \) holds. But \( f_i'(0) = 2a^{(i)} \cdot r \). Therefore, for each \( i = 1, 2, \ldots, p \), \( a^{(i)} \cdot r \geq 0 \), that is, \( A^T r \geq 0 \). Moreover, the complementarity condition \( \lambda_i (A^T r)_i = 0, i = 1, 2, \ldots, p \), holds. This proves the necessity of the proposed optimality conditions.

Suppose, on the other hand, that \( \lambda \) satisfies the given conditions. If \( \bar{\lambda} \geq 0 \) and \( u = \bar{\lambda} - \lambda \), then

\[
\lambda_i = 0 \Rightarrow u_i \geq 0, \\
\lambda_i > 0 \Rightarrow (A^T r)_i = 0,
\]

and hence, since \( (A^T r)_i \geq 0 \) for all \( i \),

\[
u \cdot (A^T r) \geq 0.
\]

But then

\[
\|A\bar{\lambda} - c\|^2 = \|A\lambda + Au - c\|^2 \\
= \|A\lambda - c\|^2 + 2Au \cdot (A\lambda - c) + \|Au\|^2 \\
= \|A\lambda - c\|^2 + 2u \cdot (A^T r) + \|Au\|^2 \\
\geq \|A\lambda - c\|^2.
\]

This proves the sufficiency of the given conditions. QED

If the constrained least-square problem (8.21–8.22) has a solution \( \lambda \), then either \( r = A\lambda - c = 0 \) or \( r \neq 0 \). In the first case, there is a solution to (8.20). In the second case, \( A^T r \geq 0 \) and

\[
c \cdot r = (A\lambda - r) \cdot r \quad (\text{since } c = A\lambda - r) \\
= (A\lambda) \cdot r - r \cdot r \\
= \lambda \cdot A^T r - r \cdot r \\
= -r \cdot r \quad (\text{since } \lambda \cdot A^T r = 0 \text{ by the complementarity condition}) \\
< 0.
\]

Then \( x = r \) is a solution of (8.19).

Therefore, to prove Farkas’s lemma, it remains only to show that the constrained least-square problem (8.21–8.22) has a solution.
Lemma 8.16. If $A \in \mathbb{R}^{n \times p}$, $c \in \mathbb{R}^n$, and the set

$$S = \{ A\lambda : y \geq 0 \}$$

is closed, then (8.21–8.22) has a solution.

**Proof:** If (8.21–8.22) has a solution $\lambda$, then $A\lambda$ lies in the ball

$$\mathbb{B}_{||c||}(c) = \{ x \in \mathbb{R}^n : ||x - c|| \leq ||c|| \},$$

(since otherwise $||A0 - c|| < ||A\lambda - c||$). Assuming $S$ is closed, the set

$$S_{[c]} = S \cap \mathbb{B}_{||c||}(c)$$

is closed and bounded, and hence by Weierstrass’s theorem the continuous function $f(x) = ||x - c||$ attains its minimum over $S_{[c]}$, say at $x^* = A\lambda$. But then $\lambda$ is a solution to (8.21–8.22). QED

It remains only to prove that $S$ is closed. The following proof is taken from Ciarlet [3].

**Lemma 8.17.** For any $A \in \mathbb{R}^{n \times p}$, the set

$$S = \{ A\lambda : y \geq 0 \}$$

is closed.

**Proof:** I will prove the result first under the assumption that the columns of $A$ are linearly independent. Suppose $\{s^{(k)}\}$ is a sequence in $S$ converging to some vector $s$. By definition of $S$, there exists a sequence $\{\lambda^{(k)}\}$ in $\mathbb{R}^p$ such that $\lambda^{(k)} \geq 0$ and $s^{(k)} = A\lambda^{(k)}$ for all $k$. Then $||A\lambda^{(k)}||^2$ is uniformly bounded, and

$$||A\lambda^{(k)}||^2 = \lambda^{(k)} : A^TA\lambda^{(k)} \geq \lambda_{\text{min}}(A^TA)||\lambda^{(k)}||^2. \quad (8.23)$$

Since $A$ has full rank by assumption, $A^TA$ is positive definite and so $\lambda_{\text{min}}(A^TA)$ is positive. Therefore, (8.23) implies that $||\lambda^{(k)}||$ is uniformly bounded. Without loss of generality, I can assume that $\lambda^{(k)} \to \lambda$ for some $\lambda$. Since $\lambda^{(k)} \geq 0$ for all $k$, it follows that $\lambda \geq 0$, and

$$A\lambda^{(k)} \to s, \quad A\lambda^{(k)} \to A\lambda$$

both hold. Therefore, $s = A\lambda$, $\lambda \geq 0$, and the result is proved in this case.

I will now show that if the columns of $A$ are linearly dependent, then $S$ can be expressed as a finite union of sets of the form $\{By : y \geq 0\}$, where $B \in \mathbb{R}^{n \times k}$ has linearly independent columns. Then, by the above reasoning, each of these sets is closed, and a finite union of closed sets is closed.

Let the columns of $A$ be denoted by $a^{(1)}, a^{(2)}, \ldots, a^{(p)}$. By assumption, there exist numbers $\mu_1, \mu_2, \ldots, \mu_p$, not all zero, such that

$$\sum_{i=1}^p \mu_i a^{(i)} = 0.$$
Clearly, multiplying by \(-1\) is necessary, the scalars \(\mu_1, \mu_2, \ldots, \mu_p\) can be chosen so that at least one of the numbers is negative. Defining \(J = \{i : \mu_i < 0\}\) and, for any \(\lambda \geq 0\),
\[
t = \min \left\{ \frac{-\lambda_i}{\mu_i} : i \in J \right\},
\]
it follows that
\[
A\lambda = \sum_{i=1}^{p} \lambda_i a^{(i)} = \sum_{i=1}^{p} (\lambda_i + t\mu_i) a^{(i)},
\]
and, by construction, \(\lambda_i + t\mu_i\) is nonnegative for each \(i\). Moreover, for at least one value of \(j\), \(\lambda_j + t\mu_j = 0\), and therefore \(A\lambda\) can be expressed using at most \(p - 1\) of the vectors \(a^{(1)}, a^{(2)}, \ldots, a^{(p)}\). This is true for every \(\lambda \geq 0\), and therefore
\[
S = \bigcup_{j=1}^{p} \left\{ \sum_{i=1}^{p} \lambda_i a^{(i)} : \lambda_i \geq 0 \text{ for all } i \neq j \right\}.
\]
This argument can be applied to each set
\[
\left\{ \sum_{i=1}^{p} \lambda_i a^{(i)} : \lambda_i \geq 0 \text{ for all } i \neq j \right\}
\]
such that \(\{a^{(i)} : i \neq j\}\) is linearly dependent, and then again as necessary, until finally \(S\) is expressed in terms of a finite number of closed sets. QED