Chapter 7

Sequential unconstrained minimization techniques

7.1 The quadratic penalty method for equality-constrained optimization

The most straightforward methods for solving a constrained optimization problem convert it to a sequence of unconstrained problems whose solutions converge to the desired solution. The idea of the quadratic penalty method is to add to the objective function a term that penalizes infeasibility. The quadratic penalty function for

$$\begin{align*}
\min f(x) \\
\text{s.t. } g(x) = 0
\end{align*}$$

is

$$Q(x; \mu) = f(x) + \frac{1}{2\mu}||g(x)||^2,$$

where $\mu$ is a positive scalar called the penalty parameter. Solving

$$\min_x Q(x; \mu)$$

yields a solution $x^*_\mu$. Since smaller values of $\mu$ make any given infeasible point $x$ less attractive when minimizing $Q(x; \mu)$, it is hoped that $x^*_\mu$ approaches the feasible set and hence that

$$x^*_\mu \to x^* \text{ as } \mu \to 0,$$

where $x^*$ is a solution of (7.1).

A minimizer $x^*_\mu$ of $Q(\cdot; \mu)$ must satisfy

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

Since

$$\nabla Q(x^*_\mu; \mu) = \nabla f(x^*_\mu) + \frac{1}{\mu} \nabla g(x^*_\mu) g(x^*_\mu),$$

125
$x_{i*}$ satisfies
\[
\nabla f(x_{i*}) = \nabla g(x_{i*}) \left( -\frac{1}{\mu} g(x_{i*}) \right).
\]
If $x^*$ is a minimizer of the NLP (7.1) and $\lambda^*$ is the Lagrange multiplier corresponding to $x^*$, then
\[
\nabla f(x^*) = \nabla g(x^*) \lambda^*.
\]
These formulas suggest that, if $x_{i*} \to x^*$, then
\[
-\frac{1}{\mu} g(x_{i*}) \to \lambda^* \text{ as } \mu \to 0.
\]
This can be proved under certain mild conditions.

**Example 7.1.** As an example, consider the NLP

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

where $f : \mathbb{R}^2 \to \mathbb{R}$ and $g : \mathbb{R}^2 \to \mathbb{R}$ are defined by $f(x) = (x_1 - 1)^2 + 2(x_2 - 2)^2$ and $g(x) = x_1^2 + x_2^2 - 1$, respectively. The contours of $f$ and the feasible set (the unit circle) are shown in Figure 7.1.

![Figure 7.1](image-url)

**Figure 7.1.** The contours of $f$ and the feasible set determined by $g(x) = 0$ (see Example 7.1). The feasible set is the unit circle, shown as the dashed curve. The constrained minimizer is indicated by an asterisk.
### Table 7.1. The results from Example 7.1.

| $\mu$  | $x_{\mu}^*$         | $|g(x_{\mu}^*)|$ | $-\frac{1}{\mu}g(x_{\mu}^*)$ |
|--------|----------------------|------------------|-------------------------------|
| $10^{-1}$ | (0.34798, 1.0326)   | $1.8737 \cdot 10^{-1}$ | $-1.8737$ |
| $10^{-2}$ | (0.31586, 0.96015)  | $2.1660 \cdot 10^{-2}$ | $-2.1660$ |
| $10^{-3}$ | (0.31234, 0.95180)  | $3.4560 \cdot 10^{-3}$ | $-3.4560$ |
| $10^{-4}$ | (0.31164, 0.95038)  | $3.3540 \cdot 10^{-4}$ | $-3.3540$ |
| $10^{-5}$ | (0.31148, 0.95023)  | $2.2095 \cdot 10^{-5}$ | $-2.2095$ |

A straightforward calculation shows that the global minimizer and corresponding Lagrange multiplier of the NLP are

$$x^* \doteq \begin{bmatrix} 0.31157 \\ 0.95022 \end{bmatrix}, \quad \lambda^* \doteq -2.2095.$$

The penalty method with $\mu = 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}$ yields the results in Table 7.1. These results exhibit the expected convergence.

One difficulty in achieving the results described above is that (7.1) and $Q(\cdot; \mu)$ can have many local minimizers. To develop a satisfactory theory, it is necessary to assume that the minimizers $x_{\mu}^*$ that are computed for various values of $\mu$ are consistent in that they are all related to the same solution $x^*$ of (7.1). One way to do this is to consider a specific strict local minimizer $x^*$ of (7.1) and to choose a neighborhood $N$ of $x^*$ in which $x^*$ is the only minimizer of the NLP. Below it is shown that, for $\mu$ sufficiently small, $Q(\cdot; \mu)$ has a unique minimizer in $N$, which will be designated as $x_{\mu}^*$.

#### 7.1.1 Analysis of the quadratic penalty method

**Theorem 7.2.** Suppose that

$$\min f(x)$$

$$s.t. g(x) = 0,$$

where $f: \mathbb{R}^n \to \mathbb{R}$ and $g: \mathbb{R}^n \to \mathbb{R}^m$, has a strict local minimizer $x^*$, and let $N$ be a neighborhood of $x^*$ such that $x^*$ is the unique solution of the NLP, subject to the further constraint that $x \in N$. Suppose further that $x_{\mu}^*$ is the unique minimizer of

$$Q(x; \mu) = f(x) + \frac{1}{2\mu}||g(x)||^2$$

over $N$ for all $\mu$ sufficiently small, say $0 < \mu \leq \hat{\mu}$. Then $0 < \mu_2 < \mu_1 \leq \hat{\mu}$ implies

1. $Q(x_{\mu_2}^*; \mu_2) \geq Q(x_{\mu_1}^*; \mu_1)$;
2. $||g(x_{\mu_2}^*)|| \leq ||g(x_{\mu_1}^*)||$;
3. $f(x_{\mu_2}^*) \geq f(x_{\mu_1}^*)$. 

Proof: First of all,

\[ \mu_2 < \mu_1 \Rightarrow Q(x^*_\mu_2; \mu_2) \geq Q(x^*_\mu_1; \mu_1) \geq Q(x^*_\mu_1; \mu_1), \]

which yields the first conclusion. Second, the optimality of \( x^*_\mu_1 \) and \( x^*_\mu_2 \) implies the inequalities

\[ Q(x^*_\mu_1; \mu_1) \leq Q(x^*_\mu_2; \mu_1), \]
\[ Q(x^*_\mu_2; \mu_2) \leq Q(x^*_\mu_1; \mu_2), \]

that is,

\[ f(x^*_\mu_1) + \frac{1}{2\mu_1}||g(x^*_\mu_2)||^2 \leq f(x^*_\mu_2) + \frac{1}{2\mu_1}||g(x^*_\mu_2)||^2; \]
\[ f(x^*_\mu_2) + \frac{1}{2\mu_2}||g(x^*_\mu_2)||^2 \leq f(x^*_\mu_1) + \frac{1}{2\mu_2}||g(x^*_\mu_1)||^2. \]

Adding these two inequalities and rearranging yields

\[ \frac{1}{2} \left( \frac{1}{\mu_1} - \frac{1}{\mu_2} \right) (||g(x^*_\mu_1)||^2 - ||g(x^*_\mu_2)||^2) \leq 0. \]

Since \( 1/\mu_1 - 1/\mu_2 < 0 \), this is possible only if

\[ ||g(x^*_\mu_1)|| \geq ||g(x^*_\mu_2)||, \]

which is the second conclusion. Finally,

\[ Q(x^*_\mu_1; \mu_1) \geq Q(x^*_\mu_1; \mu_1) \]
\[ \Rightarrow f(x^*_\mu_2) + \frac{1}{2\mu_1}||g(x^*_\mu_2)||^2 \geq f(x^*_\mu_1) + \frac{1}{2\mu_1}||g(x^*_\mu_1)||^2 \]
\[ \Rightarrow f(x^*_\mu_2) - f(x^*_\mu_1) \geq \frac{1}{2\mu_1} (||g(x^*_\mu_1)||^2 - ||g(x^*_\mu_2)||^2) \geq 0, \]

which yields the final conclusion. QED

7.1.2 The implicit function theorem

The implicit function theorem is a basic result from analysis that is critical for analyzing the convergence of the quadratic penalty method and similar algorithms.

Theorem 7.3 (The implicit function theorem). Suppose \( F : \mathbb{R}^n \times \mathbb{R}^k \to \mathbb{R}^n \) is continuously differentiable and that \( x^* \in \mathbb{R}^n, y^* \in \mathbb{R}^k \) satisfy

\[ F(x^*, y^*) = 0. \]

Let \( J^* \) be the Jacobian of \( F \) with respect to \( x \), evaluated at \((x^*, y^*)\). If \( J^* \) is nonsingular, then there exists a neighborhood \( M \) of \( y^* \), a neighborhood \( N \) of \( x^* \), and a function \( f : M \to N \) such that \( f(y^*) = x^* \) and \( f(y) \) is the only point in \( N \) satisfying \( F(f(y), y) = 0 \) for all \( y \in M \).
Moreover, \( f \) is continuously differentiable.

In other words, under the given conditions, it is possible to solve the system of \( n \) equation for the first \( n \) variables in terms of the other \( k \) variables, at least locally around \( y^* \).

### 7.1.3 Convergence of the quadratic penalty method

Convergence of the quadratic penalty method will be obtained by applying the the implicit function theorem to the first-order optimality conditions of the NLP (7.1). This requires that the condition

\[
\nabla Q(x; \mu) = 0
\]

be written in a form similar to the extended system (7.1). Introducing the variable \( \lambda \), (7.3) is equivalent to

\[
\nabla f(x) - \nabla g(x)\lambda = 0, \\
-g(x) - \mu \lambda = 0.
\]

I now define \( F : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R} \to \mathbb{R}^n \times \mathbb{R}^m \) by

\[
F(x, \lambda, \mu) = \begin{bmatrix}
\nabla f(x) - \nabla g(x)\lambda \\
-g(x) - \mu \lambda
\end{bmatrix}.
\]

If \( x^* \) is a local minimizer of NLP (7.1) and \( \lambda^* \) is the associated Lagrange multiplier, then

\[
F(x^*, \lambda^*, 0) = 0.
\]

By the implicit function theorem, \( F(x, \lambda, \mu) = 0 \) can be solved for \( x^*, \lambda^* \) as functions of \( \mu \), provided the Jacobian of \( F \) with respect to \( (x, \lambda) \) is nonsingular at \( (x^*, \lambda^*, 0) \).

The Jacobian of \( F \) with respect to \( (x, \lambda) \) is

\[
J^* = \begin{bmatrix}
\nabla^2 \ell(x^*; \lambda^*) - \nabla g(x^*) \\
\n\nabla g(x^*)^T
\end{bmatrix},
\]

(7.4)

which is a symmetric matrix. At \( (x^*, \lambda^*, 0) \), the Jacobian is

\[
J^* = \begin{bmatrix}
\nabla^2 \ell(x^*; \lambda^*) \\
\n\nabla g(x^*)^T
\end{bmatrix}. 
\]

If \( x^* \) is a nonsingular point (that is, a regular point that satisfies the second-order sufficient conditions for a local minimizer), then \( J^* \) is nonsingular. To see this, suppose \( (z, w) \in \mathbb{R}^n \times \mathbb{R}^m \) satisfies \( J^*(z, w) = 0 \). Then

\[
\nabla^2 \ell(x^*; \lambda^*) z - \nabla g(x^*) w = 0, \\
\n\nabla g(x^*)^T z = 0.
\]

(7.5)

(7.6)

Taking the dot product of both sides of (7.5) with \( z \) yields

\[
z \cdot \nabla^2 \ell(x^*; \lambda^*) z - z \cdot \nabla g(x^*) w = 0 \\
\Rightarrow z \cdot \nabla^2 \ell(x^*; \lambda^*) z = (\nabla g(x^*)^T z) \cdot w = 0 \\
\Rightarrow z \cdot \nabla^2 \ell(x^*; \lambda^*) z = 0
\]
(using the fact that \( z \in \mathcal{N}(\nabla g(x^*)^T) \)). But then
\[
z \in \mathcal{N}(\nabla g(x^*)^T), \; z \cdot \nabla^2 \ell(x^*; \lambda^*) z = 0,
\]
and therefore the second-order sufficient condition implies that \( z = 0 \). Equation (7.5) then implies that
\[
\nabla g(x^*) w = 0,
\]
which shows that \( w = 0 \) (since \( x^* \) is a regular point and hence the columns of \( \nabla g(x^*) \) are linearly independent). Since the only vector \((z, w)\) satisfying \( J^*(z, w) = 0 \) is the zero vector, \( J^* \) must be nonsingular.

Assuming that \( x^* \) is a nonsingular point, the implicit function theorem yields continuous functions \( x^* = x^*(\mu) \) and \( \lambda^* = \lambda^*(\mu) \), defined for all \( \mu \) sufficiently close to 0, such that
\[
F(x^*(\mu), \lambda^*(\mu), \mu) = 0.
\]
The continuity implies that \( x^*(\mu) \to x^* \) and \( \lambda^*(\mu) \to \lambda^* \) as \( \mu \to 0 \). By definition,
\[
\lambda^*(\mu) = -\frac{1}{\mu} g(x^*(\mu)),
\]
so
\[
-\frac{1}{\mu} g(x^*(\mu)) \to \lambda^* \text{ as } \mu \to 0.
\]

This suggests the following theorem:

**Theorem 7.4.** Suppose \( f : \mathbb{R}^n \to \mathbb{R} \) and \( g : \mathbb{R}^n \to \mathbb{R}^m \) are twice continuously differentiable and that \( x^* \) is local minimizer of
\[
\min f(x) \quad \text{s.t. } g(x) = 0
\]
with Lagrange multiplier \( \lambda^* \). If \( x^* \) is a nonsingular point, then there exists a neighborhood \( N \) of \( x^* \) such that, for all \( \mu \) sufficiently small,
\[
Q(x; \mu) = f(x) + \frac{1}{2\mu}||g(x)||^2
\]
has a unique local minimizer \( x^*_\mu \) in \( N \). Moreover, \( x^*_\mu \) depends continuously on \( \mu \), with
\[
x^*_\mu \to x^* \text{ as } \mu \to 0,
\]
and
\[
-\frac{1}{\mu} g(x^*_\mu) \to \lambda^* \text{ as } \mu \to 0.
\]
The quadratic penalty method for equality-constrained optimization

**Proof:** The above reasoning shows that $Q(\cdot; \mu)$ has a unique stationary point $x^*_\mu$ in $N$. It remains only to show that $x^*_\mu$ is actually a local minimizer of $Q(\cdot; \mu)$. The Hessian of $Q(\cdot; \mu)$ at $x^*_\mu$ can be written as

\[
\nabla^2 Q(x^*_\mu; \mu) = \nabla^2 f(x^*_\mu) + \frac{1}{\mu} \nabla g(x^*_\mu) \nabla g(x^*_\mu)^T + \frac{1}{\mu} \sum_{i=1}^m g_i(x^*_\mu) \nabla^2 g_i(x^*_\mu)
\]

\[
= \nabla^2 \ell(x^*_\mu; \lambda^*_\mu) + \frac{1}{\mu} \nabla g(x^*_\mu) \nabla g(x^*_\mu)^T,
\]

where

\[
\lambda^*_\mu = -\frac{1}{\mu} g(x^*_\mu).
\]

To prove that $\nabla^2 Q(x^*_\mu; \mu)$ is positive definite for all $\mu$ sufficiently small, I will argue by contradiction. If the desired conclusion does not hold, there exist sequences $\{\mu_k\}$ and $\{y^{(k)}\}$ such that

\[
\mu_k \to 0 \text{ as } k \to \infty,
\]

\[
||y^{(k)}|| = 1 \text{ for all } k,
\]

\[
y^{(k)} \cdot \nabla^2 Q(x^*_{\mu_k}; \mu_k) y^{(k)} \leq 0 \text{ for all } k.
\]

Since $\{y^{(k)}\}$ is bounded, a subsequence must converge to some $y \in \mathbb{R}^n$. Without loss of generality, I might as well assume that $y^{(k)} \to y$; then $||y|| = 1$ because $||y^{(k)}|| = 1$ for all $k$.

Since

\[
\nabla^2 \ell(x^*_\mu; \lambda^*_\mu) \to \nabla^2 \ell(x^*; \lambda^*) \text{ as } \mu \to 0,
\]

it follows that

\[
y^{(k)} \cdot \nabla^2 \ell(x^*_{\mu_k}; \lambda^*_{\mu_k}) y^{(k)}
\]

must be uniformly bounded, say

\[
-M \leq y^{(k)} \cdot \nabla^2 \ell(x^*_{\mu_k}; \lambda^*_{\mu_k}) y^{(k)} \leq M \text{ for all } k.
\]

It then follows that

\[
0 \leq \frac{1}{\mu_k} y^{(k)} \cdot \nabla g(x^*_{\mu_k}) \nabla g(x^*_{\mu_k})^T y^{(k)} \leq M \text{ for all } k,
\]

that is,

\[
0 \leq y^{(k)} \cdot \nabla g(x^*_{\mu_k}) \nabla g(x^*_{\mu_k})^T y^{(k)} \leq \mu_k M \text{ for all } k.
\]

This shows that

\[
y^{(k)} \cdot \nabla g(x^*_{\mu_k}) \nabla g(x^*_{\mu_k})^T y^{(k)} \to 0 \text{ as } k \to \infty.
\]

But continuity implies that

\[
y^{(k)} \cdot \nabla g(x^*_{\mu_k}) \nabla g(x^*_{\mu_k})^T y^{(k)} \to y \cdot \nabla g(x^*) \nabla g(x^*)^T y \text{ as } k \to \infty.
\]
Therefore

\[ y \cdot \nabla g(x^*) \nabla g(x^*)^T y = 0 \]

must hold, which implies that

\[ \nabla g(x^*)^T y = 0 \]

or, equivalently, that \( y \in \mathcal{N}(\nabla g(x^*)) \).

It now follows, since \( x^* \) is by assumption a nonsingular point, that

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

This implies that

\[ y^{(k)} \cdot \nabla^2 \ell(x_{\mu_k}^*; \lambda_{\mu_k}^*) y^{(k)} > 0 \]

for all \( k \) sufficiently large. But then

\[ y^{(k)} \cdot \nabla^2 Q(x_{\mu_k}^*; \mu_k) y^{(k)} \geq y^{(k)} \cdot \nabla^2 \ell(x_{\mu_k}^*; \lambda_{\mu_k}^*) y^{(k)} > 0 \]

for all \( k \) sufficiently large, a contradiction. This contradiction shows that \( \nabla^2 Q(x_{\mu}^*; \mu) \) must be positive definite for all \( \mu \) sufficiently small. QED

### 7.1.4 Ill-conditioning of the quadratic penalty method

In spite of the simplicity and strong convergence properties of the quadratic penalty method, it suffers from a significant disadvantage: The Hessian \( Q(x_{\mu}^*; \mu) \) gets arbitrarily ill-conditioned as \( \mu \to 0 \), which implies that Newton’s method and similar methods for minimizing \( Q(\cdot; \mu) \) suffer from increasingly severe numerical difficulties as \( \mu \) becomes small. Since it is necessary to take \( \mu \) small in order to obtain convergence, this ill-conditioning is unavoidable.

The ill-conditioning of \( \nabla^2 Q(x_{\mu}^*; \mu) \) is easy to demonstrate. It is necessary only to produce two unit vectors \( u, v \) such that

\[ u \cdot Q(x_{\mu}^*; \mu) u \] \[ v \cdot Q(x_{\mu}^*; \mu) v \]

differ greatly in magnitude. Since

\[ \nabla^2 Q(x_{\mu}^*; \mu) = \nabla^2 \ell(x_{\mu}^*; \lambda_{\mu}^*) + \frac{1}{\mu} \nabla g(x_{\mu}^*) \nabla g(x_{\mu}^*)^T , \]

it suffices to take \( u \in \mathcal{N}(\nabla g(x_{\mu}^*)^T) \) and \( v \in \mathcal{R}(\nabla g(x_{\mu}^*)) = \mathcal{N}(\nabla g(x_{\mu}^*)^T) \). The details are left to the reader.

### 7.2 The augmented Lagrangian method for equality-constrained optimization

The augmented Lagrangian method for solving

\[
\min \ f(x) \quad \text{s.t.} \ g(x) = 0
\]  

(7.7)
is similar to the quadratic penalty approach. However, instead of adding the penalty term to the objective function \( f \), it is added to the Lagrangian \( \ell \), resulting in the augmented Lagrangian
\[
L(x; \lambda; \mu) = f(x) - \lambda \cdot g(x) + \frac{1}{2\mu} ||g(x)||^2.
\]
(7.8)

Since \( f \) and \( \ell \) (for any \( \lambda \)) agree on the feasible set \( g(x) = 0 \), the basic idea is the same as before: a small value of \( \mu \) forces the minimizer(s) of \( L \) to lie close to the feasible set, while, at the same time, values of \( x \) that reduce \( f \) are preferred. The advantage of the augmented Lagrangian approach is that by including an explicit estimate of the Lagrange multiplier, it is not necessary to decrease \( \mu \) to zero in order to obtain convergence, and so various numerical problems are avoided.

As usual, it will be assumed that \( x^* \) is a local minimizer of (7.7) and \( \lambda^* \) is the corresponding Lagrange multiplier. If \( x^*_\mu \) is a minimizer of \( L \) for a given \( \mu \) and \( \lambda \), then
\[
\nabla L(x^*_\mu; \lambda; \mu) = \nabla f(x^*_\mu) - \nabla g(x^*_\mu) \lambda + \frac{1}{\mu} \nabla g(x^*_\mu) g(x^*_\mu) = 0,
\]
that is,
\[
\nabla f(x^*_\mu) = \nabla g(x^*_\mu) \left( \lambda - \frac{1}{\mu} g(x^*_\mu) \right).
\]

Since
\[
\nabla f(x^*) = \nabla g(x^*) \lambda^*,
\]
it follows that
\[
\lambda - \frac{1}{\mu} g(x^*_\mu)
\]
should be a better estimate of \( \lambda^* \) than is \( \lambda \). This suggests the following strategy:

Choose an initial estimate \( \lambda^{(0)} \) of \( \lambda^* \) and some \( \mu > 0 \).

For \( k = 1, 2, 3, \ldots \)

Define \( x^{(k)} \) to be a minimizer of \( L(\cdot; \lambda^{(k-1)}; \mu) \).
Define \( \lambda^{(k)} = \lambda^{(k-1)} - (1/\mu) g(x^{(k)}) \).

Under certain conditions (specified below), it is possible to prove the following:

1. If \( \mu \) is sufficiently small, then \( L(\cdot; \lambda^*; \mu) \) has a local minimizer at \( x^* \) and, for all \( \lambda \) sufficiently close to \( \lambda^* \), \( L(\cdot; \lambda; \mu) \) has a unique local minimizer in a neighborhood of \( x^* \).

2. The sequence \( (x^{(k)}, \lambda^{(k)}) \) converges to \( (x^*, \lambda^*) \) as \( k \to \infty \).

Before developing the theory, I apply the augmented Lagrangian method to the following example, which was treated earlier (in Example 7.1) by the quadratic penalty method.
<table>
<thead>
<tr>
<th>$k$</th>
<th>$x^{(k)}$</th>
<th>$g(x^*_p)$</th>
<th>$\lambda^{(k)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(0.34798, 1.0326)</td>
<td>$1.8737 \cdot 10^{-1}$</td>
<td>$-1.8737$</td>
</tr>
<tr>
<td>2</td>
<td>(0.31710, 0.96303)</td>
<td>$2.7985 \cdot 10^{-2}$</td>
<td>$-2.1535$</td>
</tr>
<tr>
<td>3</td>
<td>(0.31249, 0.95237)</td>
<td>$4.6521 \cdot 10^{-3}$</td>
<td>$-2.2000$</td>
</tr>
<tr>
<td>4</td>
<td>(0.31173, 0.95659)</td>
<td>$7.8649 \cdot 10^{-4}$</td>
<td>$-2.2079$</td>
</tr>
<tr>
<td>5</td>
<td>(0.31160, 0.95028)</td>
<td>$1.3335 \cdot 10^{-4}$</td>
<td>$-2.2093$</td>
</tr>
<tr>
<td>6</td>
<td>(0.31158, 0.95023)</td>
<td>$2.2616 \cdot 10^{-5}$</td>
<td>$-2.2095$</td>
</tr>
</tbody>
</table>

**Table 7.2.** The results from Example 7.5.

**Example 7.5.** Let $f : \mathbb{R}^2 \to \mathbb{R}$ and $g : \mathbb{R}^2 \to \mathbb{R}$ be defined by $f(x) = (x_1 - 1)^2 + 2(x_2 - 2)^2$ and $g(x) = x_1^2 + x_2^2 - 1$, respectively. The global minimizer of $f$, subject to the constraint $g(x) = 0$, and the corresponding Lagrange multiplier are

$$
\begin{align*}
    x^* &= \begin{bmatrix} 0.31157 \\ 0.95022 \end{bmatrix}, \\
    \lambda^* &= -2.2095.
\end{align*}
$$

With $\mu = 10^{-1}$ and $\lambda^{(0)} = 0$, the augmented Lagrangian method yields the results given in Table 7.2. The reader should notice that $x^{(k)} \to x^*$ and $\lambda^{(k)} \to \lambda^*$, in spite of the fact that $\mu$ is held constant at $10^{-1}$.

### 7.2.1 Convergence analysis

The convergence analysis of the augmented Lagrangian method is similar to that of the quadratic penalty method, but significantly more complicated because there are two parameters $\lambda, \mu$ instead of just one. The analysis is based on the system

$$
F(x, \lambda_+; \lambda; \mu) = \begin{bmatrix} \nabla f(x) - \nabla g(x) \lambda_+ \\ -g(x) - \mu (\lambda_+ - \lambda) \end{bmatrix} = 0,
$$

which is to be solved for $(x, \lambda_+)$ in terms of $\lambda$ and $\mu$ as parameters. First of all, assuming as usual that $x^*, \lambda^*$ is a local minimizer-Lagrange multiplier pair,

$$
F(x^*, \lambda^*; \lambda^*; \mu) = \begin{bmatrix} \nabla f(x^*) - \nabla g(x^*) \lambda^* \\ -g(x^*) - \mu (\lambda^* - \lambda^*) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}
$$

for all $\mu > 0$. Moreover, the Jacobian of $F$ (with respect to the variables $x, \lambda_+$), is

$$
J(x, \lambda_+; \lambda; \mu) = \begin{bmatrix} \nabla^2 \ell(x; \lambda_+) - \nabla g(x) \\ -\nabla g(x)^T - \mu I \end{bmatrix}.
$$

Assuming $x^*$ is a nonsingular point of the NLP, the matrix

$$
\begin{bmatrix} \nabla^2 \ell(x^*; \lambda^*) & -\nabla g(x^*) \\ -\nabla g(x^*)^T & 0 \end{bmatrix}
$$
is nonsingular (as I showed previously when discussing the quadratic penalty function),

\[
J(x^*, \lambda^*; \lambda^*; \mu) = \begin{bmatrix}
\nabla^2 \ell(x^*; \lambda^*) - \nabla g(x^*) \\
-\nabla g(x^*)^T \\
-\mu I
\end{bmatrix} \rightarrow \begin{bmatrix}
\nabla^2 \ell(x^*; \lambda^*) - \nabla g(x^*) \\
-\nabla g(x^*)^T \\
0
\end{bmatrix}
\]

as \( \mu \to 0 \). Therefore, there exists \( \tilde{\mu} > 0 \) such that \( J(x^*, \lambda^*; \lambda^*; \mu) \) is nonsingular for all \( \mu \in [0, \tilde{\mu}] \). The implicit function theorem\(^{21}\) then implies that there exists neighborhoods \( M \) of \( \lambda^* \) and \( N \) of \( x^* \) and functions \( x: M \times [0, \tilde{\mu}] \to N, \lambda_+: M \times [0, \tilde{\mu}] \to N \) such that

- \( x(\lambda^*; \mu) = x^*, \lambda_+(\lambda^*; \mu) = \lambda^* \) for all \( \mu \in [0, \tilde{\mu}] \);
- for all \( \lambda \in N, \mu \in [0, \tilde{\mu}] \),

\[
F(x(\lambda; \mu), \lambda_+(\lambda; \mu); \lambda; \mu) = 0.
\]

The functions \( x, \lambda_+ \) satisfy

\[
\nabla f(x(\lambda; \mu)) - \nabla g(x(\lambda; \mu)) \lambda_+(\lambda; \mu) = 0, \quad (7.9)
\]

\[
g(x(\lambda; \mu)) - \mu (\lambda_+(\lambda; \mu) - \lambda) = 0. \quad (7.10)
\]

Solving (7.10) for \( \lambda_+(\lambda; \mu) \) yields

\[
\lambda_+(\lambda; \mu) = \lambda - \frac{1}{\mu} g(x(\lambda; \mu));
\]

substituting this into (7.9) then produces

\[
\nabla f(x(\lambda; \mu)) - \nabla g(x(\lambda; \mu)) \left( \lambda - \frac{1}{\mu} g(x(\lambda; \mu)) \right) = 0.
\]

Rearranging this last equation shows that

\[
\nabla L(x(\lambda; \mu); \lambda; \mu) = 0;
\]

in other words, \( x(\lambda; \mu) \) is a stationary point of \( L(\cdot; \lambda \mu) \) for each \( \lambda \in N \) and each \( \mu \in [0, \tilde{\mu}] \).

Since

\[
\nabla^2 L(x(\lambda; \mu); \lambda; \mu) = \nabla^2 \ell(x(\lambda; \mu); \lambda_+(\lambda; \mu)) + \frac{1}{\mu} \nabla g(x) \nabla g(x)^T,
\]

\(^{21}\)Actually, a different version of the implicit function theorem is needed here. The parameter \( \mu \) varies over a compact set and it is possible to show that there exist neighborhoods \( M \) of \( \lambda^* \) and \( N \) of \( x^* \) such that for all \( \lambda \in M \) and for all \( \mu \in [0, \tilde{\mu}] \), there exist unique \( x(\lambda; \mu) \in N, \lambda_+(\lambda; \mu) \in M \) satisfying

\[
F(x(\lambda; \mu), \lambda_+(\lambda; \mu); \lambda; \mu) = 0.
\]

The ordinary implicit function theorem, as presented earlier, would give a possibly different neighborhoods \( M, N \) for each value of \( \mu \); as stated, it would not guarantee that the same neighborhoods work for all \( \mu \in [0, \tilde{\mu}] \). The improved version of the implicit function theorem can be found in Bertsekas [1], page 12.
and \( x(\lambda; \mu) \rightarrow x^*; \lambda_+(\lambda; \mu) \rightarrow \lambda^* \) as \( \lambda \rightarrow \lambda^* \), it is straightforward to show that
\[
\nabla^2 L(x(\lambda; \mu); \lambda; \mu)
\]
is positive definite for \( \lambda \) sufficiently close to \( \lambda^* \) and for \( \mu \) sufficiently small. (The proof is similar to the case of the quadratic penalty function.) I have therefore proved the following theorem.

**Theorem 7.6.** Suppose \( f : \mathbb{R}^n \rightarrow \mathbb{R} \) and \( g : \mathbb{R}^n \rightarrow \mathbb{R}^m \) are twice continuously differentiable and \( x^* \) is a local minimizer of the NLP
\[
\min f(x)
\]
\[s.t. \; g(x) = 0.
\]
If \( x^* \) is a nonsingular point and \( \lambda^* \) is the corresponding Lagrange multiplier, then there exist \( \bar{\mu} > 0 \), neighborhoods \( M \) of \( \lambda^* \) and \( N \) of \( x^* \), and a function \( x : M \times [0, \bar{\mu}] \rightarrow N \), with the following properties:

1. \( x \) is continuously differentiable;
2. \( x(\lambda^*; \mu) = x^* \) for all \( \mu \in [0, \bar{\mu}] \);
3. \( x(\lambda^*; \mu) \) is the unique local minimizer of \( L(\cdot; \lambda; \mu) \) in \( N \);

According to the previous theorem, if \( \mu \) is sufficiently small and \( \lambda \rightarrow \lambda^* \), then \( x(\lambda; \mu) \rightarrow x^* \). However, since \( \lambda^* \) is unknown, the condition \( \lambda \rightarrow \lambda^* \) cannot be enforced directly. Instead, the augmented Lagrangian method updates \( \lambda \) using the results of the unconstrained minimization: \( \lambda \leftarrow \lambda_+(\lambda; \mu) \). It is necessary to prove, then, that updating \( \lambda \) in this manner produces a sequence of Lagrange multiplier estimates converging to \( \lambda^* \).

Since \( \lambda_+ \) is a continuously differentiable function of \( \lambda \) and \( \lambda_+(\lambda^*; \mu) = \lambda^* \), the fundamental theorem of calculus implies
\[
\lambda_+(\lambda; \mu) = \lambda^* + \int_0^1 \nabla \lambda_+(\lambda^* + t(\lambda - \lambda^*); \mu)T(\lambda - \lambda^*) \; dt.
\]
Using the triangle inequality for integrals, it follows that
\[
\|\lambda_+(\lambda; \mu) - \lambda^*\| \leq \int_0^1 \|\nabla \lambda_+(\lambda^* + t(\lambda - \lambda^*); \mu)T\| \|\lambda - \lambda^*\| \; dt \leq C(\mu)\|\lambda - \lambda^*\|,
\]
where \( C(\mu) \) is an upper bound for \( \|\nabla \lambda_+(\cdot; \mu)T\| \). Similarly,
\[
x(\lambda; \mu) = x^* + \int_0^1 \nabla x(\lambda^* + t(\lambda - \lambda^*); \mu)T(\lambda - \lambda^*) \; dt
\]
and so
\[
\|x(\lambda; \mu) - x^*\| \leq \int_0^1 \|\nabla x(\lambda^* + t(\lambda - \lambda^*); \mu)T\| \|\lambda - \lambda^*\| \; dt \leq D(\mu)\|\lambda - \lambda^*\|,
\]
where $D(\mu)$ is an upper bound for $\|\nabla x(\cdot; \mu)^T\|$.

The functions $x, \lambda_+$ are defined by the equations

$$
\begin{align*}
\nabla f(x(\lambda; \mu)) - \nabla g(x(\lambda; \mu))\lambda_+(\lambda; \mu) &= 0, \\
g(x(\lambda; \mu)) - \mu\lambda_+(\lambda; \mu) - \lambda &= 0.
\end{align*}
$$

Differentiating these equations with respect to $\lambda$ and simplifying the results yields

$$
\begin{align*}
\nabla^2 \ell(x(\lambda; \mu); \lambda_+(\lambda; \mu))\nabla x(\lambda; \mu)^T - \nabla g(x(\lambda; \mu))\nabla \lambda_+(\lambda; \mu)^T &= 0, \\
-\nabla g(x(\lambda; \mu))^T\nabla x(\lambda; \mu)^T - \mu\nabla \lambda_+(\lambda; \mu)^T + \mu I &= 0,
\end{align*}
$$

or

$$
J(x(\lambda; \mu), \lambda_+(\lambda; \mu); \lambda; \mu) \begin{bmatrix} \nabla x(\lambda; \mu)^T \\ \nabla \lambda_+(\lambda; \mu)^T \end{bmatrix} = \begin{bmatrix} 0 \\ -\mu I \end{bmatrix}.
$$

Since $J(x(\lambda; \mu), \lambda_+(\lambda; \mu); \lambda; \mu) \rightarrow J(x^*, \lambda^*; \lambda^*; \mu)$ as $\lambda \rightarrow \lambda^*$, it follows that

$$
\|J(x(\lambda; \mu), \lambda_+(\lambda; \mu); \lambda; \mu)^{-1}\|
$$

is bounded above for all $\lambda$ sufficiently close to $\lambda^*$. Therefore, from

$$
\begin{bmatrix} \nabla x(\lambda; \mu)^T \\ \nabla \lambda_+(\lambda; \mu)^T \end{bmatrix} = \mu J(x(\lambda; \mu), \lambda_+(\lambda; \mu); \lambda; \mu)^{-1} \begin{bmatrix} 0 \\ -I \end{bmatrix},
$$

Therefore, there exists $K > 0$ such that, for all $\mu \in (0, \hat{\mu})$,

$$
\|\nabla x(\lambda; \mu)^T\| \leq \mu K, \quad \|\nabla \lambda_+(\lambda; \mu)^T\| \leq \mu K.
$$

Replacing $C(\mu)$ and $D(\mu)$ above by $\mu K$, it follows that

$$
\|\lambda_+(\lambda; \mu) - \lambda^*\| \leq \mu M\|\lambda - \lambda^*\|,
$$

$$
\|x(\lambda; \mu) - x^*\| \leq \mu M\|\lambda - \lambda^*\|
$$

for all $\mu \in (0, \hat{\mu})$. I have therefore proved the following theorem:

**Theorem 7.7.** Suppose $f : \mathbb{R}^n \rightarrow \mathbb{R}$ and $g : \mathbb{R}^n \rightarrow \mathbb{R}^m$ are twice continuously differentiable and $x^*$ is a local minimizer of the NLP

$$
\begin{align*}
\min f(x) \\
\text{s.t. } g(x) &= 0.
\end{align*}
$$

If $x^*$ is a nonsingular point and $\lambda^*$ is the corresponding Lagrange multiplier, then there exist $\hat{\mu} > 0$, neighborhoods $M$ of $\lambda^*$ and $N$ of $x^*$, and a function $x : M \times [0, \hat{\mu}] \rightarrow N$, such that

1. $x$ is continuously differentiable;
2. $x(\lambda^*; \mu) = x^*$ for all $\mu \in [0, \hat{\mu}]$;
3. $x(\lambda^*; \mu)$ is the unique local minimizer of $L(\cdot; \lambda; \mu)$ in $N$;
Moreover, defining $\lambda_+(\lambda; \mu) = \lambda - \hat{g}(\lambda; \mu)/\mu$ for $\mu \in (0, \hat{\mu})$, there exists a constant $K > 0$ such that

\begin{align*}
\|\lambda_+(\lambda; \mu) - \lambda^*\| &\leq \mu K \|\lambda - \lambda^*\|, \\
\|x(\lambda; \mu) - x^*\| &\leq \mu K \|\lambda - \lambda^*\| 
\end{align*}

(7.13)  
(7.14)

hold for all $\mu \in (0, \hat{\mu}]$.

By taking $\hat{\mu}$ sufficiently small, it follows that $\mu K < 1$ for all $\mu \in [0, \hat{\mu}]$. Therefore, inequalities (7.13) and (7.14) that the augmented Lagrangian method produces sequences $\{x^{(k)}\}$ and $\{\lambda^{(k)}\}$ converging to $x^*$ and $\lambda^*$, provided $\mu$ is sufficiently small (but fixed) and $\lambda^{(0)}$ is sufficiently close to $\lambda^*$. 