Chapter 15
Primal-Dual Methods

This chapter discusses methods that work simultaneously with primal and dual variables, in essence seeking to satisfy the first-order necessary conditions for optimality. The methods employ many of the concepts used in earlier chapters, including those related to active set methods, various first and second order methods, penalty methods, and barrier methods. Indeed, a study of this chapter is in a sense a review and extension of what has been presented earlier.

The first several sections of the chapter discuss methods for solving the standard nonlinear programming structure that has been treated in the Parts II and III of the text. These sections provide alternatives to the methods discussed earlier.

15.1 The Standard Problem

Consider again the standard nonlinear program

$$\begin{aligned} \text{minimize} \quad & f(\mathbf{x}) \\ \text{subject to} \quad & \mathbf{h}(\mathbf{x}) = \mathbf{0}, \ \mathbf{g}(\mathbf{x}) \leqslant \mathbf{0}. \end{aligned} \tag{15.1}$$

Together with the feasibility, the first-order necessary conditions for optimality are, as we know,

$$\begin{aligned} \nabla f(\mathbf{x}) + \boldsymbol{\lambda}^T \nabla \mathbf{h}(\mathbf{x}) + \boldsymbol{\mu}^T \nabla \mathbf{g}(\mathbf{x}) &= \mathbf{0} \\ \boldsymbol{\mu} &\geqslant \mathbf{0} \\ \boldsymbol{\mu}^T \mathbf{g}(\mathbf{x}) &= \mathbf{0} \end{aligned} \tag{15.2}$$

The last requirement is the complementary slackness condition. If it is known which of the inequality constraints is active at the solution, these active constraints can be

D.G. Luenberger, Y. Ye, *Linear and Nonlinear Programming*, International Series in Operations Research & Management Science 228,
DOI 10.1007/978-3-319-18842-3_15

rolled into the equality constraints $\mathbf{h}(\mathbf{x}) = \mathbf{0}$, and the inactive inequalities along with the complementary slackness condition dropped, to obtain a problem with equality constraints only. This indeed is the structure of the problem near the solution.

If in this structure the vector $\mathbf{x}$ is n-dimensional and $\mathbf{h}$ is m-dimensional, then $\boldsymbol{\lambda}$ will also be m-dimensional. The system (15.1) will, in this reduced form, consist of $n + m$ equations and $n + m$ unknowns, which is an indication that the system may be well defined, and hence that there is a solution for the pair $(\mathbf{x}, \boldsymbol{\lambda})$. In essence, primal-dual methods amount to solving this system of equations, and use additional strategies to account for inequality constraints.

In view of the above observation it is natural to consider whether in fact the system of necessary conditions is in fact well conditioned, possessing a unique solution $(\mathbf{x}, \boldsymbol{\lambda})$. We investigate this question by considering a linearized version of the conditions.

A useful and somewhat more generally useful approach is to consider the quadratic program

$$\begin{aligned} &\text{minimize} \quad \frac{1}{2}\mathbf{x}^T\mathbf{Q}\mathbf{x} + \mathbf{c}^T\mathbf{x} \\ &\text{subject to} \quad \mathbf{A}\mathbf{x} = \mathbf{b}, \end{aligned} \tag{15.3}$$

where $\mathbf{x}$ is n-dimensional and $\mathbf{b}$ is m-dimensional.

The first-order conditions for this problem are

$$\begin{aligned} \mathbf{Q}\mathbf{x} + \mathbf{A}^T\boldsymbol{\lambda} + \mathbf{c} &= \mathbf{0} \\ \mathbf{A}\mathbf{x} \qquad\quad - \mathbf{b} &= \mathbf{0}. \end{aligned} \tag{15.4}$$

These correspond to the necessary conditions (15.2) for equality constraints only. The following proposition gives conditions under which the system is nonsingular.

Proposition. *Let* $\mathbf{Q}$ *and* $\mathbf{A}$ *be* $n \times n$ *and* $m \times n$ *matrices, respectively. Suppose that* $\mathbf{A}$ *has rank* m *and that* $\mathbf{Q}$ *is positive definite on the subspace* $M = \{\mathbf{x} : \mathbf{A}\mathbf{x} = \mathbf{0}\}$. *Then the matrix*

$$\begin{bmatrix} \mathbf{Q} & \mathbf{A}^T \\ \mathbf{A} & \mathbf{0} \end{bmatrix} \tag{15.5}$$

is nonsingular.

Proof. Suppose $(\mathbf{x}, \mathbf{y}) \in E^{n+m}$ is such that

$$\begin{aligned} \mathbf{Q}\mathbf{x} + \mathbf{A}^T\mathbf{y} &= \mathbf{0} \\ \mathbf{A}\mathbf{x} \qquad\quad &= \mathbf{0}. \end{aligned} \tag{15.6}$$

Multiplication of the first equation by $\mathbf{x}^T$ yields

$$\mathbf{x}^T\mathbf{Q}\mathbf{x} + \mathbf{x}^T\mathbf{A}^T\mathbf{y} = 0,$$

and substitution of $\mathbf{A}\mathbf{x} = \mathbf{0}$ yields $\mathbf{x}^T\mathbf{Q}\mathbf{x} = \mathbf{0}$. However, clearly $\mathbf{x} \in M$, and thus the hypothesis on $\mathbf{Q}$ together with $\mathbf{x}^T\mathbf{Q}\mathbf{x} = \mathbf{0}$ implies that $\mathbf{x} = \mathbf{0}$. It then follows from the first equation that $\mathbf{A}^T\mathbf{y} = \mathbf{0}$. The full-rank condition on $\mathbf{A}$ then implies that $\mathbf{y} = \mathbf{0}$. Thus the only solution to (15.6) is $\mathbf{x} = \mathbf{0}$, $\mathbf{y} = \mathbf{0}$. ∎

If, as is often the case, the matrix $\mathbf{Q}$ is actually positive definite (over the whole space), then an explicit formula for the solution of the system can be easily derived as follows: From the first equation in (15.4) we have

$$\mathbf{x} = -\mathbf{Q}^{-1}\mathbf{A}^T\boldsymbol{\lambda} - \mathbf{Q}^{-1}\mathbf{c}.$$

Substitution of this into the second equation then yields

$$-\mathbf{A}\mathbf{Q}^{-1}\mathbf{A}^T\boldsymbol{\lambda} - \mathbf{A}\mathbf{Q}^{-1}\mathbf{c} - \mathbf{b} = \mathbf{0},$$

from which we immediately obtain

$$\boldsymbol{\lambda} = -(\mathbf{A}\mathbf{Q}^{-1}\mathbf{A}^T)^{-1}[\mathbf{A}\mathbf{Q}^{-1}\mathbf{c} + \mathbf{b}] \tag{15.7}$$

and

$$\begin{aligned}\mathbf{x} &= \mathbf{Q}^{-1}\mathbf{A}^T(\mathbf{A}\mathbf{Q}^{-1}\mathbf{A}^T)^{-1}[\mathbf{A}\mathbf{Q}^{-1}\mathbf{c} + \mathbf{b}] - \mathbf{Q}^{-1}\mathbf{c} \\ &= -\mathbf{Q}^{-1}[\mathbf{I} - \mathbf{A}^T(\mathbf{A}\mathbf{Q}^{-1}\mathbf{A}^T)^{-1}\mathbf{A}\mathbf{Q}^{-1}]\mathbf{c} \\ &\quad + \mathbf{Q}^{-1}\mathbf{A}^T(\mathbf{A}\mathbf{Q}^{-1}\mathbf{A}^T)^{-1}\mathbf{b}.\end{aligned} \tag{15.8}$$

Strategies

There are some general strategies that guide the development of the primal-dual methods of this chapter.

1. **Descent Measures.** A fundamental concept that we have frequently used is that of assuring that progress is made at each step of an iterative algorithm. It is this that is used to guarantee global convergence. In primal methods this measure of descent is the objective function. Even the simplex method of linear programming is founded on this idea of making progress with respect to the objective function. For primal minimization methods, one typically arranges that the objective function decreases at each step.
   The objective function is not the only possible way to measure progress. We have, for example, when minimizing a function f, considered the quantity $(1/2)|\nabla f(\mathbf{x})|^2$, seeking to monotonically reduce it to zero.
   In general, a function used to measure progress is termed a *merit function*. Typically, it is defined so as to decrease as progress is made toward the solution of a minimization problem, but the sign may be reversed in some definitions. For primal-dual methods, the merit function may depend on both $\mathbf{x}$ and $\boldsymbol{\lambda}$. One especially useful merit function for equality constrained problems is

$$m(\mathbf{x}, \boldsymbol{\lambda}) = \frac{1}{2}|\nabla f(\mathbf{x}) + \boldsymbol{\lambda}^T\nabla\mathbf{h}(\mathbf{x})|^2 + \frac{1}{2}|\mathbf{h}(\mathbf{x}))|^2.$$

   It is examined in the next section.

We shall examine other merit functions later in the chapter. With interior point methods or semidefinite programming, we shall use a potential function that serves as a merit function.

2. **Active Set Methods.** Inequality constraints can be treated using active set methods that treat the active constraints as equality constraints, at least for the current iteration. However, in primal-dual methods, both $\mathbf{x}$ and $\boldsymbol{\lambda}$ are changed. We shall consider variations of steepest descent, conjugate directions, and Newton's method where movement is made in the $(\mathbf{x}, \boldsymbol{\lambda})$ space.
3. **Penalty Functions.** In some primal-dual methods, a penalty function can serve as a merit function, even though the penalty function depends only on $\mathbf{x}$. This is particularly attractive for recursive quadratic programming methods where a quadratic program is solved at each stage to determine the direction of change in the pair $(\mathbf{x}, \boldsymbol{\lambda})$.
4. **Interior (Barrier) Methods.** Barrier methods lead to methods that move within the relative interior of the inequality constraints. This approach leads to the concept of the primal-dual central path. These methods are used for semidefinite programming since these problems are characterized as possessing a special form of inequality constraint.

15.2 A Simple Merit Function

It is very natural, when considering the system of necessary conditions (15.2), to form the function

$$m(\mathbf{x}, \boldsymbol{\lambda}) = \frac{1}{2}|\nabla f(\mathbf{x}) + \boldsymbol{\lambda}^T \nabla \mathbf{h}(\mathbf{x})|^2 + \frac{1}{2}|\mathbf{h}(\mathbf{x})|^2, \tag{15.9}$$

and use it as a measure of how close a point $(\mathbf{x}, \boldsymbol{\lambda})$ is to a solution.

It must be noted, however, that the function $m(\mathbf{x}, \boldsymbol{\lambda})$ is not always well-behaved; it may have local minima, and these are of no value in a search for a solution. The following theorem gives the conditions under which the function $m(\mathbf{x}, \boldsymbol{\lambda})$ can serve as a well-behaved merit function. Basically, the main requirement is that the Hessian of the Lagrangian be positive definite. As usual, we define $l(\mathbf{x}, \boldsymbol{\lambda}) = f(\mathbf{x}) + \boldsymbol{\lambda}^T \mathbf{h}(\mathbf{x})$.

Theorem. *Let f and $\mathbf{h}$ be twice continuously differentiable functions on E^n of dimension 1 and m, respectively. Suppose that $\mathbf{x}^*$ and $\boldsymbol{\lambda}^*$ satisfy the first-order necessary conditions for a local minimum of $m(\mathbf{x}, \boldsymbol{\lambda}) = \frac{1}{2}|\nabla f(\mathbf{x})+\boldsymbol{\lambda}^T \nabla \mathbf{h}(\mathbf{x})|^2 + \frac{1}{2}|\mathbf{h}(\mathbf{x})|^2$ with respect to $\mathbf{x}$ and $\boldsymbol{\lambda}$. Suppose also that at $\mathbf{x}^*$, $\boldsymbol{\lambda}^*$, (i) the rank of $\nabla \mathbf{h}(\mathbf{x}^*)$ is m and (ii) the Hessian matrix $\mathbf{L}(\mathbf{x}^*, \boldsymbol{\lambda}^*) = \mathbf{F}(\mathbf{x}^*)+\boldsymbol{\lambda}^{*T}\mathbf{H}(\mathbf{x}^*)$ is positive definite. Then, $\mathbf{x}^*$, $\boldsymbol{\lambda}^*$ is a (possibly nonunique) global minimum point of $m(\mathbf{x}, \boldsymbol{\lambda})$, with value $m(\mathbf{x}^*, \boldsymbol{\lambda}^*) = 0$.*

Proof. Since $\mathbf{x}^*$, $\boldsymbol{\lambda}^*$ satisfies the first-order conditions for a local minimum point of $m(\mathbf{x}, \boldsymbol{\lambda})$, we have

$$[\nabla f(\mathbf{x}^*) + \boldsymbol{\lambda}^{*T} \nabla \mathbf{h}(\mathbf{x}^*)]\mathbf{L}(\mathbf{x}^*, \boldsymbol{\lambda}^*) + \mathbf{h}(\mathbf{x}^*)^T \nabla \mathbf{h}(\mathbf{x}^*) = \mathbf{0} \tag{15.10}$$

$$[\nabla f(\mathbf{x}^*) + \boldsymbol{\lambda}^{*T} \nabla \mathbf{h}(\mathbf{x}^*)]\nabla \mathbf{h}(\mathbf{x}^*)^T = \mathbf{0}. \tag{15.11}$$

Multiplying (15.10) on the right by $[\nabla f(\mathbf{x}^*) + \boldsymbol{\lambda}^{*T}\nabla \mathbf{h}(\mathbf{x}^*)]^T$ and using (15.11) we obtain†

$$\nabla l(\mathbf{x}^*, \boldsymbol{\lambda}^*)\mathbf{L}(\mathbf{x}^*, \boldsymbol{\lambda}^*)\nabla l(\mathbf{x}^*, \boldsymbol{\lambda}^*)^T = 0.$$

Since $\mathbf{L}(\mathbf{x}^*, \boldsymbol{\lambda}^*)$ is positive definite, this implies that $\nabla l(\mathbf{x}^*, \boldsymbol{\lambda}^*) = \mathbf{0}$. Using this in (15.10), we find that $\mathbf{h}(\mathbf{x}^*)^T\nabla \mathbf{h}(\mathbf{x}^*) = \mathbf{0}$, which, since $\nabla \mathbf{h}(\mathbf{x}^*)$ is of rank m, implies that $\mathbf{h}(\mathbf{x}^*) = \mathbf{0}$. ∎

The requirement that the Hessian of the Lagrangian $\mathbf{L}(\mathbf{x}^*, \boldsymbol{\lambda}^*)$ be positive definite at a stationary point of the merit function m is actually not too restrictive. This condition will be satisfied in the case of a convex programming problem where f is strictly convex and $\mathbf{h}$ is linear. Furthermore, even in nonconvex problems one can often arrange for this condition to hold, at least near a solution to the original constrained minimization problem. If it is assumed that the second-order sufficiency conditions for a constrained minimum hold at $\mathbf{x}^*, \boldsymbol{\lambda}^*$, then $\mathbf{L}(\mathbf{x}^*, \boldsymbol{\lambda}^*)$ is positive definite on the subspace that defines the tangent to the constraints; that is, on the subspace defined by $\nabla \mathbf{h}(\mathbf{x}^*)\mathbf{x} = \mathbf{0}$. Now if the original problem is modified with a penalty term to the problem

$$\begin{aligned} &\text{minimize} \quad f(\mathbf{x}) + \frac{1}{2}c|\mathbf{h}(\mathbf{x})|^2 \\ &\text{subject to} \quad \mathbf{h}(\mathbf{x}) = \mathbf{0}, \end{aligned} \tag{15.12}$$

the solution point $\mathbf{x}^*$ will be unchanged. However, as discussed in Chap. 14, the Hessian of the Lagrangian of this new problem (15.12) at the solution point is $\mathbf{L}(\mathbf{x}^*, \boldsymbol{\lambda}^*) + c\nabla \mathbf{h}(\mathbf{x}^*)^T\nabla \mathbf{h}(\mathbf{x}^*)$. For sufficiently large c, this matrix will be positive definite. Thus a problem can be “convexified” (at least locally) before the merit function method is employed.

An extension to problems with inequality constraints can be defined by partitioning the constraints into the two groups *active* and *inactive*. However, at this point the simple merit function for problems with equality constraints is adequate for the purpose of illustrating the general idea.

15.3 Basic Primal-Dual Methods

Many primal-dual methods are patterned after some of the methods used in earlier chapters, except of course that the emphasis is on equation solving rather than explicit optimization.

† Unless explicitly indicated to the contrary, the notation $\nabla l(\mathbf{x}, \boldsymbol{\lambda})$ refers to the gradient of l with respect to $\mathbf{x}$, that is, $\nabla_{\mathbf{x}} l(\mathbf{x}, \boldsymbol{\lambda})$.

First-Order Method

We consider first a simple straightforward approach, which in a sense parallels the idea of steepest descent in that it uses only a first-order approximation to the primal–dual equations. It is defined by

$$\begin{aligned} \mathbf{x}_{k+1} &= \mathbf{x}_k - \alpha_k \nabla l(\mathbf{x}_k, \boldsymbol{\lambda}_k)^T \\ \boldsymbol{\lambda}_{k+1} &= \boldsymbol{\lambda}_k + \alpha_k \mathbf{h}(\mathbf{x}_k), \end{aligned} \tag{15.13}$$

where α_k is not yet determined. This is based on the error in satisfying (15.2). Assume that the Hessian of the Lagrangian $\mathbf{L}(\mathbf{x}, \boldsymbol{\lambda})$ is positive definite in some compact region of interest, and consider the simple merit function

$$m(\mathbf{x}, \boldsymbol{\lambda}) = \frac{1}{2}|\nabla l(\mathbf{x}, \boldsymbol{\lambda})|^2 + \frac{1}{2}|\mathbf{h}(\mathbf{x})|^2 \tag{15.14}$$

discussed above. We would like to determine whether the direction of change in (15.13) is a descent direction with respect to this merit function. The gradient of the merit function has components corresponding to $\mathbf{x}$ and $\boldsymbol{\lambda}$ of

$$\begin{aligned} &\nabla l(\mathbf{x}, \boldsymbol{\lambda})\mathbf{L}(\mathbf{x}, \boldsymbol{\lambda}) + \mathbf{h}(\mathbf{x})^T \nabla \mathbf{h}(\mathbf{x}) \\ &\nabla l(\mathbf{x}, \boldsymbol{\lambda}) \nabla \mathbf{h}(\mathbf{x})^T. \end{aligned} \tag{15.15}$$

Thus the inner product of this gradient with the direction vector having components $-\nabla l(\mathbf{x}, \boldsymbol{\lambda})^T$, $\mathbf{h}(\mathbf{x})$ is

$$\begin{aligned} -\nabla l(\mathbf{x}, \boldsymbol{\lambda})\mathbf{L}(\mathbf{x}, \boldsymbol{\lambda})\nabla l(\mathbf{x}, \boldsymbol{\lambda})^T - \mathbf{h}(\mathbf{x})^T \nabla \mathbf{h}(\mathbf{x}) \nabla l(\mathbf{x}, \boldsymbol{\lambda})^T + \nabla l(\mathbf{x}, \boldsymbol{\lambda}) \nabla \mathbf{h}(\mathbf{x})^T \mathbf{h}(\mathbf{x}) \\ = -\nabla l(\mathbf{x}, \boldsymbol{\lambda})\mathbf{L}(\mathbf{x}, \boldsymbol{\lambda})\nabla l(\mathbf{x}, \boldsymbol{\lambda})^T \leqslant 0. \end{aligned}$$

This shows that the search direction is in fact a descent direction for the merit function, unless $\nabla l(\mathbf{x}, \boldsymbol{\lambda}) = \mathbf{0}$. Thus by selecting α_k to minimize the merit function in the search direction at each step, the process will converge to a point where $\nabla l(\mathbf{x}, \boldsymbol{\lambda}) = \mathbf{0}$. However, there is no guarantee that $\mathbf{h}(\mathbf{x}) = \mathbf{0}$ at that point.

We can try to improve the method either by changing the way in which the direction is selected or by changing the merit function. In this case a slight modification of the merit function will work. Let

$$w(\mathbf{x}, \boldsymbol{\lambda}, \gamma) = m(\mathbf{x}, \boldsymbol{\lambda}) - \gamma[f(\mathbf{x}) + \boldsymbol{\lambda}^T \mathbf{h}(\mathbf{x})]$$

for some $\gamma > 0$. We then calculate that the gradient of w has the two components corresponding to $\mathbf{x}$ and $\boldsymbol{\lambda}$

$$\begin{aligned} &\nabla l(\mathbf{x}, \boldsymbol{\lambda})\mathbf{L}(\mathbf{x}, \boldsymbol{\lambda}) + \mathbf{h}(\mathbf{x})^T \nabla \mathbf{h}(\mathbf{x}) - \gamma \nabla l(\mathbf{x}, \boldsymbol{\lambda}) \\ &\nabla l(\mathbf{x}, \boldsymbol{\lambda}) \nabla \mathbf{h}(\mathbf{x})^T - \gamma \mathbf{h}(\mathbf{x})^T, \end{aligned}$$

and hence the inner product of the gradient with the direction $-\nabla l(\mathbf{x}, \boldsymbol{\lambda})^T$, $\mathbf{h}(\mathbf{x})$ is

$$-\nabla l(\mathbf{x}, \boldsymbol{\lambda})[\mathbf{L}(\mathbf{x}, \boldsymbol{\lambda}) - \gamma \mathbf{I}]\nabla l(\mathbf{x}, \boldsymbol{\lambda})^T - \gamma|\mathbf{h}(\mathbf{x})|^2.$$

Now since we are assuming that $\mathbf{L}(\mathbf{x}, \boldsymbol{\lambda})$ is positive definite in a compact region of interest, there is a $\gamma > 0$ such that $\mathbf{L}(\mathbf{x}, \boldsymbol{\lambda}) - \gamma\mathbf{I}$ is positive definite in this region. Then according to the above calculation, the direction $-\nabla l(\mathbf{x}, \boldsymbol{\lambda})^T, \mathbf{h}(\mathbf{x})$ is a descent direction, and the standard descent method will converge to a solution. This method will not converge very rapidly however. (See Exercise 2 for further analysis of this method.)

Conjugate Directions

One may also use the conjugate direction. Let us consider the quadratic program

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\mathbf{x}^T\mathbf{Q}\mathbf{x} - \mathbf{b}^T\mathbf{x} \\ \text{subject to} \quad & \mathbf{A}\mathbf{x} = \mathbf{c}. \end{aligned} \tag{15.16}$$

The first-order necessary conditions for this problem are

$$\begin{aligned} \mathbf{Q}\mathbf{x} + \mathbf{A}^T\boldsymbol{\lambda} &= \mathbf{b} \\ \mathbf{A}\mathbf{x} \qquad &= \mathbf{c}. \end{aligned} \tag{15.17}$$

As discussed in the previous section, this problem is equivalent to solving a system of linear equations whose coefficient matrix is

$$\mathbf{M} = \begin{bmatrix} \mathbf{Q} & \mathbf{A}^T \\ \mathbf{A} & \mathbf{0} \end{bmatrix}. \tag{15.18}$$

This matrix is symmetric, but it is *not* positive definite (nor even semidefinite). However, it is possible to formally generalize the conjugate gradient method to systems of this type by just applying the conjugate-gradient formulae (15.17)–(15.20) of Sect. 9.3 with $\mathbf{Q}$ replaced by $\mathbf{M}$. A difficulty is that *singular directions* (defined as directions $\mathbf{p}$ such that $\mathbf{p}^T\mathbf{M}\mathbf{p} = 0$) may occur and cause the process to break down. Procedures for overcoming this difficulty have been developed, however. Also, as in the ordinary conjugate gradient method, the approach can be generalized to treat nonquadratic problems as well. Overall, however, the application of conjugate direction methods to the Lagrange system of equations, although very promising, is not currently considered practical.

Second-Order Method: Newton's Method

Newton's method for solving systems of equations can be easily applied to the Lagrange equations. In its most straightforward form, the method solves the system

$$\begin{aligned} \nabla l(\mathbf{x}, \boldsymbol{\lambda}) &= \mathbf{0} \\ \mathbf{h}(\mathbf{x}) &= \mathbf{0} \end{aligned} \tag{15.19}$$

by solving the linearized version recursively. That is, given $\mathbf{x}_k,\ \boldsymbol{\lambda}_k$ the new point $\mathbf{x}_{k+1},\ \boldsymbol{\lambda}_{k+1}$ is determined from the equations

$$\begin{aligned} \nabla l(\mathbf{x}_k,\ \boldsymbol{\lambda}_k)^T + \mathbf{L}(\mathbf{x}_k,\ \boldsymbol{\lambda}_k)\mathbf{d}_k + \nabla \mathbf{h}(\mathbf{x}_k)^T \mathbf{y}_k &= \mathbf{0} \\ \mathbf{h}(\mathbf{x}_k) + \nabla \mathbf{h}(\mathbf{x}_k)\mathbf{d}_k &= \mathbf{0} \end{aligned} \tag{15.20}$$

by setting $\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{d}_k,\ \boldsymbol{\lambda}_{k+1} = \boldsymbol{\lambda}_k + \mathbf{y}_k$. In matrix form the above Newton equations are

$$\begin{bmatrix} \mathbf{L}(\mathbf{x}_k, \boldsymbol{\lambda}_k) & \nabla \mathbf{h}(\mathbf{x}_k)^T \\ \nabla \mathbf{h}(\mathbf{x}_k) & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{d}_k \\ \mathbf{y}_k \end{bmatrix} = \begin{bmatrix} -\nabla l(\mathbf{x}_k, \boldsymbol{\lambda}_k)^T \\ -\mathbf{h}(\mathbf{x}_k) \end{bmatrix}. \tag{15.21}$$

The Newton equations have some important structural properties. First, we observe that by adding $\nabla \mathbf{h}(\mathbf{x}_k)^T \boldsymbol{\lambda}_k$ to the top equation, the system can be transformed to the form

$$\begin{bmatrix} \mathbf{L}(\mathbf{x}_k, \boldsymbol{\lambda}_k) & \nabla \mathbf{h}(\mathbf{x}_k)^T \\ \nabla \mathbf{h}(\mathbf{x}_k) & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{d}_k \\ \boldsymbol{\lambda}_{k+1} \end{bmatrix} = \begin{bmatrix} -\nabla f(\mathbf{x}_k)^T \\ -\mathbf{h}(\mathbf{x}_k) \end{bmatrix}, \tag{15.22}$$

where again $\boldsymbol{\lambda}_{k+1} = \boldsymbol{\lambda}_k + \mathbf{y}_k$. In this form $\boldsymbol{\lambda}_k$ appears only in the matrix $\mathbf{L}(\mathbf{x}_k,\ \boldsymbol{\lambda}_k)$. This conversion between (15.21) and (15.22) will be useful later.

Next we note that the structure of the coefficient matrix of (15.21) or (15.22) is identical to that of the Proposition of Sect. 15.1. The standard second-order sufficiency conditions imply that $\nabla \mathbf{h}(\mathbf{x}^*)$ is of full rank and that $\mathbf{L}(\mathbf{x}^*,\ \boldsymbol{\lambda}^*)$ is positive definite on $M = \{\mathbf{x} : \nabla \mathbf{h}(\mathbf{x}^*)\mathbf{x} = \mathbf{0}\}$ at the solution. By continuity these conditions can be assumed to hold in a region near the solution as well. Under these assumptions it follows from Proposition 1 that the Newton equation (15.21) has a unique solution.

It is again worthwhile to point out that, although the Hessian of the Lagrangian need be positive definite only on the tangent subspace in order for the system (15.21) to be nonsingular, it is possible to alter the original problem by incorporation of a quadratic penalty term so that the new Hessian of the Lagrangian is $\mathbf{L}(\mathbf{x},\ \boldsymbol{\lambda}) + c\nabla \mathbf{h}(\mathbf{x})^T \nabla \mathbf{h}(\mathbf{x})$. For sufficiently large c, this new Hessian will be positive definite over the entire space.

If $\mathbf{L}(\mathbf{x},\ \boldsymbol{\lambda})$ is positive definite (either originally or through the incorporation of a penalty term), it is possible to write an explicit expression for the solution of the system (15.21). Let us define $\mathbf{L}_k = \mathbf{L}(\mathbf{x}_k,\ \boldsymbol{\lambda}_k),\ \mathbf{A}_k = \nabla \mathbf{h}(\mathbf{x}_k),\ \mathbf{l}_k = \nabla l(\mathbf{x}_k,\ \boldsymbol{\lambda}_k)^T,\ \mathbf{h}_k = \mathbf{h}(\mathbf{x}_k)$. The system then takes the form

$$\begin{aligned} \mathbf{L}_k \mathbf{d}_k + \mathbf{A}_k^T \mathbf{y}_k &= -\mathbf{l}_k \\ \mathbf{A}_k \mathbf{d}_k \qquad\quad &= -\mathbf{h}_k. \end{aligned} \tag{15.23}$$

The solution is readily found, as in (15.7) and (15.8) for quadratic programming, to be

$$\mathbf{y}_k = (\mathbf{A}_k \mathbf{L}_k^{-1} \mathbf{A}_k^T)^{-1} [\mathbf{h}_k - \mathbf{A}_k \mathbf{L}_k^{-1} \mathbf{l}_k] \tag{15.24}$$

$$\mathbf{d}_k = -\mathbf{L}_k^{-1} [\mathbf{I} - \mathbf{A}_k^T (\mathbf{A}_k \mathbf{L}_k^{-1} \mathbf{A}_k^T)^{-1} \mathbf{A}_k \mathbf{L}_k^{-1}] \mathbf{l}_k - \mathbf{L}_k^{-1} \mathbf{A}_k^T (\mathbf{A}_k \mathbf{L}_k^{-1} \mathbf{A}_k^T)^{-1} \mathbf{h}_k. \tag{15.25}$$

There are standard results concerning Newton's method applied to a system of nonlinear equations that are applicable to the system (15.19). These results state that if the linearized system is nonsingular at the solution (as is implied by our assumptions) and if the initial point is sufficiently close to the solution, the method will in fact converge to the solution and the convergence will be of order at least two. To guarantee convergence from remote initial points and hence be more broadly applicable, it is desirable to use the method as a descent process. Fortunately, we can show that the direction generated by Newton's method is a descent direction for the simple merit function

$$m(\mathbf{x},\ \boldsymbol{\lambda}) = \frac{1}{2}|\boldsymbol{\nabla} l(\mathbf{x},\ \boldsymbol{\lambda})|^2 + \frac{1}{2}|\mathbf{h}(\mathbf{x})|^2.$$

Given $\mathbf{d}_k,\ \mathbf{y}_k$ satisfying (15.23), the inner product of this direction with the gradient of m at $\mathbf{x}_k,\ \boldsymbol{\lambda}_k$ is, referring to (15.15),

$$\begin{aligned}[\mathbf{L}_k\mathbf{l}_k + \mathbf{A}_k^T\mathbf{h}_k,\ \mathbf{A}_k\mathbf{l}_k]^T[\mathbf{d}_k,\ \mathbf{y}_k] &= \mathbf{l}_k^T\mathbf{L}_k\mathbf{d}_k + \mathbf{h}_k^T\mathbf{A}_k\mathbf{d}_k + \mathbf{l}_k^T\mathbf{A}_k^T\mathbf{y}_k \\ &= -|\mathbf{l}_k|^2 - |\mathbf{h}_k|^2.\end{aligned}$$

This is strictly negative unless both $\mathbf{l}_k = \mathbf{0}$ and $\mathbf{h}_k = \mathbf{0}$. Thus Newton's method has desirable global convergence properties when executed as a descent method with variable step size.

Note that the calculation above does not employ the explicit formulae (15.24) and (15.25), and hence it is not necessary that $\mathbf{L}(\mathbf{x},\ \boldsymbol{\lambda})$ be positive definite, as long as the system (15.21) is invertible. We summarize the above discussion by the following theorem.

Theorem. *Define the Newton process by*

$$\begin{aligned}\mathbf{x}_{k+1} &= \mathbf{x}_k + \alpha_k\mathbf{d}_k \\ \boldsymbol{\lambda}_{k+1} &= \boldsymbol{\lambda}_k + \alpha_k\mathbf{y}_k,\end{aligned}$$

where $\mathbf{d}_k,\ \mathbf{y}_k$ *are solutions to (15.24) and where* α_k *is selected to minimize the merit function*

$$m(\mathbf{x},\ \boldsymbol{\lambda}) = \frac{1}{2}|\boldsymbol{\nabla} l(\mathbf{x},\ \boldsymbol{\lambda})|^2 + \frac{1}{2}|\mathbf{h}(\mathbf{x})|^2.$$

Assume that $\mathbf{d}_k,\ \mathbf{y}_k$ *exist and that the points generated lie in a compact set. Then any limit point of these points satisfies the first-order necessary conditions for a solution to the constrained minimization problem (15.1).*

Proof. Most of this follows from the above observations and the Global Convergence Theorem. The one-dimensional search process is well-defined, since the merit function m is bounded below. ∎

In view of this result, it is worth pursuing Newton's method further. We would like to extend it to problems with inequality constraints. We would also like to avoid the necessity of evaluating $\mathbf{L}(\mathbf{x}_k,\ \boldsymbol{\lambda}_k)$ at each step and to consider alternative merit functions—perhaps those that might distinguish a local maximum from a local minimum, which the simple merit function does not do. These considerations guide the developments of the next several sections.

Relation to Sequential Quadratic Programming

It is clear from the development of the preceding discussion that Newton's method is closely related to quadratic programming with equality constraints. We explore this relationship more fully here, which will lead to a generalization of Newton's method to problems with inequality constraints.

Consider the problem

$$\begin{aligned} &\text{minimize} \quad \mathbf{l}_k^T \mathbf{d}_k + \frac{1}{2}\mathbf{d}_k^T \mathbf{L}_k \mathbf{d}_k \\ &\text{subject to} \quad \mathbf{A}_k \mathbf{d}_k + \mathbf{h}_k = \mathbf{0}. \end{aligned} \tag{15.26}$$

The first-order necessary conditions of this problem are exactly (15.21), or equivalently (15.23), where $\mathbf{y}_k$ corresponds to the Lagrange multiplier of (15.26). Thus, the solution of (15.26) produces a Newton step.

Alternatively, we may consider the quadratic program

$$\begin{aligned} &\text{minimize} \quad \nabla f(\mathbf{x}_k)\mathbf{d}_k + \frac{1}{2}\mathbf{d}_k^T \mathbf{L}_k \mathbf{d}_k \\ &\text{subject to} \quad \mathbf{A}_k \mathbf{d}_k + \mathbf{h}_k = \mathbf{0}. \end{aligned} \tag{15.27}$$

The necessary conditions of this problem are exactly (15.22), where $\boldsymbol{\lambda}_{k+1}$ now corresponds to the Lagrange multiplier of (15.27). The program (15.27) is obtained from (15.26) by merely subtracting $\boldsymbol{\lambda}_k^T \mathbf{A}_k \mathbf{d}_k$ from the objective function; and this change has no influence on $\mathbf{d}_k$, since $\mathbf{A}_k \mathbf{d}_k$ is fixed.

The connection with quadratic programming suggests a procedure for extending Newton's method to minimization problems with inequality constraints. Consider the problem

$$\begin{aligned} &\text{minimize} \quad f(\mathbf{x}) \\ &\text{subject to} \quad \mathbf{h}(\mathbf{x}) = \mathbf{0} \\ &\qquad\qquad\quad \mathbf{g}(\mathbf{x}) \leqslant \mathbf{0}. \end{aligned}$$

Given an estimated solution point $\mathbf{x}_k$ and estimated Lagrange multipliers $\boldsymbol{\lambda}_k$, $\boldsymbol{\mu}_k$, one solves the quadratic program

$$\begin{aligned} &\text{minimize} \quad \nabla f(\mathbf{x}_k)\mathbf{d}_k + \tfrac{1}{2}\mathbf{d}_k^T \mathbf{L}_k \mathbf{d}_k \\ &\text{subject to} \quad \nabla \mathbf{h}(\mathbf{x}_k)\mathbf{d}_k + \mathbf{h}_k = \mathbf{0} \\ &\qquad\qquad\quad \nabla \mathbf{g}(\mathbf{x}_k)\mathbf{d}_k + \mathbf{g}_k \leqslant \mathbf{0}, \end{aligned} \tag{15.28}$$

where $\mathbf{L}_k = \mathbf{F}(\mathbf{x}_k) + \boldsymbol{\lambda}_k^T \mathbf{H}(\mathbf{x}_k) + \boldsymbol{\mu}_k^T \mathbf{G}(\mathbf{x}_k)$, $\mathbf{h}_k = \mathbf{h}(\mathbf{x}_k)$, $\mathbf{g}_k = \mathbf{g}(\mathbf{x}_k)$. The new point is determined by $\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{d}_k$, and the new Lagrange multipliers are the Lagrange multipliers of the quadratic program (15.28). This is the essence of an early method for nonlinear programming termed SOLVER. It is a very attractive procedure, since it applies directly to problems with inequality as well as equality constraints without the use of an active set strategy (although such a strategy might be used to solve the required quadratic program). Methods of this general type, where a quadratic program is solved at each step, are referred to as *recursive quadratic programming* methods, and several variations are considered in this chapter.

As presented here the recursive quadratic programming method extends Newton's method to problems with inequality constraints, but the method has limitations. The quadratic program may not always be well-defined, the method requires second-order derivative information, and the simple merit function is not a descent function for the case of inequalities. Of these, the most serious is the requirement of second-order information, and this is addressed in the next section.

15.4 Modified Newton Methods

A modified Newton method is based on replacing the actual linearized system by an approximation.

First, we concentrate on the equality constrained optimization problem

$$\begin{array}{ll} \text{minimize} & f(\mathbf{x}) \\ \text{subject to} & \mathbf{h}(\mathbf{x}) = \mathbf{0} \end{array} \tag{15.29}$$

in order to most clearly describe the relationships between the various approaches. Problems with inequality constraints can be treated within the equality constraint framework by an active set strategy or, in some cases, by recursive quadratic programming.

The basic equations for Newton's method can be written

$$\begin{bmatrix} \mathbf{x}_{k+1} \\ \boldsymbol{\lambda}_{k+1} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_k \\ \boldsymbol{\lambda}_k \end{bmatrix} - \alpha_k \begin{bmatrix} \mathbf{L}_k & \mathbf{A}_k^T \\ \mathbf{A}_k & \mathbf{0} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{l}_k \\ \mathbf{h}_k \end{bmatrix},$$

where as before $\mathbf{L}_k$ is the Hessian of the Lagrangian, $\mathbf{A}_k = \nabla \mathbf{h}(\mathbf{x}_k)$, $\mathbf{l}_k = [\nabla f(\mathbf{x}_k) + \boldsymbol{\lambda}_k^T \nabla \mathbf{h}(\mathbf{x}_k)]^T$, $\mathbf{h}_k = \mathbf{h}(\mathbf{x}_k)$. A *structured modified Newton method* is a method of the form

$$\begin{bmatrix} \mathbf{x}_{k+1} \\ \boldsymbol{\lambda}_{k+1} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_k \\ \boldsymbol{\lambda}_k \end{bmatrix} - \alpha_k \begin{bmatrix} \mathbf{B}_k & \mathbf{A}_k^T \\ \mathbf{A}_k & \mathbf{0} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{l}_k \\ \mathbf{h}_k \end{bmatrix}, \tag{15.30}$$

where $\mathbf{B}_k$ is an approximation to $\mathbf{L}_k$. The term "structured" derives from the fact that only second-order information in the original system of equations is approximated; the first-order information is kept intact.

Of course the method is implemented by solving the system

$$\begin{aligned} \mathbf{B}_k \mathbf{d}_k + \mathbf{A}_k^T \mathbf{y}_k &= -\mathbf{l}_k \\ \mathbf{A}_k \mathbf{d}_k \qquad &= -\mathbf{h}_k \end{aligned} \tag{15.31}$$

for $\mathbf{d}_k$ and $\mathbf{y}_k$ and then setting $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{d}_k$, $\boldsymbol{\lambda}_{k+1} = \boldsymbol{\lambda}_k + \alpha_k \mathbf{y}_k$ for some value of α_k. In this section we will not consider the procedure for selection of α_k, and thus for simplicity we take $\alpha_k = 1$. The simple transformation used earlier can be applied to write (15.31) in the form

$$\begin{aligned} \mathbf{B}_k\mathbf{d}_k + \mathbf{A}_k^T\boldsymbol{\lambda}_{k+1} &= -\nabla f(\mathbf{x}_k)^T \\ \mathbf{A}_k\mathbf{d}_k &= -\mathbf{h}_k. \end{aligned} \tag{15.32}$$

Then $\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{d}_k$, and $\boldsymbol{\lambda}_{k+1}$ is found directly as a solution to system (15.32).

There are, of course, various ways to choose the approximation $\mathbf{B}_k$. One is to use a fixed, constant matrix throughout the iterative process. A second is to base $\mathbf{B}_k$ on some readily accessible information in $\mathbf{L}(\mathbf{x}_k, \ \boldsymbol{\lambda}_k)$, such as setting $\mathbf{B}_k$ equal to the diagonal of $\mathbf{L}(\mathbf{x}_k, \ \boldsymbol{\lambda}_k)$. Finally, a third possibility is to update $\mathbf{B}_k$ using one of the various quasi-Newton formulae.

One important advantage of the structured method is that $\mathbf{B}_k$ can be taken to be positive definite even though $\mathbf{L}_k$ is not. If this is done, we can write the explicit solution

$$\mathbf{y}_k = (\mathbf{A}_k\mathbf{B}_k^{-1}\mathbf{A}_k^T)^{-1}[\mathbf{h}_k - \mathbf{A}_k\mathbf{B}_k^{-1}\mathbf{I}_k] \tag{15.33}$$

$$\mathbf{d}_k = -\mathbf{B}_k^{-1}[\mathbf{I} - \mathbf{A}_k^T(\mathbf{A}_k\mathbf{B}_k^{-1}\mathbf{A}_k^T)^{-1}\mathbf{A}_k\mathbf{B}_k^{-1}]\mathbf{I}_k - \mathbf{B}_k^{-1}\mathbf{A}_k^T(\mathbf{A}_k\mathbf{B}_k^{-1}\mathbf{A}_k^T)^{-1}\mathbf{h}_k. \tag{15.34}$$

Consider the quadratic program

$$\begin{aligned} &\text{minimize} \quad \nabla f(\mathbf{x}_k)\mathbf{d}_k + \frac{1}{2}\mathbf{d}_k^T\mathbf{B}_k\mathbf{d}_k \\ &\text{subject to} \quad \mathbf{A}_k\mathbf{d}_k + \mathbf{h}(\mathbf{x}_k) = \mathbf{0}. \end{aligned} \tag{15.35}$$

The first-order necessary conditions for this problem are

$$\begin{aligned} \mathbf{B}_k\mathbf{d}_k + \mathbf{A}_k^T\boldsymbol{\lambda}_{k+1} &= -\nabla f(\mathbf{x}_k)^T \\ \mathbf{A}_k\mathbf{d}_k &= -\mathbf{h}(\mathbf{x}_k), \end{aligned} \tag{15.36}$$

which are again identical to the system of equations of the structured modified Newton method—in this case in the form (15.33). The Lagrange multiplier of the quadratic program is $\boldsymbol{\lambda}_{k+1}$. The equivalence of (15.35) and (15.36) leads to a recursive quadratic programming method, where at each $\mathbf{x}_k$ the quadratic program (15.35) is solved to determine the direction $\mathbf{d}_k$. In this case an arbitrary symmetric matrix $\mathbf{B}_k$ is used in place of the Hessian of the Lagrangian. Note that the problem (15.35) does not explicitly depend on $\boldsymbol{\lambda}_k$; but $\mathbf{B}_k$, often being chosen to approximate the Hessian of the Lagrangian, may depend on $\boldsymbol{\lambda}_k$.

As before, a principal advantage of the quadratic programming formulation is that there is an obvious extension to problems with inequality constraints: One simply employs a linearized version of the inequalities.

15.5 Descent Properties

In order to ensure convergence of the structured modified Newton methods of the previous section, it is necessary to find a suitable merit function—a merit function that is compatible with the direction-finding algorithm in the sense that it decreases along the direction generated. We must abandon the simple merit function at this point, since it is not compatible with these methods when $\mathbf{B}_k \neq \mathbf{L}_k$. However, two other penalty functions considered earlier, the absolute-value exact penalty function and the quadratic penalty function, *are* compatible with the modified Newton approach.

Absolute-Value Penalty Function

Let us consider the constrained minimization problem

$$\begin{aligned} &\text{minimize} \quad f(\mathbf{x}) \\ &\text{subject to} \quad \mathbf{g}(\mathbf{x}) \leqslant \mathbf{0}, \end{aligned} \tag{15.37}$$

where $\mathbf{g}(\mathbf{x})$ is r-dimensional. For notational simplicity we consider the case of inequality constraints only, since it is, in fact, the most difficult case. The extension to equality constraints is straightforward. In accordance with the recursive quadratic programming approach, given a current point $\mathbf{x}$, we select the direction of movement $\mathbf{d}$ by solving the quadratic programming problem

$$\begin{aligned} &\text{minimize} \quad \frac{1}{2}\mathbf{d}^T\mathbf{B}\mathbf{d} + \nabla f(\mathbf{x})\mathbf{d} \\ &\text{subject to} \quad \nabla \mathbf{g}(\mathbf{x})\mathbf{d} + \mathbf{g}(\mathbf{x}) \leqslant \mathbf{0}, \end{aligned} \tag{15.38}$$

where $\mathbf{B}$ is positive definite.

The first-order necessary conditions for a solution to this quadratic program are

$$\mathbf{B}\mathbf{d} + \nabla f(\mathbf{x})^T + \nabla \mathbf{g}(\mathbf{x})^T \boldsymbol{\mu} = \mathbf{0} \tag{15.39a}$$

$$\nabla \mathbf{g}(\mathbf{x})\mathbf{d} + \mathbf{g}(\mathbf{x}) \leqslant \mathbf{0} \tag{15.39b}$$

$$\boldsymbol{\mu}^T[\nabla \mathbf{g}(\mathbf{x})\mathbf{d} + \mathbf{g}(\mathbf{x})] = 0 \tag{15.39c}$$

$$\boldsymbol{\mu} \geqslant \mathbf{0}. \tag{15.39d}$$

Note that if the solution to the quadratic program has $\mathbf{d} = \mathbf{0}$, then the point $\mathbf{x}$, together with $\boldsymbol{\mu}$ from (15.39), satisfies the first-order necessary conditions for the original minimization problem (15.37). The following proposition is the fundamental result concerning the compatibility of the absolute-value penalty function and the quadratic programming method for determining the direction of movement.

Proposition 1. *Let* $\mathbf{d}$, $\boldsymbol{\mu}$ *(with* $\mathbf{d} \neq \mathbf{0}$*) be a solution of the quadratic program (15.38). Then if* $c \geqslant \max_j(\mu_j)$*, the vector* $\mathbf{d}$ *is a descent direction for the penalty function*

$$P(\mathbf{x}) = f(\mathbf{x}) + c\sum_{j=1}^{r} g_j(\mathbf{x})^+.$$

Proof. Let $J(\mathbf{x}) = \{j : g_j(\mathbf{x}) > 0\}$. Now for $\alpha > 0$,

$$\begin{aligned} P(\mathbf{x} + \alpha\mathbf{d}) &= f(\mathbf{x} + \alpha\mathbf{d}) + c\sum_{j=1}^{r} g_j(\mathbf{x} + \alpha\mathbf{d})^+ \\ &= f(\mathbf{x}) + \alpha\nabla f(\mathbf{x})\mathbf{d} + c\sum_{j=1}^{r}[g_j(\mathbf{x}) + \alpha\nabla g_j(\mathbf{x})\mathbf{d}]^+ + o(\alpha) \\ &= f(\mathbf{x}) + \alpha\nabla f(\mathbf{x})\mathbf{d} + c\sum_{j=1}^{r} g_j(\mathbf{x})^+ + \alpha c\sum_{j\in J(\mathbf{x})}\nabla g_j(\mathbf{x})\mathbf{d} + o(\alpha) \\ &= P(\mathbf{x}) + \alpha\nabla f(\mathbf{x})\mathbf{d} + \alpha c\sum_{j\in J(\mathbf{x})}\nabla g_j(\mathbf{x})\mathbf{d} + o(\alpha). \end{aligned} \tag{15.40}$$

Where (15.39b) was used in the third line to infer that $\nabla g_j(\mathbf{x}) \le 0$ if $g_j(\mathbf{x}) = 0$. Again using (15.39b) we have

$$c\sum_{j\in J(\mathbf{x})}\nabla g_j(\mathbf{x})\mathbf{d} \leqslant c\sum_{j\in J(\mathbf{x})} -g_j(\mathbf{x}) = -c\sum_{j=1}^{r} g_j(\mathbf{x})^+. \tag{15.41}$$

Using (15.39a) we have

$$\nabla\mathbf{f}(\mathbf{x})\mathbf{d} = -\mathbf{d}^T\ \mathbf{B}\mathbf{d}\ - \sum_{j=1}^{r}\mu_j\nabla g_j(\mathbf{x})\mathbf{d},$$

which by using the complementary slackness condition (15.39c) leads to

$$\begin{aligned} \nabla f(\mathbf{x})\mathbf{d} = -\mathbf{d}^T\mathbf{B}\mathbf{d} + \sum_{j=1}^{r}\mu_j g_j(\mathbf{x}) &\leqslant -\mathbf{d}^T\mathbf{B}\mathbf{d} + \sum_{j=1}^{r}\mu_j g_j(\mathbf{x})^+ \\ &\le -\mathbf{d}^T\ \mathbf{B}\mathbf{d}\ + \max\,(\mu_j)\sum_{j=1}^{r} g_j(\mathbf{x})^+. \end{aligned} \tag{15.42}$$

Finally, substituting (15.41) and (15.42) in (15.40), we find

$$P(\mathbf{x} + \alpha\mathbf{d}) \leqslant P(\mathbf{x}) + \alpha\{-\mathbf{d}^T\mathbf{B}\mathbf{d} - [c - \max(\mu_j)]\sum_{j=1}^{r} g_j(\mathbf{x})^+\} + o(\alpha),$$

Since $\mathbf{B}$ is positive definite and $c \ge \max(\mu_j)$, it follows that for α sufficiently small, $P(\mathbf{x} + \alpha\mathbf{d}) < P(\mathbf{x})$. ∎

The above proposition is exceedingly important, for it provides a basis for establishing the global convergence of modified Newton methods, including recursive quadratic programming. The following is a simple global convergence result based on the descent property.

Theorem. *Let* $\mathbf{B}$ *be positive definite and assume that throughout some compact region* $\subset E^n$*, the quadratic program (15.38) has a unique solution* $\mathbf{d}$, $\boldsymbol{\mu}$ *such that at each point the Lagrange multipliers satisfy* $\max_j(\mu_j) \leqslant c$*. Let the sequence* $\{\mathbf{x}_k\}$ *be generated by*

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{d}_k,$$

where $\mathbf{d}_k$ *is the solution to (15.38) at* $\mathbf{x}_k$ *and where* α_k *minimizes* $P(\mathbf{x}_{k+1})$*. Assume that each* $\mathbf{x}_k \in \Omega$*. Then every limit point* $\overline{\mathbf{x}}$ *of* $\{\mathbf{x}_k\}$ *satisfies the first-order necessary conditions for the constrained minimization problem (15.37).*

Proof. The solution to a quadratic program depends continuously on the data, and hence the direction determined by the quadratic program (15.38) is a continuous function of $\mathbf{x}$. The function $P(\mathbf{x})$ is also continuous, and by Proposition 1, it follows that P is a descent function at every point that does not satisfy the first-order conditions. The result thus follows from the Global Convergence Theorem. ∎

In view of the above result, recursive quadratic programming in conjunction with the absolute-value penalty function is an attractive technique. There are, however, some difficulties to be kept in mind. First, the selection of the parameter α_k requires a one-dimensional search with respect to a nondifferentiable function. Thus the efficient curve-fitting search methods of Chap. 8 cannot be used without significant modification. Second, use of the absolute-value function requires an estimate of an upper bound for μ_j's, so that c can be selected properly. In some applications a suitable bound can be obtained from previous experience, but in general one must develop a method for revising the estimate upward when necessary.

Another potential difficulty with the quadratic programming approach above is that the quadratic program (15.38) may be infeasible at some point $\mathbf{x}_k$, even though the original problem (15.37) is feasible. If this happens, the method breaks down. However, see Exercise 8 for a method that avoids this problem.

The Quadratic Penalty Function

Another penalty function that is compatible with the modified Newton method approach is the standard quadratic penalty function. It has the added technical advantage that, since this penalty function is differentiable, it is possible to apply our earlier analytical principles to study the rate of convergence of the method. This leads to an analytical comparison of primal-dual methods with the methods of other chapters.

We shall restrict attention to the problem with equality constraints, since that is all that is required for a rate of convergence analysis. The method can be extended to problems with inequality constraints either directly or by an active set method. Thus we consider the problem

$$\begin{aligned} &\text{minimize} \quad f(\mathbf{x}) \\ &\text{subject to} \quad \mathbf{h}(\mathbf{x}) = \mathbf{0} \end{aligned} \tag{15.43}$$

and the standard quadratic penalty objective

$$P(\mathbf{x}) = f(\mathbf{x}) + \frac{1}{2}c|\mathbf{h}(\mathbf{x})|^2. \tag{15.44}$$

From the theory in Chap. 13, we know that minimization of the objective with a quadratic penalty function will *not* yield an exact solution to (15.43). In fact, the minimum of the penalty function (15.44) will have $c\mathbf{h}(\mathbf{x}) \simeq \boldsymbol{\lambda}$, where $\boldsymbol{\lambda}$ is the Lagrange multiplier of (15.43). Therefore, it seems appropriate in this case to consider the quadratic programming problem

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\mathbf{d}^T\mathbf{B}\mathbf{d} + \nabla f(\mathbf{x})\mathbf{d} \\ \text{subject to} \quad & \nabla\mathbf{h}(\mathbf{x})\mathbf{d} + \mathbf{h}(\mathbf{x}) = \hat{\boldsymbol{\lambda}}/c, \end{aligned} \tag{15.45}$$

where $\hat{\boldsymbol{\lambda}}$ is an estimate of the Lagrange multiplier of the original problem. A particularly good choice is

$$\hat{\boldsymbol{\lambda}} = [(1/c)\mathbf{I} + \mathbf{Q}]^{-1}[\mathbf{h}(\mathbf{x}) - \ \mathbf{A}\mathbf{B}^{-1}\nabla f(\mathbf{x})^T], \tag{15.46}$$

where $\mathbf{A} = \nabla\mathbf{h}(\mathbf{x})$, $\mathbf{Q} = \mathbf{A}\mathbf{B} - 1\mathbf{A}^T$ which is the Lagrange multiplier that would be obtained by the quadratic program with the penalty method. The proposed method requires that $\hat{\boldsymbol{\lambda}}$ be first estimated from (15.46) and then used in the quadratic programming problem (15.45).

The following proposition shows that this procedure produces a descent direction for the quadratic penalty objective.

Proposition 2. *For any $c > 0$, let $\mathbf{d}$, $\boldsymbol{\lambda}$ (with $\mathbf{d} \neq \mathbf{0}$) be a solution to the quadratic program (15.45). Then $\mathbf{d}$ is a descent direction of the function $P(\mathbf{x}) = f(\mathbf{x}) + (1/2)c|\mathbf{h}(\mathbf{x})|^2$.*

Proof. We have from the constraint equation

$$\mathbf{A}\mathbf{d} = (1/c)\hat{\boldsymbol{\lambda}} - \mathbf{h}(\mathbf{x}),$$

which yields

$$c\mathbf{A}^T\mathbf{A}\mathbf{d} = \mathbf{A}^T\hat{\boldsymbol{\lambda}} - \mathbf{c}\mathbf{A}^T\mathbf{h}(\mathbf{x}).$$

Solving the necessary conditions for (15.45) yields (see the top part of (15.9) for a similar expression with $\mathbf{Q} = \mathbf{B}$ there)

$$\mathbf{B}\mathbf{d} \ = \mathbf{A}^T\mathbf{Q}^{-1}[\mathbf{A}\mathbf{B}^{-1}\nabla\mathbf{f}(\mathbf{x})^T + (1/c)\hat{\boldsymbol{\lambda}} - \mathbf{h}(\mathbf{x})] - \nabla f(\mathbf{x})^T.$$

Therefore,

$$\begin{aligned} (\mathbf{B} + c\mathbf{A}^T\mathbf{A})\mathbf{d} &= \mathbf{A}^T\mathbf{Q}^{-1}[\mathbf{A}\mathbf{B}^{-1}\nabla f(\mathbf{x})^T - \mathbf{h}(\mathbf{x})] \\ &\quad + \mathbf{A}^T[(1/c)\mathbf{Q}^{-1} + \mathbf{I}]\hat{\boldsymbol{\lambda}} - \nabla f(\mathbf{x})^T - c\mathbf{A}^T\mathbf{h}(\mathbf{x}) \\ &= \mathbf{A}^T\mathbf{Q}^{-1}\{\mathbf{A}\mathbf{B}^{-1}\nabla f(\mathbf{x})^\tau - \mathbf{h}(\mathbf{x}) + ((1/c)\mathbf{I} + \mathbf{Q})\hat{\boldsymbol{\lambda}}\} \\ &\quad - \nabla f(\mathbf{x})^T - c\mathbf{A}^T\mathbf{h}(\mathbf{x}) \\ &= -\nabla f(\mathbf{x})^T - c\mathbf{A}^T\mathbf{h}(\mathbf{x}) = -\nabla P(\mathbf{x})^T. \end{aligned}$$

The matrix $(\mathbf{B} + c\mathbf{A}^T\mathbf{A})$ is positive definite for any $c \geqslant 0$. It follows that $\nabla P(\mathbf{x})\mathbf{d} < 0$. ∎

*15.6 *Rate of Convergence

It is now appropriate to apply the principles of convergence analysis that have been repeatedly emphasized in previous chapters to the recursive quadratic programming approach. We expect that, if this new approach is well founded, then the rate of convergence of the algorithm should be related to the familiar canonical rate, which we have learned is a fundamental measure of the complexity of the problem. If it is not so related, then some modification of the algorithm is probably required. Indeed, we shall find that a small but important modification *is* required.

From the proof of Proposition 2 of Sect. 15.5, we have the formula

$$(\mathbf{B} + c\mathbf{A}^T\mathbf{A})\mathbf{d} = -\nabla P(\mathbf{x})^T,$$

which can be written as

$$\mathbf{d} = -(\mathbf{B} + c\mathbf{A}^T\mathbf{A})^{-1}\nabla P(\mathbf{x})^T.$$

This shows that the method is a modified Newton method applied to the unconstrained minimization of $P(\mathbf{x})$. From the Modified Newton Method Theorem of Sect. 10.1, we see immediately that the rate of convergence is determined by the eigenvalues of the matrix that is the product of the coefficient matrix $(\mathbf{B} + c\mathbf{A}^T\mathbf{A})^{-1}$ and the Hessian of the function P at the solution point. The Hessian of P is $(\mathbf{L} + c\mathbf{A}^T\mathbf{A})$, where $\mathbf{L} = \mathbf{F}(\mathbf{x}) + c\mathbf{h}(\mathbf{x})^T\mathbf{H}(\mathbf{x})$. We know that the vector $c\mathbf{h}(\mathbf{x})$ at the solution of the penalty problem is equal to $\boldsymbol{\lambda}_c$, where $\nabla f(\mathbf{x}) + \boldsymbol{\lambda}_c^T\nabla\mathbf{h}(\mathbf{x}) = \mathbf{0}$. Therefore, the rate of convergence is determined by the eigenvalues of

$$(\mathbf{B} + c\mathbf{A}^T\mathbf{A})^{-1}(\mathbf{L} + c\mathbf{A}^T\mathbf{A}), \tag{15.47}$$

where all quantities are evaluated at the solution to the penalty problem and $\mathbf{L} = \mathbf{F} + \boldsymbol{\lambda}_c^T\mathbf{H}$. For large values of c, all quantities are approximately equal to the values at the optimal solution to the constrained problem.

Now what we wish to show is that as $c \to \infty$, the matrix (15.47) looks like $\mathbf{B}_M^{-1}\mathbf{L}_M$ on the subspace, M, and like the identity matrix on $M^{\perp}$, the subspace orthogonal to M. To do this in detail, let $\mathbf{C}$ be an $n \times (n-m)$ matrix whose columns form an orthonormal basis for M, the tangent subspace $\{\mathbf{x} : \mathbf{A}\mathbf{x} = \mathbf{0}\}$. Let $\mathbf{D} = \mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}$. Then $\mathbf{AC} = \mathbf{0}$, $AD = \mathbf{I}$, $\mathbf{C}^T\mathbf{C} = \mathbf{I}$, $\mathbf{C}^T\mathbf{D} = \mathbf{0}$.

The eigenvalues of $(\mathbf{B} + c\mathbf{A}^T\mathbf{A})^{-1}(\mathbf{L} + c\mathbf{A}^T\mathbf{A})$ are equal to those of

$$\begin{aligned}&[\mathbf{C},\ \mathbf{D}]^{-1}(\mathbf{B} + c\mathbf{A}^T\mathbf{A})^{-1}\{[\mathbf{C},\ \mathbf{D}]^T\}^{-1}[\mathbf{C},\ \mathbf{D}]^T(\mathbf{L} + c\mathbf{A}^T\mathbf{A})[\mathbf{C},\ \mathbf{D}]\\ &= \begin{bmatrix}\mathbf{C}^T\mathbf{BC} & \mathbf{C}^T\mathbf{BD}\\ \mathbf{D}^T\mathbf{BC} & \mathbf{D}^T\mathbf{BC} + c\mathbf{I}\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{C}^T\mathbf{LC} & \mathbf{C}^T\mathbf{LD}\\ \mathbf{D}^T\mathbf{LC} & \mathbf{D}^T\mathbf{LD} + c\mathbf{I}\end{bmatrix}.\end{aligned}$$

Now as $c \to \infty$, the matrix above approaches

$$\begin{bmatrix}\mathbf{B}_M^{-1}\mathbf{L}_M & \mathbf{B}_M\mathbf{C}^T(\mathbf{L} - \mathbf{B})\mathbf{D}\\ \mathbf{0} & \mathbf{I}\end{bmatrix},$$

where $\mathbf{B}_M = \mathbf{C}^T\mathbf{B}\mathbf{C}$, $\mathbf{L}_M = \mathbf{C}^T\mathbf{L}\mathbf{C}$ (see Exercise 6). The eigenvalues of this matrix are those of $\mathbf{B}_M^{-1}\mathbf{L}_M$ together with those of $\mathbf{I}$. This analysis leads directly to the following conclusion:

Theorem. *Let a, A be the smallest and largest eigenvalues, respectively, of $\mathbf{B}_M^{-1}\mathbf{L}_M$ and assume that $a \leqslant 1 \leqslant A$. Then the structured modified Newton method with quadratic penalty function has a rate of convergence no greater than $[(A-a)/(A+a)]^2$ as $c \to \infty$.*

In the special case of $\mathbf{B} = \mathbf{I}$, the rate in the above proposition is precisely the canonical rate, defined by the eigenvalues of $\mathbf{L}$ restricted to the tangent plane. It is important to note, however, that in order for the rate of the theorem to be achieved, the eigenvalues of $\mathbf{B}_M^{-1}\mathbf{L}_M$ must be spread around unity; if not, the rate will be poorer. Thus, even if $\mathbf{L}_M$ is well-conditioned, but the eigenvalues differ greatly from unity, the choice $\mathbf{B} = \mathbf{I}$ may be poor. This is an instance where proper scaling is vital. (We also point out that the above analysis is closely related to that of Sect. 13.4, where a similar conclusion is obtained.)

There is a geometric explanation for the scaling property. Take $\mathbf{B} = \mathbf{I}$ for simplicity. Then the direction of movement $\mathbf{d}$ is $\mathbf{d} = -\nabla f(\mathbf{x})^T + \mathbf{A}^T\boldsymbol{\lambda}$ for some $\boldsymbol{\lambda}$. Using the fact that the projected gradient is $\mathbf{p} = \nabla f(\mathbf{x})^T + \mathbf{A}^T\boldsymbol{\mu}$ for some $\boldsymbol{\mu}$, we see that $\mathbf{d} = -\mathbf{p} + \mathbf{A}^T(\boldsymbol{\lambda} + \boldsymbol{\mu})$. Thus $\mathbf{d}$ can be decomposed into two components: one in the direction of the projected negative gradient, the other in a direction orthogonal to the tangent plane (see Fig. 15.1). Ideally, these two components should be in proper proportions so that the constraint surface is reached at the same point as would be reached by minimization in the direction of the projected negative gradient. If they are not, convergence will be poor.

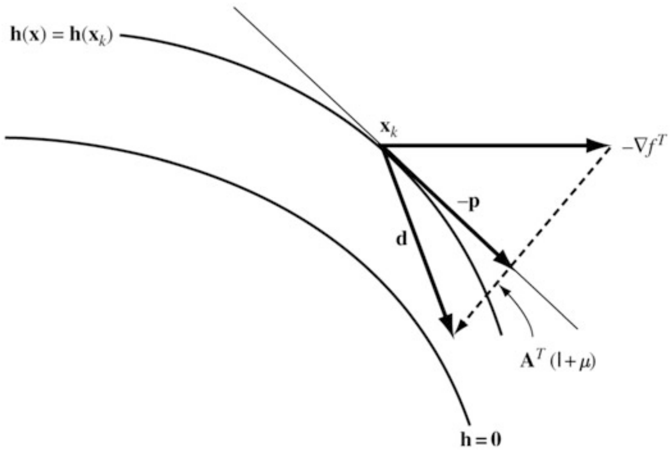

Fig. 15.1 Decomposition of the direction **d**

15.7 Primal-Dual Interior Point Methods

The primal-dual interior-point methods discussed for linear programming in Chap. 5 are, as mentioned there, closely related to the barrier methods presented in Chap. 13 and the primal-dual methods of the current chapter. They can be naturally extended to solve nonlinear programming problems while maintaining both theoretical and practical efficiency.

Consider the inequality constrained problem

$$\begin{aligned} \text{minimize} \quad & f(\mathbf{x}) \\ \text{subject to} \quad & \mathbf{A}\mathbf{x} = \mathbf{b}, \\ & \mathbf{g}(\mathbf{x}) \le \mathbf{0}, \end{aligned} \tag{15.48}$$

In general, a weakness of the active constraint method for such a problem is the combinatorial nature of determining which constraints should be active.

Logarithmic Barrier Function

A method that avoids the necessity to explicitly select a set of active constraints is based on the logarithmic barrier method, which solves a sequence of equality constrained minimization problems. Specifically,

$$\begin{aligned} \text{minimize} \quad & f(\mathbf{x}) - \boldsymbol{\mu} \sum_{i=1}^{p} \log(-g_j(\mathbf{x})) \\ \text{subject to} \quad & \mathbf{A}\mathbf{x} = \mathbf{b}, \end{aligned} \tag{15.49}$$

where $\mu = \mu^k > 0,\ k = 1, \ldots,, \mu^k > \mu^{k+1},\ \mu^k \to 0$. The μ^k**s** can be pre-determined. Typically, we have $\mu^{k+1} = \gamma\mu^k$ for some constant $0 < \gamma < 1$. Here, we also assume that the original problem has a feasible interior-point $\mathbf{x}^0$; that is,

$$\mathbf{A}\mathbf{x}^0 = \mathbf{b} \quad \text{and} \quad \mathbf{g}(\mathbf{x}^0) < \mathbf{0},$$

and **A** has full row rank.

For fixed μ, and using $s_i = \mu/g_i$, the first-order optimality conditions of the barrier problem (15.49) are:

$$\begin{aligned} -\mathbf{S}\mathbf{g}(\mathbf{x}) \quad & = \mu\mathbf{1} \\ \mathbf{A}\mathbf{x} \quad & = \mathbf{b} \\ -\mathbf{A}^T\mathbf{y} + \nabla f(\mathbf{x})^T + \nabla \mathbf{g}(\mathbf{x})^T\mathbf{s} \quad & = \mathbf{0}, \end{aligned} \tag{15.50}$$

where $\mathbf{S} = \text{diag}(\mathbf{s})$; that is, a diagonal matrix whose diagonal entries are **s**, and $\nabla\mathbf{g}(\mathbf{x})$ is the Jacobian matrix of $\mathbf{g}(\mathbf{x})$.

If $f(\mathbf{x})$ and $g_i(\mathbf{x})$ are convex functions for all i, $f(\mathbf{x}) - \mu \sum_i \log(-g_i(\mathbf{x}))$ is strictly convex in the interior of the feasible region, and the objective level set is bounded, then there is a unique minimizer for the barrier problem. Let $(\mathbf{x}(\mu) > \mathbf{0},\ \mathbf{y}(\mu),\ \mathbf{s}(\mu) > \mathbf{0})$ be the (unique) solution of (15.50). Then, these values form the *primal-dual central path* of (15.48):

$$\mathcal{C} = \{(\mathbf{x}(\mu),\ \mathbf{y}(\mu),\ \mathbf{s}(\mu) > \mathbf{0}) : 0 < \mu < \infty\}.$$

This can be summarized in the following theorem.

Theorem 1. *Let $(\mathbf{x}(\mu),\ \mathbf{y}(\mu),\ \mathbf{s}(\mu))$ be on the central path.*

i) If $f(\mathbf{x})$ and $g_i(\mathbf{x})$ are convex functions for all i, then $\mathbf{s}(\mu)$ is unique.
ii) Furthermore, if $f(\mathbf{x}) - \mu \sum_i \log(-g_i(\mathbf{x}))$ is strictly convex, $(\mathbf{x}(\mu),\ \mathbf{y}(\mu),\ \mathbf{s}(\mu))$ are unique, and they are bounded for $0 < \mu \leqslant \mu^0$ for any given $\mu^0 > 0$.
iii) For $0 < \mu' < \mu$, $f(\mathbf{x}(\mu')) < f(\mathbf{x}(\mu))$ if $\mathbf{x}(\mu') \neq \mathbf{x}(\mu)$.
iv) $(\mathbf{x}(\mu),\ \mathbf{y}(\mu),\ \mathbf{s}(\mu))$ converges to a point satisfying the first-order necessary conditions for a solution of (15.48) as $\mu \to 0$.

Once we have an approximate solution point $(\mathbf{x},\ \mathbf{y},\ \mathbf{s}) = (\mathbf{x}_k,\ \mathbf{y}_k,\ \mathbf{s}_k)$ for (15.50) for $\mu = \mu^k > 0$, we can again use the primal-dual methods described for linear programming to generate a new approximate solution to (15.50) for $\mu = \mu^{k+1} < \mu^k$. The Newton direction vectors $(\mathbf{d_x},\ \mathbf{d_y},\ \mathbf{d_s})$ is found from the system of linear equations:

$$\begin{aligned} -\mathbf{S}\nabla\mathbf{g}(\mathbf{x})\mathbf{d_x} - \mathbf{G}(\mathbf{x})\mathbf{d_s} &= \mu\mathbf{1} + \mathbf{S}\mathbf{g}(\mathbf{x}), \\ \mathbf{A}\mathbf{d_x} &= \mathbf{b} - \mathbf{A}\mathbf{x}, \\ -\mathbf{A}^T\mathbf{d_y} + \left(\nabla^2 f(\mathbf{x}) + \sum_i s_i \nabla^2 g_i(\mathbf{x})\right)\mathbf{d_x} & \\ +\nabla\mathbf{g}(\mathbf{x})^T\mathbf{d_s} &= \mathbf{A}^T\mathbf{y} - \nabla f(\mathbf{x})^T - \nabla\mathbf{g}(\mathbf{x})^T\mathbf{s}, \end{aligned} \tag{15.51}$$

where $\mathbf{G}(\mathbf{x}) = \mathrm{diag}(\mathbf{g}(\mathbf{x}))$. Then, the new iterate is update to:

$$(\mathbf{x}_{k+1},\ \mathbf{y}_{k+1}, \mathbf{s}_{k+1}) = (\mathbf{x}_k,\ \mathbf{y}_k, \mathbf{s}_k) + \alpha_k(\mathbf{d_x},\ \mathbf{d_y}, \mathbf{d_s})$$

for a stepsize α_k. Recently, this approach has also been used to find points satisfying the first-order conditions for problems when $f(\mathbf{x})$ and $g_i(\mathbf{x})$ are not generally convex functions.

Interior Point Method for Convex Quadratic Programming

Let $f(\mathbf{x}) = (1/2)\mathbf{x}^T\mathbf{Q}\mathbf{x} + \mathbf{c}^T\mathbf{x}$ and $g_i(\mathbf{x}) = -x_i$ for $i = 1, \ldots, n$, and consider the quadratic program

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\mathbf{x}^T\mathbf{Q}\mathbf{x} + \mathbf{c}^T\mathbf{x} \\ \text{subject to} \quad & \mathbf{A}\mathbf{x} = \mathbf{b}, \\ & \mathbf{x} \geqslant \mathbf{0}, \end{aligned} \tag{15.52}$$

where the given matrix $\mathbf{Q} \in E^{n\times n}$ is positive semidefinite (that is, the objective is a convex function), $\mathbf{A} \in E^{n\times m}$, $\mathbf{c} \in E^n$ and $\mathbf{b} \in E^m$. The problem reduces to finding $\mathbf{x} \in E^n$, $\mathbf{y} \in E^m$ and $\mathbf{s} \in E^n$ satisfying the following optimality conditions:

$$\begin{aligned} \mathbf{Sx} &= \mathbf{0} \\ \mathbf{Ax} &= \mathbf{b} \\ -\mathbf{A}^T\mathbf{y} + \mathbf{Qx} - \mathbf{s} &= -\mathbf{c} \\ (\mathbf{x},\ \mathbf{s}) &\geq \mathbf{0}. \end{aligned} \tag{15.53}$$

The optimality conditions with the logarithmic barrier function with parameter μ are be:

$$\begin{aligned} \mathbf{Sx} &= \mu\mathbf{1} \\ \mathbf{Ax} &= \mathbf{b} \\ -\mathbf{A}^T\mathbf{y} + \mathbf{Qx} - \mathbf{s} &= -\mathbf{c}. \end{aligned} \tag{15.54}$$

Note that the bottom two sets of constraints are linear equalities.

Thus, once we have an interior feasible point $(\mathbf{x},\ \mathbf{y},\ \mathbf{s})$ for (15.54), with $\boldsymbol{\mu} = \mathbf{x}^T\mathbf{s}/n$, we can apply Newton's method to compute a new (approximate) iterate $(\mathbf{x}^+,\ \mathbf{y}^+,\ \mathbf{s}^+)$ by solving for $(\mathbf{d_x},\ \mathbf{d_y},\ \mathbf{d_s})$ from the system of linear equations:

$$\begin{aligned} \mathbf{Sd_x} + \mathbf{Xd_s} &= \gamma\mu\mathbf{1} - \mathbf{Xs}, \\ \mathbf{Ad_x} &= \mathbf{0}, \\ -\mathbf{A}^T\mathbf{d_y} + \mathbf{Qd_x} - \mathbf{d_s} &= \mathbf{0}, \end{aligned} \tag{15.55}$$

where $\mathbf{X}$ and $\mathbf{S}$ are two diagonal matrices whose diagonal entries are $\mathbf{x} > \mathbf{0}$ and $\mathbf{s} > \mathbf{0}$, respectively. Here, γ is a fixed positive constant less than 1, which implies that our targeted μ is reduced by the factor γ at each step.

Potential Function as a Merit Function

For any interior feasible point $(\mathbf{x},\ \mathbf{y},\ \mathbf{s})$ of (15.52) and its dual, a suitable merit function is the potential function introduced in Chap. 5 for linear programming:

$$\psi_{n+\rho}(\mathbf{x},\ \mathbf{s}) = (n+\rho)\log(\mathbf{x}^T\mathbf{s}) - \sum_{j=1}^{n}\log(x_j s_j).$$

The main result for this is stated in the following theorem.

Theorem 2. *In solving (15.55) for* $(\mathbf{d_x},\ \mathbf{d_y},\ \mathbf{d_s})$*, let* $\gamma = n/(n+\rho) < 1$ *for fixed* $\rho \geqslant \sqrt{n}$ *and assign* $\mathbf{x}^+ = \mathbf{x} + \alpha\mathbf{d_x}$, $\mathbf{y}^+ = \mathbf{y} + \alpha\mathbf{d_y}$*, and* $\mathbf{s}^+ = \mathbf{s} + \alpha\mathbf{d}_s$ *where*

$$\alpha = \frac{\bar{\alpha}\sqrt{\min(\mathbf{Xs})}}{|(\mathbf{XS})^{-1/2}(\frac{\mathbf{x}^T\mathbf{s}}{n+\rho}\mathbf{1} - \mathbf{Xs})|},$$

where $\overline{\alpha}$ is any positive constant less than 1. (Again **X** *and* **S** *are matrices with components on the diagonal being those of* **x** *and* **s**, *respectively.) Then,*

$$\psi_{n+\rho}(\mathbf{x}^+,\ \mathbf{s}^+) - \psi_{n+\rho}(\mathbf{x},\ \mathbf{s}) \leqslant -\overline{\alpha}\sqrt{3/4} + \frac{\overline{\alpha}^2}{2(1-\overline{\alpha})}.$$

The proof of the theorem is also similar to that for linear programming; see Exercise 12. Notice that, since **Q** is positive semidefinite, we have

$$\mathbf{d_x}^T\mathbf{d_s} = (\mathbf{d_x},\ \mathbf{d_y})^T(\mathbf{d_s},\ \mathbf{0}) = \mathbf{d_x}^T\mathbf{Q}\mathbf{d_x} \geqslant 0$$

while $\mathbf{d}_\mathbf{x}^T\mathbf{d_s} = 0$ in the linear programming case.

We outline the algorithm here:

Given any interior feasible $(\mathbf{x}_0,\ \mathbf{y}_0,\ \mathbf{s}_0)$ *of* (15.52) *and its dual. Set* $\rho \geqslant \sqrt{n}$ *and* $k = 0$.

1. *Set* $(\mathbf{x},\ \mathbf{s}) = (\mathbf{x}_k,\ \mathbf{s}_k)$ *and* $\gamma = n/(n+\rho)$ *and compute* $(\mathbf{d_x},\ \mathbf{d_y},\ \mathbf{d_s})$ *from* (15.55).
2. *Let* $\mathbf{x}_{k+1} = \mathbf{x}_k + \overline{\alpha}\mathbf{d_x},\ \mathbf{y}_{k+1} = \mathbf{y}_k + \overline{\alpha}\mathbf{d_y}$, and $\mathbf{s}_{k+1} = \mathbf{s}_k + \overline{\alpha}\mathbf{d_s}$ *where*

$$\overline{\alpha} = \arg\min_{\alpha\geqslant 0}\psi_{n+\rho}(\mathbf{x}_k + \alpha\mathbf{d_x},\ \mathbf{s}_k + \alpha\mathbf{d_s}).$$

3. *Let* $k = k+1$. If $\mathbf{s}_k^T\mathbf{x}_k/\mathbf{s}_0^T\mathbf{x}_0 \leq \varepsilon$, *stop. Otherwise, return to Step 1.*

This algorithm exhibits an iteration complexity bound that is identical to that of linear programming expressed in Theorem 1, Sect. 5.6.

15.8 Summary

A constrained optimization problem can be solved by directly solving the equations that represent the first-order necessary conditions for a solution. For a quadratic programming problem with linear constraints, these equations are linear and thus can be solved by standard linear procedures. Quadratic programs with inequality constraints can be solved by an active set method in which the direction of movement is toward the solution of the corresponding equality constrained problem. This method will solve a quadratic program in a finite number of steps.

For general nonlinear programming problems, many of the standard methods for solving systems of equations can be adapted to the corresponding necessary equations. One class consists of first-order methods that move in a direction related to the residual (that is, the error) in the equations. Another class of methods is based on extending the method of conjugate directions to nonpositive-definite systems. Finally, a third class is based on Newton's method for solving systems of nonlinear equations, and solving a linearized version of the system at each iteration. Under appropriate assumptions, Newton's method has excellent global as well as local convergence properties, since the simple merit function, $\frac{1}{2}|\nabla f(\mathbf{x})+\boldsymbol{\lambda}^T\nabla\mathbf{h}(\mathbf{x})|^2+\frac{1}{2}|\mathbf{h}(\mathbf{x})|^2$, decreases in the Newton direction. An individual step of Newton's method is equivalent to solving a quadratic programming problem, and thus Newton's method can be extended to problems with inequality constraints through recursive quadratic programming.

More effective methods are developed by accounting for the special structure of the linearized version of the necessary conditions and by introducing approximations to the second-order information. In order to assure global convergence of these methods, a penalty (or merit) function must be specified that is compatible with the method of direction selection, in the sense that the direction is a direction of descent for the merit function. The absolute-value penalty function and the standard quadratic penalty function are both compatible with some versions of recursive quadratic programming.

The best of the primal-dual methods take full account of special structure, and are based on direction-finding procedures that are closely related to methods described in earlier chapters. It is not surprising therefore that the convergence properties of these methods are also closely related to those of other chapters. Again we find that the canonical rate is fundamental for properly designed first-order methods.

Interior point methods in the primal-dual model are very effective for treating problems with inequality constraints, for they avoid (or at least minimize) the difficulties associated with determining which constraints will be active at the solution. Applied to general nonlinear programming problems, these methods closely parallel the interior point methods for linear programming. There is again a central path, and Newton's method is a good way to follow the path.

15.9 Exercises

1. Solve the quadratic program

$$\begin{aligned} \text{minimize } & x^2 - xy + y^2 - 3x \\ \text{subject to } & x \geqslant 0 \\ & y \geqslant 0 \\ & x + y \leqslant 4 \end{aligned}$$

   by use of the active set method starting at $x = y = 0$.

2. Suppose $\mathbf{x}^*$, $\boldsymbol{\lambda}^*$ satisfy

$$\begin{aligned} \nabla f(\mathbf{x}^*) + \boldsymbol{\lambda}^{*T} \nabla \mathbf{h}(\mathbf{x}^*) &= \mathbf{0} \\ \mathbf{h}(\mathbf{x}^*) &= \mathbf{0}. \end{aligned}$$

   Let

$$\mathbf{C} = \begin{bmatrix} \mathbf{L}(\mathbf{x}^*, \boldsymbol{\lambda}^*) & \nabla \mathbf{h}(\mathbf{x}^*)^T \\ \nabla \mathbf{h}(\mathbf{x}^*) & \mathbf{0} \end{bmatrix}.$$

   Assume that $\mathbf{L}(\mathbf{x}^*, \boldsymbol{\lambda}^*)$ is positive definite and that $\nabla \mathbf{h}(\mathbf{x}^*)$ is of full rank.

(a) Show that the real part of each eigenvalue of **C** is positive.
(b) Using the result of Part (a), show that for some $\alpha > 0$ the iterative process

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha \nabla l(\mathbf{x}_k, \boldsymbol{\lambda}_k)^T \text{ and } \boldsymbol{\lambda}_{k+1} = \boldsymbol{\lambda}_k + \alpha \mathbf{h}(\mathbf{x}_k)$$

converges locally to $\mathbf{x}^*, \boldsymbol{\lambda}^*$. (That is, if started sufficiently close to $\mathbf{x}^*, \boldsymbol{\lambda}^*$, the process converges to $\mathbf{x}^*, \boldsymbol{\lambda}^*$.) *Hint*: Use Ostroski's Theorem: Let $\mathbf{A(z)}$ be a continuously differentiable mapping from E^p to E^p, assume $\mathbf{A}(\mathbf{z}^*) = \mathbf{0}$, and let $\nabla \mathbf{A}(\mathbf{z}^*)$ have all eigenvalues strictly inside the unit circle of the complex plane. Then $\mathbf{z}_{k+1} = \mathbf{z}_k + \mathbf{A}(\mathbf{z}_k)$ converges locally to $\mathbf{z}^*$.

3. Let **A** be a real symmetric matrix. A vector **x** is *singular* if $\mathbf{x}^T\mathbf{A}\mathbf{x} = 0$. A pair of vectors $\mathbf{x}, \mathbf{y}$ is a *hyperbolic pair* if both **x** and **y** are singular and $\mathbf{x}^T\mathbf{A}\mathbf{y} \neq 0$. Hyperbolic pairs can be used to generalize the conjugate gradient method to the nonpositive definite case.

(a) If $\mathbf{p}_k$ is singular, show that if $\mathbf{p}_{k+1}$ is defined as

$$\mathbf{p}_{k+1} = \mathbf{A}\mathbf{p}_k - \frac{(\mathbf{A}\mathbf{p}_k)^T\mathbf{A}^2\mathbf{p}_k}{2|\mathbf{A}\mathbf{p}_k|^2}\mathbf{p}_k,$$

then $\mathbf{p}_k, \mathbf{p}_{k+1}$ is a hyperbolic pair.
(b) Consider a modification of the conjugate gradient process of Sect. 8.1, where if $\mathbf{p}_k$ is singular, $\mathbf{p}_{k+1}$ is generated as above, and then

$$\begin{aligned}
\mathbf{x}_{k+1} &= \mathbf{x}_k + \alpha_k \mathbf{p}_k \\
\mathbf{x}_{k+2} &= \mathbf{x}_{k+1} + \alpha_{k+1} \mathbf{p}_{k+1} \\
\alpha_k &= \frac{\mathbf{r}_k^T\mathbf{p}_{k+1}}{\mathbf{p}_k^T\mathbf{A}\mathbf{p}_{k+1}}, \quad \alpha_{k+1} = \frac{\mathbf{r}_k^T\mathbf{p}_k}{\mathbf{p}_k^T\mathbf{A}\mathbf{p}_{k+1}} \\
\mathbf{p}_{k+2} &= \mathbf{r}_{k+2} - \frac{\mathbf{r}_{k+2}^T\mathbf{A}\mathbf{p}_{k+1}}{\mathbf{p}_k\mathbf{A}\mathbf{p}_{k+1}}\mathbf{p}_k.
\end{aligned}$$

Show that if $\mathbf{p}_{k+1}$ is the second member of a hyperbolic pair and $\mathbf{r}_k \neq \mathbf{0}$, then $\mathbf{x}_{k+2} \neq \mathbf{x}_{k+1}$, which means the process does not get "stuck."

4. Another method for solving a system $\mathbf{A}\mathbf{x} = \mathbf{b}$ when A is nonsingular and symmetric is the *conjugate residual method*. In this method the direction vectors are constructed to be an $\mathbf{A}^2$-orthogonalized version of the residuals $\mathbf{r}_k = \mathbf{b} - \mathbf{A}\mathbf{x}_k$. The error function $E(\mathbf{x}) = |\mathbf{A}\mathbf{x} - \mathbf{b}|^2$ decreases monotonically in this process. Since the directions are based on $\mathbf{r}_k$ rather than the gradient of E, which is $2\mathbf{A}r_k$, the method extends the simplicity of the conjugate gradient method by implicit use of the fact that $\mathbf{A}^2$ is positive definite. The method is this: Set $\mathbf{p}_1 = \mathbf{r}_1 = \mathbf{b} - \mathbf{A}\mathbf{x}_1$ and repeat the following steps, omitting (a, b) on the first step.

If $\alpha_{k-1} \neq 0$,

$$\mathbf{p}_k = \mathbf{r}_k - \beta_k \mathbf{p}_{k-1}, \quad \beta_k = \frac{\mathbf{r}_k^T \mathbf{A}^2 \mathbf{p}_{k-1}}{\mathbf{p}_{k-1}^T \mathbf{A}^2 \mathbf{p}_{k-1}}. \tag{15.56a}$$

If $\alpha_{k-1} = 0$,

$$\begin{aligned} \mathbf{p}_k &= \mathbf{A}\mathbf{r}_k - \gamma_k \mathbf{p}_{k-1} - \delta_k \mathbf{p}_{k-2} \\ \gamma_k &= \frac{\mathbf{r}_k^T \mathbf{A}^3 \mathbf{p}_{k-1}}{\mathbf{p}_{k-1}^T \mathbf{A}^2 \mathbf{p}_{k-1}}, \qquad \delta_k = \frac{\mathbf{r}_k^T \mathbf{A}^3 \mathbf{p}_{k-2}}{\mathbf{p}_{k-2}^T \mathbf{A}^3 \mathbf{p}_{k-2}} \end{aligned} \tag{15.56b}$$

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{p}_k, \qquad \alpha_k = \frac{\mathbf{r}_k^T \mathbf{A} \mathbf{p}_k}{\mathbf{p}_k^T \mathbf{A}^2 \mathbf{p}_k} \tag{15.56c}$$

$$\mathbf{r}_{k+1} = \mathbf{b} - \mathbf{A}\mathbf{x}_{k+1}. \tag{15.56d}$$

Show that the directions $\mathbf{p}_k$ are $\mathbf{A}^2$-orthogonal.

5. Consider the $(n + m)$-dimensional system of equations

$$\begin{bmatrix} \mathbf{L} & \mathbf{A}^T \\ \mathbf{A} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \boldsymbol{\lambda} \end{bmatrix} = \begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix}.$$

Suppose that $\mathbf{A} = [\mathbf{B}, \ \mathbf{C}]$, where $\mathbf{B}$ is $m \times m$ and invertible. Let $\mathbf{x} = (\mathbf{x}_B, \ \mathbf{x}_c)$, where $\mathbf{x}_B$ is the first m components of $\mathbf{x}$. The system can then be written

$$\begin{bmatrix} \mathbf{L_{BB}} & \mathbf{L_{BC}} & \mathbf{B}^T \\ \mathbf{L_{CB}} & \mathbf{L_{cc}} & \mathbf{C}^T \\ \mathbf{B} & \mathbf{C} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{x_B} \\ \mathbf{x_C} \\ \boldsymbol{\lambda} \end{bmatrix} = \begin{bmatrix} \mathbf{a_B} \\ \mathbf{a_C} \\ \mathbf{b} \end{bmatrix}$$

(a) Assume that $\mathbf{L}$ is positive definite on the tangent space $\{\mathbf{x} : \mathbf{A}\mathbf{x} = \mathbf{0}\}$. Derive an explicit statement equivalent to this assumption in terms of the positive definiteness of some $(n - m) \times (n - m)$ matrix.
(b) Solve the system in terms of the submatrices of the partitioned form.

6. Consider the partitioned square matrix $\mathbf{M}$ of the form

$$\mathbf{M} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}.$$

Show that

$$\mathbf{M}^{-1} = \begin{bmatrix} \mathbf{Q} & -\mathbf{Q}\mathbf{B}\mathbf{D}^{-1} \\ -\mathbf{D}^{-1}\mathbf{C}\mathbf{Q} & \mathbf{D}^{-1} + \mathbf{D}^{-1}\mathbf{C}\mathbf{Q}\mathbf{B}\mathbf{D}^{-1} \end{bmatrix},$$

where $\mathbf{Q} = (\mathbf{A} - \ \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}$, provided that all indicated inverses exist. Use this result to verify the rate of convergence result in Sect. 15.6.

7. For the problem

$$\begin{aligned} &\text{minimize} \quad f(\mathbf{x}) \\ &\text{subject to} \quad \mathbf{g}(\mathbf{x}) \leqslant \mathbf{0}, \end{aligned}$$

where $\mathbf{g}(\mathbf{x})$ is r-dimensional, define the penalty function

$$p(\mathbf{x}) = f(\mathbf{x}) + c\ \max\{0,\ g_1(\mathbf{x}),\ g_2(\mathbf{x}),\ \ldots,\ g_r(\mathbf{x})\}.$$

Let $\mathbf{d}$, $(\mathbf{d} \neq \mathbf{0})$ be a solution to the quadratic program

$$\begin{aligned} &\text{minimize} && \frac{1}{2}\mathbf{d}^T\mathbf{B}\mathbf{d} + \nabla f(\mathbf{x})\mathbf{d} \\ &\text{subject to} && \mathbf{g}(\mathbf{x}) + \nabla\mathbf{g}(\mathbf{x})\mathbf{d} \leqslant \mathbf{0}, \end{aligned}$$

where $\mathbf{B}$ is positive definite. Show that $\mathbf{d}$ is a descent direction for p for sufficiently large c.

8. Suppose the quadratic program of Exercise 7 is not feasible. In that case one may solve

$$\begin{aligned} &\text{minimize} && \frac{1}{2}\mathbf{d}^T\mathbf{B}\mathbf{d} + \nabla f(\mathbf{x})\mathbf{d} + c\xi \\ &\text{subject to} && \mathbf{g}(\mathbf{x}) + \nabla\mathbf{g}(\mathbf{x})\mathbf{d} \leqslant \xi\mathbf{1} \\ &&& \xi \geqslant 0. \end{aligned}$$

   (a) Show that if $\mathbf{d} \neq \mathbf{0}$ is a solution, then $\mathbf{d}$ is a descent direction for p.
   (b) If $\mathbf{d} = \mathbf{0}$ is a solution, show that $\mathbf{x}$ is a critical point of p in the sense that for any $\mathbf{d} \neq \mathbf{0}$, $p(\mathbf{x} + \alpha\mathbf{d}) > p(\mathbf{x}) + o(\alpha)$.

9. For the equality constrained problem, consider the function

$$\phi(\mathbf{x}) = f(\mathbf{x}) + \boldsymbol{\lambda}(\mathbf{x})^T\mathbf{h}(\mathbf{x}) + c\mathbf{h}(\mathbf{x})^T\mathbf{C}(\mathbf{x})\mathbf{C}(\mathbf{x})^T\mathbf{h}(\mathbf{x}),$$

where

$$\mathbf{C}(\mathbf{x}) = [\nabla\mathbf{h}(\mathbf{x})\nabla\mathbf{h}(\mathbf{x})^T]^{-1}\nabla\mathbf{h}(\mathbf{x}) \quad \text{and} \quad \boldsymbol{\lambda}(\mathbf{x}) = \mathbf{C}(\mathbf{x})\nabla f(\mathbf{x})^T.$$

   (a) Under standard assumptions on the original problem, show that for sufficiently large c, ϕ is (locally) an exact penalty function.
   (b) Show that $\phi(\mathbf{x})$ can be expressed as

$$\phi(\mathbf{x}) = \mathbf{f}(\mathbf{x}) + \boldsymbol{\pi}(\mathbf{x})^T\mathbf{h}(\mathbf{x}),$$

   where $\boldsymbol{\pi}(\mathbf{x})$ is the Lagrange multiplier of the problem

$$\begin{aligned} &\text{minimize} && \frac{1}{2}c\mathbf{d}^T\mathbf{d} + \nabla f(\mathbf{x})\mathbf{d} \\ &\text{subject to} && \nabla\mathbf{h}(\mathbf{x})\mathbf{d} + \mathbf{h}(\mathbf{x}) = \mathbf{0}. \end{aligned}$$

   (c) Indicate how ϕ can be defined for problems with inequality constraints.

10. Let $\{\mathbf{B}_k\}$ be a sequence of positive definite symmetric matrices, and assume that there are constants $a > 0$, $b > 0$ such that $a|\mathbf{x}|^2 \leqslant \mathbf{x}^T\mathbf{B}_k\mathbf{x} \leqslant b|\mathbf{x}|^2$ for all $\mathbf{x}$. Suppose that $\mathbf{B}$ is replaced by $\mathbf{B}_k$ in the kth step of the recursive quadratic

programming procedure of the theorem in Sect. 15.4. Show that the conclusions of that theorem are still valid. *Hint*: Note that the set of allowable $\mathbf{B}_k$'s is closed.

11. (Central path theorem) Prove the central path theorem, Theorem 1 of Sect. 15.7, for convex optimization.
12. Prove the potential reduction theorem, Theorem 2 of Sect. 15.7, for convex quadratic programming. This theorem can be generalized to non-quadratic convex objective functions $f(\mathbf{x})$ satisfying the following condition: let

$$u : (0, 1) \rightarrow (1, \infty)$$

be a monotone increasing function; then

$$|\mathbf{X}(\nabla f(\mathbf{x} + \mathbf{d_x}) - \nabla f(\mathbf{x}) - \nabla^2 f(\mathbf{x})\mathbf{d_x})|_1 \leq u(\alpha)\mathbf{d_x}^T \nabla f(\mathbf{x})\mathbf{d_x}$$

whenever

$$\mathbf{x} > 0, \ |\mathbf{X}^{-1}\mathbf{d_x}|_\infty \leq \alpha < 1.$$

Such condition is called the scaled Lipschitz condition in $\{\mathbf{x} : \mathbf{x} > \mathbf{0}\}$.

References

15.1 An early method for solving quadratic programming problems is the principal pivoting method of Dantzig and Wolfe; see Dantzig [D6]. For a discussion of factorization methods applied to quadratic programming, see Gill, Murray, and Wright [G7].

15.2–15.4 Arrow and Hurwicz [A9] proposed a continuous process (represented as a system of differential equations) for solving the Lagrange equations. This early paper showed the value of the simple merit function in attacking the equations. A formal discussion of the properties of the simple merit function may be found in Luenberger [L17]. The first-order method was examined in detail by Polak [P4]. Also see Zangwill [Z2] for an early analysis of a method for inequality constraints. The conjugate direction method was first extended to nonpositive definite cases by the use of hyperbolic pairs and then by employing conjugate residuals. (See Exercises 3 and 4, and Luenberger [L9, L11].) Additional methods with somewhat better numerical properties were later developed by Paige and Saunders [P1] and by Fletcher [F8]. It is perhaps surprising that Newton's method was analyzed in this form only recently, well after the development of the SOLVER method discussed in Sect. 15.2. For a comprehensive account of Newton methods, see Bertsekas, Chap. 4 [B11]. The SOLVER method was proposed by Wilson [W2] for convex programming problems and was later interpreted by Beale [B7]. Garcia-Palomares and Mangasarian [G3] proposed a quadratic programming approach to the solution of the first-order equations. See Fletcher [F10] for a good overview discussion.

15.5–15.6 The discovery that the absolute-value penalty function is compatible with recursive quadratic programming was made by Pshenichny (see Pshenichny and Danilin [P10]) and later by Han [H3], who also suggested that the method be combined with a quasi-Newton update procedure.

The development of recursive quadratic programming for the standard quadratic penalty function is due to Biggs [B14, B15]. The convergence rate analysis of Sect. 15.6 first appeared in the second edition of this text.

15.7 Many researchers have applied interior-point algorithms to convex quadratic problems. These algorithms can be divided into three groups: the primal algorithm, the dual algorithm, and the primal-dual algorithm. Relations among these algorithms can be seen in den Hertog [H6], Anstreicher et al [A6], Sun and Qi [S12], Tseng [T12], and Ye [Y3]. For results similar to those of Exercises 2,7, and 8, see Bertsekas [B11]. For discussion of Exercise 9, see Fletcher [F10].