Chapter 14
Duality and Dual Methods

We first derive the duality theory of for constrained optimization, which is based on our earlier *zero-order optimality conditions* and the Lagrangian relaxations. The variables of the dual are typically the Lagrange multipliers associated with the constraints in the primal problem—the original constrained optimization problem.

Thus, dual methods are based on the viewpoint that it is the Lagrange multipliers which are the fundamental unknowns associated with a constrained problem; once these multipliers are known determination of the solution point is simple (at least in some situations). Dual methods, therefore, do not attack the original constrained problem directly but instead attack an alternate problem, the dual problem, whose unknowns are the Lagrange multipliers of the first problem. For a problem with n variables and m equality constraints, dual methods thus work in the m-dimensional space of Lagrange multipliers. Because Lagrange multipliers measure sensitivities and hence often have meaningful intuitive interpretations as prices associated with constraint resources, searching for these multipliers, is often, in the context of a given practical problem, as appealing as searching for the values of the original problem variables.

The study of dual methods, and more particularly the introduction of the dual problem, precipitates some extensions of earlier concepts. One interesting feature of this chapter is the calculation of the Hessian of the dual problem and the discovery of a *dual canonical convergence ratio* associated with a constrained problem that governs the convergence of steepest ascent applied to the dual.

The convergence ratio theory lead to a popular method, the method of multipliers based on the augmented Lagrangian, in which the Hessian condition would be significantly improved to facilitate faster convergence.

The alternate direction method of multipliers is based on an idea resembling that in the coordinate descent method. Here, the gradient of the dual is calculated approximately in a block coordinate fashion using primal variables. This method is particularly effective for large-scale optimization.

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_14

Cutting plane algorithms, exceedingly elementary in principle, develop a series of ever-improving approximating linear programs, whose solutions converge to the solution of the original problem. The methods differ only in the manner by which an improved approximating problem is constructed once a solution to the old approximation is known. The theory associated with these algorithms is, unfortunately, scant and their convergence properties are not particularly attractive. They are, however, often very easy to implement.

14.1 Global Duality

Duality in nonlinear programming takes its most elegant form when it is formulated globally in terms of sets and hyperplanes that touch those sets. This theory makes clear the role of Lagrange multipliers as defining hyperplanes which can be considered as dual to points in a vector space. The theory provides a symmetry between primal and dual problems and this symmetry can be considered as perfect for convex problems. For non-convex problems the "imperfection" is made clear by the duality gap which has a simple geometric interpretation. The global theory, which is presented in this section, serves as useful background when later we specialize to a local duality theory that can be used even without convexity and which is central to the understanding of the convergence of dual algorithms.

As a counterpoint to Sect. 11.9 where equality constraints were considered before inequality constraints, here we shall first consider a problem with inequality constraints. In particular, consider the problem

$$\begin{aligned} &\text{minimize} && f(\mathbf{x}) \\ &\text{subject to} && \mathbf{g}(\mathbf{x}) \leq \mathbf{0} \\ &&& \mathbf{x} \in \boldsymbol{\Omega}. \end{aligned} \tag{14.1}$$

$\boldsymbol{\Omega} \subset E^n$ is a convex set, and the functions f and $\mathbf{g}$ are defined on $\boldsymbol{\Omega}$. The function $\mathbf{g}$ is p-dimensional. The problem is not necessarily convex, but we assume that there is a feasible point. Recall that the primal function associated with (14.1) is defined for $\mathbf{z} \in E^p$ as

$$\omega(\mathbf{z}) = \inf\{f(\mathbf{x}) : \mathbf{g}(\mathbf{x}) \leq \mathbf{z},\ \mathbf{x} \in \boldsymbol{\Omega}\}, \tag{14.2}$$

defined by letting the right hand side of inequality constraint take on arbitrary values. It is understood that (14.2) is defined on the set $D = \{\mathbf{z} : \mathbf{g}(\mathbf{x}) \leq \mathbf{z}$, for some $\mathbf{x} \in \boldsymbol{\Omega}\}$.

If problem (14.1) has a solution $\mathbf{x}^*$ with value $f^* = f(\mathbf{x}^*)$, then f^* is the point on the vertical axis in E^{p+1} where the primal function passes through the axis. If (14.1) does not have a solution, then $f^* = \inf\{f(\mathbf{x}) : \mathbf{g}(\mathbf{x}) \leq \mathbf{0},\ \mathbf{x} \in \boldsymbol{\Omega}\}$ is the intersection point.

The duality principle is derived from consideration of all hyperplanes that lie below the primal function. As illustrated in Fig. 14.1 the intercept with the vertical axis of such a hyperplanes lies below (or at) the value f^*.

To express this property we define the *dual function* defined on the positive cone in E^p as

$$\phi(\boldsymbol{\mu}) = \inf\{f(\mathbf{x}) + \boldsymbol{\mu}^T \mathbf{g}(\mathbf{x}) : \mathbf{x} \in \boldsymbol{\Omega}\}. \tag{14.3}$$

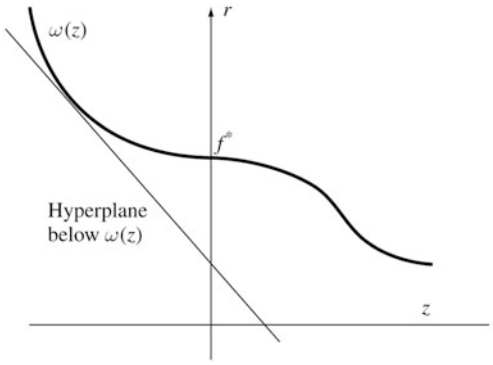

Fig. 14.1 Hyperplane below $\omega(\mathbf{z})$

In general, ϕ may not be finite throughout the positive orthant E^p_+ but the region where it is finite is convex.

Proposition 1. *The dual function is concave on the region where it is finite.*

Proof. Suppose $\boldsymbol{\mu}_1,\ \boldsymbol{\mu}_2$ are in the finite region, and let $0 \le \alpha \le 1$. Then

$$\begin{aligned}
\phi(\alpha\boldsymbol{\mu}_1 + (1 - \alpha\boldsymbol{\mu}_2)) &= \inf\{f(\mathbf{x}) + (\alpha\boldsymbol{\mu}_1 + (1-\alpha)\boldsymbol{\mu}_2)^T \mathbf{g}(\mathbf{x}) : \mathbf{x} \in \boldsymbol{\Omega}\} \\
&\ge \inf\{\alpha f(\mathbf{x}_1) + \alpha\boldsymbol{\mu}_1^T \mathbf{g}(\mathbf{x}_1) : \mathbf{x}_1 \in \boldsymbol{\Omega}\} \\
&+ \inf\{(1-\alpha) f(\mathbf{x}_2) + (1-\alpha)\boldsymbol{\mu}_2^T \mathbf{g}(\mathbf{x}_2) : \mathbf{x}_2 \in \boldsymbol{\Omega}\} \\
&= \alpha\phi(\boldsymbol{\mu}_1) + (1-\alpha)\phi(\boldsymbol{\mu}_2). \blacksquare
\end{aligned}$$

We define $\phi^* = \sup\{\phi(\boldsymbol{\mu}) : \boldsymbol{\mu} \ge \mathbf{0}\}$ where it is understood that the supremum is taken over the region where ϕ is finite. We can now state the weak form of global duality.

Weak Duality Proposition. $\phi^* \le f^*$.

Proof. For every $\boldsymbol{\mu} \ge \mathbf{0}$ we have

$$\begin{aligned}
\phi(\boldsymbol{\mu}) &= \inf\{f(\mathbf{x}) + \boldsymbol{\mu}^T \mathbf{g}(\mathbf{x}) : \mathbf{x} \in \boldsymbol{\Omega}\} \\
&\le \inf\{f(\mathbf{x}) + \boldsymbol{\mu}^T \mathbf{g}(\mathbf{x}) : \mathbf{g}(\mathbf{x}) \le \mathbf{0},\ \mathbf{x} \in \boldsymbol{\Omega}\} \\
&\le \inf\{f(\mathbf{x}) : \mathbf{g}(\mathbf{x}) \le \mathbf{0},\ \mathbf{x} \in \boldsymbol{\Omega}\} = f^*.
\end{aligned}$$

Taking the supremum over the left hand side gives $\phi^* \le f^*$. $\blacksquare$

Hence the dual function gives lower bounds on the optimal value f^*.

This dual function has a strong geometric interpretation. Consider a $p + 1$-dimensional vector $(1, \boldsymbol{\mu}) \in E^{p+1}$ with $\boldsymbol{\mu} \geq \mathbf{0}$ and a constant c. The set of vectors $(r, \mathbf{z})$ such that the inner product $(1, \boldsymbol{\mu})^T(r, \mathbf{z}) \equiv r + \boldsymbol{\mu}^T\mathbf{z} = c$ defines a hyperplane in E^{p+1}. Different values of c give different hyperplanes, all of which are parallel.

For a given $(1, \boldsymbol{\mu})$ we consider the lowest possible hyperplane of this form that just barely touches (supports) the region above the primal function of problem (14.1). Suppose $\mathbf{x}_1$ defines the touching point with values $r = f(\mathbf{x}_1)$ and $\mathbf{z} = \mathbf{g}(\mathbf{x}_1)$. Then $c = f(\mathbf{x}_1) + \boldsymbol{\mu}^T\mathbf{g}(\mathbf{x}_1) = \phi(\boldsymbol{\mu})$.

The hyperplane intersects the vertical axis at a point of the form $(r_0, \mathbf{0})$. This point also must satisfy $(1, \boldsymbol{\mu})^T(r_0, \mathbf{0}) = c = \phi(\boldsymbol{\mu})$. This gives $c = r_0$. Thus the intercept gives $\phi(\boldsymbol{\mu})$ directly. Thus the dual function at $\boldsymbol{\mu}$ is equal to the intercept of the hyperplane defined by $\boldsymbol{\mu}$ that just touches the epigraph of the primal function.

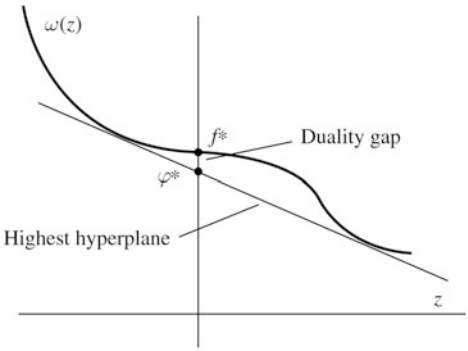

Fig. 14.2 The highest hyperplane

Furthermore, this intercept (and dual function value) is maximized by the Lagrange multiplier which corresponds to the largest possible intercept, at a point no higher than the optimal value f^*. See Fig. 14.2.

By introducing convexity assumptions, the foregoing analysis can be strengthened to give the strong duality theorem, with no duality gap when the intercept is at f^*. See Fig. 14.3.

We shall state the result for the more general problem that includes equality constraints of the form $\mathbf{h}(\mathbf{x}) = \mathbf{0}$, as in Sect. 11.9. Specifically, we consider the problem

$$\begin{aligned} &\text{maximize} && f(\mathbf{x}) \\ &\text{subject to} && \mathbf{h}(\mathbf{x}) = \mathbf{0},\ \mathbf{g}(\mathbf{x}) \leq \mathbf{0} \\ &&& \mathbf{x} \in \boldsymbol{\Omega} \end{aligned} \tag{14.4}$$

where $\mathbf{h}$ is affine of dimension m, $\mathbf{g}$ is convex of dimension p, and $\boldsymbol{\Omega}$ is a convex set.

In this case the dual function is

$$\phi(\boldsymbol{\lambda}, \boldsymbol{\mu}) = \inf\{f(\mathbf{x}) + \boldsymbol{\lambda}^T\mathbf{h}(\mathbf{x}) + \boldsymbol{\mu}^T\mathbf{g}(\mathbf{x}) : \mathbf{x} \in \boldsymbol{\Omega}\}.$$

And let

$$\phi^* = \sup\{\phi(\boldsymbol{\lambda},\ \boldsymbol{\mu}) : \boldsymbol{\lambda} \in E^m,\ \boldsymbol{\mu} \in E^p,\ \boldsymbol{\mu} \geq \mathbf{0}\}.$$

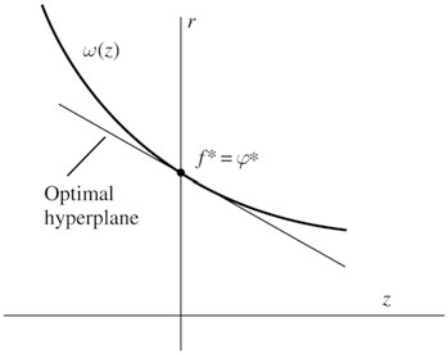

Fig. 14.3 The strong duality theorem. There is no duality gap

Strong Duality Theorem. *Suppose in the problem* (14.4), $\mathbf{h}$ *is affine and regular with respect to* $\boldsymbol{\Omega}$ *and there is a point* $\mathbf{x}_1 \in \boldsymbol{\Omega}$ *with that* $\mathbf{h}(\mathbf{x}) = \mathbf{0}$ *and* $\mathbf{g}(\mathbf{x}) < \mathbf{0}$.

Suppose the problem has solution $\mathbf{x}^*$ *with value* $f(\mathbf{x}^*) = f^*$. *Then for every* $\boldsymbol{\lambda}$ *and* $\boldsymbol{\mu} \geq \mathbf{0}$ *there holds*

$$\phi^* \leq f^*.$$

Furthermore, there are $\boldsymbol{\lambda},\ \boldsymbol{\mu} \geq \mathbf{0}$ *such that*

$$\phi(\boldsymbol{\lambda},\ \boldsymbol{\mu}) = f^*$$

and hence $\phi^* = f^*$. *Moreover, the* $\boldsymbol{\lambda}$ *and* $\boldsymbol{\mu}$ *above are Lagrange multipliers for the original problem.*

Proof. The proof follows almost immediately from the zero-order Lagrange theorem of Sect. 11.9. The Lagrange multipliers of that theorem give

$$\begin{aligned} f^* &= \max\{f(\mathbf{x}) + \boldsymbol{\lambda}^T \mathbf{h}(\mathbf{x}) + \boldsymbol{\mu}^T \mathbf{g}(\mathbf{x}) : \mathbf{x} \in \boldsymbol{\Omega}\} \\ &= \phi(\boldsymbol{\lambda},\ \boldsymbol{\mu}) \leq \phi^* \leq f^*. \end{aligned}$$

Equality must hold across the inequalities, which establishes the results. ∎

As a nice summary we can place the primal and dual problems together for the problem with inequality constraints.

Primal	Dual
$f^* = \text{minimize } \omega(\mathbf{z})$	$\phi^* = \text{maximize } \phi(\boldsymbol{\mu})$
$\text{subject to } \mathbf{z} \leq \mathbf{0}$	$\text{subject to } \boldsymbol{\mu} \geq \mathbf{0}.$

Example 1 (Quadratic Program). Consider the problem

$$\begin{aligned} &\text{minimize } \frac{1}{2}\mathbf{x}^T\mathbf{Q}\mathbf{x} \\ &\text{subject to } \mathbf{B}\mathbf{x} - \mathbf{b} \leq \mathbf{0}. \end{aligned} \tag{14.5}$$

The dual function is

$$\phi(\boldsymbol{\mu}) = \min_{\mathbf{x}} \frac{1}{2}\mathbf{x}^T\mathbf{Q}\mathbf{x} + \boldsymbol{\mu}^T(\mathbf{B}\mathbf{x} - \mathbf{b}).$$

This gives the necessary conditions

$$\mathbf{Q}\mathbf{x} + \mathbf{B}^T\boldsymbol{\mu} = \mathbf{0}$$

and hence $\mathbf{x} = -\mathbf{Q}^{-1}\mathbf{B}^T\boldsymbol{\mu}$. Substituting this into $\phi(\boldsymbol{\mu})$ gives

$$\phi(\boldsymbol{\mu}) = -\frac{1}{2}\boldsymbol{\mu}^T\mathbf{B}\mathbf{Q}^{-1}\mathbf{B}^T\boldsymbol{\mu} - \boldsymbol{\mu}^T\mathbf{b}.$$

Hence the dual problem is

$$\begin{aligned} &\text{maximize } -\frac{1}{2}\boldsymbol{\mu}^T\mathbf{B}\mathbf{Q}^{-1}\mathbf{B}^T\boldsymbol{\mu} - \boldsymbol{\mu}^T\mathbf{b} \\ &\text{subject to } \quad \boldsymbol{\mu} \geq \mathbf{0}, \end{aligned} \tag{14.6}$$

which is also a quadratic programming problem. If this problem is solved for $\boldsymbol{\mu}$, that $\boldsymbol{\mu}$ will be the Lagrange multiplier for the primal problem (14.5).

Note that the first-order conditions for the dual problem (14.6) imply

$$\boldsymbol{\mu}^T[-\mathbf{B}\mathbf{Q}^{-1}\mathbf{B}^T\boldsymbol{\mu} - \mathbf{b}] = 0,$$

which by substituting the formula for $\mathbf{x}$ is equivalent to

$$\boldsymbol{\mu}^T[\mathbf{B}\mathbf{x} - \mathbf{b}] = 0.$$

This is the complementary slackness condition for the original (primal) problem (14.5).

Example 2 (Integer Solutions). Duality gaps may arise if the object function or the constraint functions are not convex. A gap may also arise if the underlying set is not convex. This is characteristic, for example, of problems in which the components of the solution vector are constrained to be integers. For instance, consider the problem

$$\begin{aligned} &\text{minimize} \quad x_1^2 + 2x_2^2 \\ &\text{subject to} \quad x_1 + x_2 \geq 1/2 \\ &\qquad\qquad\ \ x_1,\ x_2 \text{ nonnegative integers} \end{aligned}$$

It is clear that the solution is $x_1 = 1,\ x_2 = 0$, with objective value $f^* = 1$. To put this problem in the standard form we have discussed, we write the constraint as

$$-x_1 - x_2 + 1/2 \le z, \quad \text{where } z = 0.$$

The primal function $\omega(z)$ is equal to 0 for $z \ge 1/2$ since then $x_1 = x_2 = 0$ is feasible. The entire primal function has steps as z steps negatively integer by integer, as shown in Fig. 14.4.

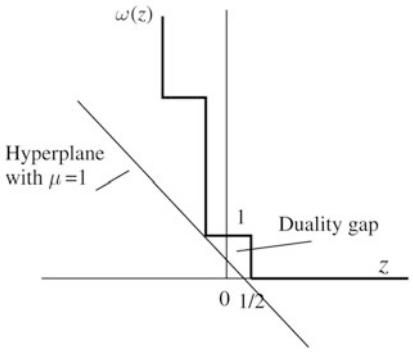

Fig. 14.4 Duality for an integer problem

The dual function is

$$\phi(\mu) = \max\{x_1^2 + x_2^2 - \lambda(x_1 + x_2 - 1/2)\}$$

where the maximum is taken with respect to the integer constraint. Analytically, the solution for small values of $\boldsymbol{\mu}$ is

$$\begin{aligned} \phi(\mu) &= \mu/2 && \text{for } 0 \le \mu \le 1, \\ &= 1 - \mu/2 && \text{for } 1 \le \mu \le 2, \\ &\vdots && \text{and more} \end{aligned}$$

The maximum value of $\phi(\mu)$ is the maximum intercept of the corresponding hyperplanes (lines, in this case) with the vertical axis. This occurs for $\mu = 1$ with a corresponding value of $\phi^* = \phi(1) = 1/2$. We have $\phi^* < f^*$ and the difference $f^* - \phi^* = 1/2$ is the duality gap.

14.2 Local Duality

In practice the mechanics of duality are frequently carried out locally, by setting derivatives to zero, or moving in the direction of a gradient. For these operations the beautiful global theory can in large measure be replaced by a weaker but often

more useful local theory. This theory requires a minimum of convexity assumptions defined locally. We present such a theory in this section, since it is in keeping with the spirit of the earlier chapters and is perhaps the simplest way to develop computationally useful duality results.

As often done before for convenience, we again consider nonlinear programming problems of the form

$$\begin{aligned} &\text{minimize} \quad f(\mathbf{x}) \\ &\text{subject to} \quad \mathbf{h}(\mathbf{x}) = \mathbf{0}, \end{aligned} \tag{14.7}$$

where $\mathbf{x} \in E^n$, $\mathbf{h}(\mathbf{x}) \in E^n$ and f, $\mathbf{h} \in C^2$. Global convexity is not assumed here. Everything we do can be easily extended to problems having inequality as well as equality constraints, for the price of a somewhat more involved notation.

We focus attention on a local solution $\mathbf{x}^*$ of (14.7). Assuming that $\mathbf{x}^*$ is a regular point of the constraints, then, as we know, there will be a corresponding Lagrange multiplier (row) vector $\boldsymbol{\lambda}^*$ such that

$$\nabla f(\mathbf{x}^*) + (\boldsymbol{\lambda}^*)^T \nabla \mathbf{h}(\mathbf{x}^*) = \mathbf{0}, \tag{14.8}$$

and the Hessian of the Lagrangian

$$\mathbf{L}(\mathbf{x}^*) = \mathbf{F}(\mathbf{x}^*) + (\boldsymbol{\lambda}^*)^T \mathbf{H}(\mathbf{x}^*) \tag{14.9}$$

must be positive semidefinite on the tangent subspace

$$M = \{\mathbf{x} : \nabla \mathbf{h}(\mathbf{x}^*)\mathbf{x} = \mathbf{0}\}.$$

At this point we introduce the special local convexity assumption necessary for the development of the local duality theory. Specifically, we assume that the Hessian $\mathbf{L}(\mathbf{x}^*)$ is positive definite. Of course, it should be emphasized that by this we mean $\mathbf{L}(\mathbf{x}^*)$ is positive definite on the whole space E^n, not just on the subspace M. The assumption guarantees that the Lagrangian $l(\mathbf{x}) = f(\mathbf{x}) + (\boldsymbol{\lambda}^*)^T \mathbf{h}(\mathbf{x})$ is locally convex at $\mathbf{x}^*$.

With this assumption, the point $\mathbf{x}^*$ is not only a local solution to the constrained problem (14.7); it is also a local solution to the unconstrained problem

$$\text{minimize} \quad f(\mathbf{x}) + (\boldsymbol{\lambda}^*)^T \mathbf{h}(\mathbf{x}), \tag{14.10}$$

since it satisfies the first- and second-order sufficiency conditions for a local minimum point. Furthermore, for any $\boldsymbol{\lambda}$ sufficiently close to $\boldsymbol{\lambda}^*$ the function $f(\mathbf{x}) + \boldsymbol{\lambda}^T \mathbf{h}(\mathbf{x})$ will have a local minimum point at a point $\mathbf{x}$ near $\mathbf{x}^*$. This follows by noting that, by the Implicit Function Theorem, the equation

$$\nabla f(\mathbf{x}) + \boldsymbol{\lambda}^T \nabla \mathbf{h}(\mathbf{x}) = \mathbf{0} \tag{14.11}$$

has a solution $\mathbf{x}$ near $\mathbf{x}^*$ when $\boldsymbol{\lambda}$ is near $\boldsymbol{\lambda}^*$, because $\mathbf{L}^*$ is nonsingular; and by the fact that, at this solution $\mathbf{x}$, the Hessian $\mathbf{F}(\mathbf{x}) + \boldsymbol{\lambda}^T\mathbf{H}(\mathbf{x})$ is positive definite. Thus locally there is a unique correspondence between $\boldsymbol{\lambda}$ and $\mathbf{x}$ through solution of the unconstrained problem

$$\text{minimize} \quad f(\mathbf{x}) + \boldsymbol{\lambda}^T\mathbf{h}(\mathbf{x}). \tag{14.12}$$

Furthermore, this correspondence is continuously differentiable.

Near $\boldsymbol{\lambda}^*$ we define the *dual function* ϕ by the equation

$$\phi(\boldsymbol{\lambda}) = \text{minimum}\,[f(\mathbf{x}) + \boldsymbol{\lambda}^T\mathbf{h}(\mathbf{x})], \tag{14.13}$$

where here it is understood that the minimum is taken locally with respect to $\mathbf{x}$ near $\mathbf{x}^*$. We are then able to show (and will do so below) that locally the original constrained problem (14.7) is equivalent to unconstrained local maximization of the dual function ϕ with respect to $\boldsymbol{\lambda}$. Hence we establish an equivalence between a constrained problem in $\mathbf{x}$ and an unconstrained problem in $\boldsymbol{\lambda}$.

To establish the duality relation we must prove two important lemmas. In the statements below we denote by $\mathbf{x}(\boldsymbol{\lambda})$ the unique solution to (14.12) in the neighborhood of $\mathbf{x}^*$.

Lemma 1. *The dual function ϕ has gradient*

$$\nabla\phi(\boldsymbol{\lambda}) = \mathbf{h}(\mathbf{x}(\boldsymbol{\lambda}))^T \tag{14.14}$$

Proof. We have explicitly, from (14.13),

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

Thus

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

Since the first term on the right vanishes by definition of $\mathbf{x}(\boldsymbol{\lambda})$, we obtain (14.14). ∎

Lemma 1 is of extreme practical importance, since it shows that the gradient of the dual function is simple to calculate. Once the dual function itself is evaluated, by minimization with respect to $\mathbf{x}$, the corresponding $\mathbf{h}(\mathbf{x})^T$, which is the gradient, can be evaluated without further calculation.

The Hessian of the dual function can be expressed in terms of the Hessian of the Lagrangian. We use the notation $\mathbf{L}(\mathbf{x},\ \boldsymbol{\lambda}) = \mathbf{F}(\mathbf{x}) + \boldsymbol{\lambda}^T\mathbf{H}(\mathbf{x})$, explicitly indicating the dependence on $\boldsymbol{\lambda}$. (We continue to use $\mathbf{L}(\mathbf{x}^*)$ when $\boldsymbol{\lambda} = \boldsymbol{\lambda}^*$ is understood.) We then have the following lemma.

Lemma 2. *The Hessian of the dual function is*

$$\boldsymbol{\Phi}(\boldsymbol{\lambda}) = -\nabla\mathbf{h}(\mathbf{x}(\boldsymbol{\lambda}))\mathbf{L}^{-1}(\mathbf{x}(\boldsymbol{\lambda}),\ \boldsymbol{\lambda})\nabla\mathbf{h}(\mathbf{x}(\boldsymbol{\lambda}))^T. \tag{14.15}$$

Proof. The Hessian is the derivative of the gradient. Thus, by Lemma 1,

$$\Phi(\boldsymbol{\lambda}) = \nabla\mathbf{h}(\mathbf{x}(\boldsymbol{\lambda}))\nabla\mathbf{x}(\boldsymbol{\lambda}). \tag{14.16}$$

By definition we have

$$\nabla f(\mathbf{x}(\boldsymbol{\lambda})) + \boldsymbol{\lambda}^T \nabla \mathbf{h}(\mathbf{x}(\boldsymbol{\lambda})) = \mathbf{0},$$

and differentiating this with respect to $\boldsymbol{\lambda}$ we obtain

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

Solving for $\nabla \mathbf{x}(\boldsymbol{\lambda})$ and substituting in (14.16) we obtain (14.15). ∎

Since $\mathbf{L}^{-1}(\mathbf{x}(\boldsymbol{\lambda}))$ is positive definite, and since $\nabla \mathbf{h}(\mathbf{x}(\boldsymbol{\lambda}))$ is of full rank near $\mathbf{x}^*$, we have as an immediate consequence of Lemma 2 that the $m \times m$ Hessian of ϕ is negative definite. As might be expected, this Hessian plays a dominant role in the analysis of dual methods.

Local Duality Theorem. Suppose that the problem

$$\begin{aligned} &\text{minimize} \quad f(\mathbf{x}) \\ &\text{subject to} \quad \mathbf{h}(\mathbf{x}) = \mathbf{0} \end{aligned} \tag{14.17}$$

has a local solution at $\mathbf{x}^*$ with *corresponding value* r^* and *Lagrange multiplier* $\boldsymbol{\lambda}^*$. *Suppose also that* $\mathbf{x}^*$ *is a regular point of the constraints and that the corresponding Hessian of the Lagrangian* $\mathbf{L}^* = \mathbf{L}(\mathbf{x}^*)$ *is positive definite. Then the dual problem*

$$\text{maximize} \quad \phi(\boldsymbol{\lambda}) \tag{14.18}$$

has a local solution at $\boldsymbol{\lambda}^*$ *with corresponding value* r^* *and* $\mathbf{x}^*$ *as the point corresponding to* $\boldsymbol{\lambda}^*$ *in the definition of* ϕ.

Proof. It is clear that $\mathbf{x}^*$ corresponds to $\boldsymbol{\lambda}^*$ in the definition of ϕ. Now at $\boldsymbol{\lambda}^*$ we have by Lemma 1

$$\nabla \phi(\boldsymbol{\lambda}^*) = \mathbf{h}(\mathbf{x}^*)^T = \mathbf{0},$$

and by Lemma 2 the Hessian of ϕ is negative definite. Thus $\boldsymbol{\lambda}^*$ satisfies the first- and second-order sufficiency conditions for an unconstrained maximum point of ϕ. The corresponding value $\phi(\boldsymbol{\lambda}^*)$ is found from the definition of ϕ to be r^*. ∎

Example 1. Consider the problem in two variables

$$\begin{aligned} &\text{minimize} \quad -xy \\ &\text{subject to} \quad (x-3)^2 + y^2 = 5. \end{aligned}$$

The first-order necessary conditions are

$$\begin{aligned} -y + (2x - 6)\lambda &= 0 \\ -x + 2y\lambda &= 0 \end{aligned}$$

together with the constraint. These equations have a solution at

$$x = 4, \quad y = 2, \quad \lambda = 1.$$

The Hessian of the corresponding Lagrangian is

$$\mathbf{L} = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}.$$

Since this is positive definite, we conclude that the solution obtained is a local minimum. (It can be shown, in fact, that it is the global solution.)

Since $\mathbf{L}$ is positive definite, we can apply the local duality theory near this solution. We define

$$\phi(\lambda) = \min\{-xy + \lambda[(x-3)^2 + y^2 - 5]\},$$

which leads to

$$\phi(\lambda) = \frac{4\lambda + 4\lambda^3 - 80\lambda^5}{(4\lambda^2 - 1)^2}$$

valid for $\lambda > \frac{1}{2}$. It can be verified that ϕ has a local maximum at $\lambda = 1$.

Inequality Constraints

For problems having inequality constraints as well as equality constraints the above development requires only minor modification. Consider the problem

$$\begin{aligned} &\text{minimize} && f(\mathbf{x}) \\ &\text{subject to} && \mathbf{h}(\mathbf{x}) = \mathbf{0} \\ &&& \mathbf{g}(\mathbf{x}) \leq \mathbf{0}, \end{aligned} \tag{14.19}$$

where $\mathbf{g}(\mathbf{x}) \in E^p$, $\mathbf{g} \in C^2$ and everything else is as before. Suppose $\mathbf{x}^*$ is a local solution of (14.19) and is a regular point of the constraints. Then, as we know, there are Lagrange multipliers $\boldsymbol{\lambda}^*$ and $\boldsymbol{\mu}^* \geq \mathbf{0}$ such that

$$\nabla f(\mathbf{x}^*) + (\boldsymbol{\lambda}^*)^T \nabla \mathbf{h}(\mathbf{x}^*) + (\boldsymbol{\mu}^*)^T \nabla \mathbf{g}(\mathbf{x}^*) = \mathbf{0} \tag{14.20}$$

$$(\boldsymbol{\mu}^*)^T \mathbf{g}(\mathbf{x}^*) = 0. \tag{14.21}$$

We impose the local convexity assumptions that the Hessian of the Lagrangian

$$\mathbf{L}(\mathbf{x}^*) = \mathbf{F}(\mathbf{x}^*) + (\boldsymbol{\lambda}^*)^T \mathbf{H}(\mathbf{x}^*) + (\boldsymbol{\mu}^*)^T \mathbf{G}(\mathbf{x}^*) \tag{14.22}$$

is positive definite (on the whole space).

For $\boldsymbol{\lambda}$ and $\boldsymbol{\mu} \geq \mathbf{0}$ near $\boldsymbol{\lambda}^*$ and $\boldsymbol{\mu}^*$ we define the dual function

$$\phi(\boldsymbol{\lambda},\ \boldsymbol{\mu}) = \min[f(\mathbf{x}) + \boldsymbol{\lambda}^T \mathbf{h}(\mathbf{x}) + \boldsymbol{\mu}^T \mathbf{g}(\mathbf{x})], \tag{14.23}$$

where the minimum is taken locally near $\mathbf{x}^*$. Then, it is easy to show, paralleling the development above for equality constraints, that ϕ achieves a local maximum with respect to $\boldsymbol{\lambda},\ \boldsymbol{\mu} \geq \mathbf{0}$ at $\boldsymbol{\lambda}^*,\ \boldsymbol{\mu}^*$.

Partial Duality

It is not necessary to include the Lagrange multipliers of all the constraints of a problem in the definition of the dual function. In general, if the local convexity assumption holds, local duality can be defined with respect to any subset of functional constraints. Thus, for example, in the problem

$$\begin{aligned} &\text{minimize} && f(\mathbf{x}) \\ &\text{subject to} && \mathbf{h}(\mathbf{x}) = \mathbf{0} \\ & && \mathbf{g}(\mathbf{x}) \leq \mathbf{0}, \end{aligned} \tag{14.24}$$

we might define the dual function with respect to only the equality constraints. In this case we would define

$$\phi(\boldsymbol{\lambda}) = \min_{\mathbf{g}(\mathbf{x}) \leq \mathbf{0}} \{f(\mathbf{x}) + \boldsymbol{\lambda}^T \mathbf{h}(\mathbf{x})\}, \tag{14.25}$$

where the minimum is taken locally near the solution $\mathbf{x}^*$ but constrained by the remaining constraints $\mathbf{g}(\mathbf{x}) \leq \mathbf{0}$. Again, the dual function defined in this way will achieve a local maximum at the optimal Lagrange multiplier $\boldsymbol{\lambda}^*$. The partial dual is especially useful when constraints $\mathbf{g}(\mathbf{x}) \leq \mathbf{0}$ are simple such as $\mathbf{x} \leq \mathbf{0}$ or in a box.

14.3 Canonical Convergence Rate of Dual Steepest Ascent

Constrained problems satisfying the local convexity assumption can be solved by solving the associated unconstrained dual problem, and any of the standard algorithms discussed in Chaps. 7 through 10 can be used for this purpose. Of course, the method that suggests itself immediately is the method of steepest ascent. It can be implemented by noting that, according to Lemma 1. Section 14.2, the gradient of ϕ is available almost without cost once ϕ itself is evaluated. Without some special properties, however, the method as a whole can be extremely costly to execute, since every evaluation of ϕ requires the solution of an unconstrained problem in the unknown $\mathbf{x}$. Nevertheless, as shown in the next section, many important problems do have a structure which is suited to this approach.

The method of steepest ascent, and other gradient-based algorithms, when applied to the dual problem will have a convergence rate governed by the eigenvalue structure of the Hessian of the dual function ϕ. At the Lagrange multiplier $\boldsymbol{\lambda}^*$ corresponding to a solution $\mathbf{x}^*$ this Hessian is (according to Lemma 2, Sect. 13.1)

$$\Phi = -\nabla \mathbf{h}(\mathbf{x}^*)(\mathbf{L}^*)^{-1} \nabla \mathbf{h}(\mathbf{x}^*)^T.$$

This expression shows that Φ is in some sense a restriction of the matrix $(\mathbf{L}^*)^{-1}$ to the subspace spanned by the gradients of the constraint functions, which is the orthogonal complement of the tangent subspace M. This restriction is not the orthogonal

restriction of $(\mathbf{L}^*)^{-1}$ onto the complement of M since the particular representation of the constraints affects the structure of the Hessian. We see, however, that while the convergence of primal methods is governed by the restriction of $\mathbf{L}^*$ to M, the convergence of dual methods is governed by a restriction of $(\mathbf{L}^*)^{-1}$ to the orthogonal complement of M.

The *dual canonical convergence rate* associated with the original constrained problem, which is the rate of convergence of steepest ascent applied to the dual, is $(B - b)^2/(B + b)^2$ where b and B are, respectively, the smallest and largest eigenvalues of

$$-\Phi = \nabla\mathbf{h}(\mathbf{x}^*)(\mathbf{L}^*)^{-1}\nabla\mathbf{h}(\mathbf{x}^*)^T.$$

For locally convex programming problems, this rate is as important as the primal canonical rate.

Scaling

We conclude this section by pointing out a kind of complementarity that exists between the primal and dual rates. Suppose one calculates the primal and dual canonical rates associated with the locally convex constrained problem

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

If a change of primal variables $\mathbf{x}$ is introduced, the primal rate will in general change but the dual rate will not. On the other hand, if the constraints are transformed (by replacing them by $\mathbf{Th}(\mathbf{x}) = \mathbf{0}$ where $\mathbf{T}$ is a nonsingular $m \times m$ matrix), the dual rate will change but the primal rate will not.

14.4 Separable Problems and Their Duals

A structure that arises frequently in mathematical programming applications is that of the separable problem:

$$\text{minimize} \quad \sum_{i=1}^{q} f_i(\mathbf{x}_i) \tag{14.26}$$

$$\text{subject to} \quad \sum_{i=1}^{q} \mathbf{h}_i(\mathbf{x}_i) = \mathbf{0} \tag{14.27}$$

$$\sum_{i=1}^{q} \mathbf{g}_i(\mathbf{x}_i) \leq \mathbf{0}. \tag{14.28}$$

In this formulation the components of the n-vector $\mathbf{x}$ are partitioned into q disjoint groups, $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_q)$ where the groups may or may not have the same number of components. Both the objective function and the constraints separate into sums of functions of the individual groups. For each i, the functions f_i, $\mathbf{h}_i$, and $\mathbf{g}_i$ are twice continuously differentiable functions of dimensions 1, m, and p, respectively.

Example 1. Suppose that we have a fixed budget of, say, A dollars that may be allocated among n activities. If x_i dollars is allocated to the ith activity, then there will be a benefit (measured in some units) of $f_i(x_i)$. To obtain the maximum benefit within our budget, we solve the separable problem

$$\begin{aligned} &\text{maximize} && \sum_{i=1}^{n} f_i(x_i) \\ &\text{subject to} && \sum_{i=1}^{n} x_i \leqslant A \\ & && x_i \geqslant 0. \end{aligned} \tag{14.29}$$

In the example $\mathbf{x}$ is partitioned into its individual components.

Example 2. Problems involving a series of decisions made at distinct times are often separable. For illustration, consider the problem of scheduling water release through a dam to produce as much electric power as possible over a given time interval while satisfying constraints on acceptable water levels. A discrete-time model of this problem is to

$$\begin{aligned} &\text{maximize} && \sum_{k=1}^{N} f(y(k),\ u(k)) \\ &\text{subject to} && y(k) = y(k-1) - u(k) + s(k),\ k = 1, \ldots, N \\ & && c \leqslant y(k) \leqslant d,\ k = 1, \ldots, N \\ & && 0 \leqslant u(k),\ k = 1, \ldots, N. \end{aligned}$$

Here $y(k)$ represents the water volume behind the dam at the end of period k, $u(k)$ represents the volume flow through the dam during period k, and $s(k)$ is the volume flowing into the lake behind the dam during period k from upper streams. The function f gives the power generation, and c and d are bounds on lake volume. The initial volume $y(0)$ is given.

In this example we consider $\mathbf{x}$ as the $2N$-dimensional vector of unknowns $y(k)$, $u(k)$, $k = 1, 2, \ldots, N$. This vector is partitioned into the pairs $\mathbf{x}_k = (y(k), u(k))$. The objective function is then clearly in separable form. The constraints can be viewed as being in the form (14.27) with $\mathbf{h}_k(\mathbf{x}_k)$ having dimension N and such that $\mathbf{h}_k(\mathbf{x}_k)$ is identically zero except in the k and $k + 1$ components.

Decomposition

Separable problems are ideally suited to dual methods, because the required unconstrained minimization decomposes into small subproblems. To see this we recall that the generally most difficult aspect of a dual method is evaluation of the dual function. For a separable problem, if we associate $\boldsymbol{\lambda}$ with the equality constraints (14.27) and $\boldsymbol{\mu} \geqslant \mathbf{0}$ with the inequality constraints (14.28), the required dual function is

$$\phi(\boldsymbol{\lambda},\ \boldsymbol{\mu}) = \min \sum_{i=1}^{q} \left(f_i(\mathbf{x}_i) + \boldsymbol{\lambda}^T \mathbf{h}_i(\mathbf{x}_i) + \boldsymbol{\mu}^T \mathbf{g}_i(\mathbf{x}_i) \right).$$

This minimization problem decomposes into the q separate problems

$$\min_{\mathbf{x}_i}\ f_i(\mathbf{x}_i) + \boldsymbol{\lambda}^T \mathbf{h}_i(\mathbf{x}_i) + \boldsymbol{\mu}^T \mathbf{g}_i(\mathbf{x}_i).$$

The solution of these subproblems can usually be accomplished relatively efficiently, since they are of smaller dimension than the original problem.

Example 3. In Example 1 using duality with respect to the budget constraint, the ith subproblem becomes, for $\mu > 0$

$$\max_{x_i \geqslant 0} f_i(x_i) - \mu x_i,$$

which is only a one-dimensional problem. It can be interpreted as setting a benefit value μ for dollars and then maximizing total benefit from activity i, accounting for the dollar expenditure.

Example 4. In Example 1 using duality with respect to the equality constraints we denote the dual variables by $\lambda(k),\ k = 1, 2, \ldots, N$. The kth subproblem becomes

$$\max_{\substack{c \leqslant y(k) \leqslant d \\ 0 \leqslant u(k)}} \{ f(y(k),\ u(k)) + [\lambda(k+1) - \lambda(k)]y(k) - \lambda(k)[u(k) - s(k)] \}$$

which is a two-dimensional optimization problem. Selection of $\boldsymbol{\lambda} \in E^N$ decomposes the problem into separate problems for each time period. The variable $\lambda(k)$ can be regarded as a value, measured in units of power, for water at the beginning of period k. The kth subproblem can then be interpreted as that faced by an entrepreneur who leased the dam for one period. He can buy water for the dam at the beginning of the period at price $\lambda(k)$ and sell what he has left at the end of the period at price $\lambda(k+1)$. His problem is to determine $y(k)$ and $u(k)$ so that his net profit, accruing from sale of generated power and purchase and sale of water, is maximized.

Example 5 (The Hanging Chain). Consider again the problem of finding the equilibrium position of the hanging chain considered in Example 4, Sect. 11.3, and Example 1, Sect. 12.7. The problem is

$$\begin{aligned} \text{minimize} \quad & \sum_{i=1}^{n} c_i y_i \\ \text{subject to} \quad & \sum_{i=1}^{n} y_i = 0 \\ & \sum_{i=1}^{n} \sqrt{1 - y_i^2} = L, \end{aligned}$$

where $c_i = n - i + \frac{1}{2}$, $L = 16$. This problem is locally convex, since as shown in Sect. 12.7 the Hessian of the Lagrangian is positive definite. The dual function is accordingly

$$\phi(\lambda, \mu) = \min \sum_{i=1}^{n} \left\{ c_i y_i + \lambda y_i + \mu \sqrt{1 - y_i^2} \right\} - L\mu.$$

Since the problem is separable, the minimization divides into a separate minimization for each y_i, yielding the equations

$$c_i + \lambda - \frac{\mu y_i}{\sqrt{1 - y_i^2}} = 0$$

or

$$(c_i + \lambda)^2 (1 - y_i^2) = \boldsymbol{\mu}^2 y_i^2.$$

This yields

$$y_i = \frac{-(c_i + \lambda)}{[(c_i + \lambda)^2 + \mu^2]^{1/2}}. \tag{14.30}$$

The above represents a local minimum point provided $\mu < 0$; and the minus sign must be taken for consistency.

The dual function is then

$$\phi(\lambda, \mu) = \sum_{i=1}^{n} \left\{ \frac{-(c_i + \lambda)^2}{[(c_i + \lambda)^2 + \mu^2]^{1/2}} + \mu \left[\frac{\mu^2}{[(c_i + \lambda)^2 + \mu^2]} \right]^{1/2} \right\} - L\mu$$

or finally, using $\sqrt{\mu^2} = -\mu$ for $\mu < 0$,

$$\phi(\lambda, \mu) = -L\mu - \sum_{i=1}^{n} \sqrt{(c_i + \lambda)^2 + \mu^2}.$$

The correct values of λ and μ can be found by maximizing $\phi(\lambda, \mu)$. One way to do this is to use steepest ascent. The results of this calculation, starting at $\lambda = \mu = 0$, are shown in Table 14.1. The values of y_i can then be found from (14.30).

Table 14.1 Results of dual of chain problem

Iteration	Value	Final solution $\lambda = -10.00048$ $\mu = -6.761136$
0	−200.00000	$y_1 = -0.8147154$
1	−66.94638	$y_2 = -0.7825940$
2	−66.61959	$y_3 = -0.7427243$
3	−66.55867	$y_4 = -0.6930215$
4	−66.54845	$y_5 = -0.6310140$
5	−66.54683	$y_6 = -0.5540263$
6	−66.54658	$y_7 = -0.4596696$
7	−66.54654	$y_8 = -0.3467526$
8	−66.54653	$y_9 = -0.2165239$
9	−66.54653	$y_{10} = -0.0736802$

14.5 Augmented Lagrangian

One of the most effective general classes of nonlinear programming methods is the *augmented Lagrangian* methods, alternatively referred to as *methods of multiplier*. These methods can be viewed as a combination of penalty functions and local duality methods; the two concepts work together to eliminate many of the disadvantages associated with either method alone. The augmented Lagrangian for the equality constrained problem

$$\begin{aligned} &\text{minimize} \quad f(\mathbf{x}) \\ &\text{subject to} \quad \mathbf{h}(\mathbf{x}) = \mathbf{0}, \quad \mathbf{x} \in \Omega \end{aligned} \tag{14.31}$$

is the function

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

for some positive constant c. We shall briefly indicate how the augmented Lagrangian can be viewed as either a special penalty function or as the basis for a dual problem. These two viewpoints are then explored further in this and the next section.

From a penalty function viewpoint the augmented Lagrangian, for a fixed value of the vector $\boldsymbol{\lambda}$, is simply the standard quadratic penalty function for the problem

$$\begin{aligned} &\text{minimize} \quad f(\mathbf{x}) + \boldsymbol{\lambda}^T \mathbf{h}(\mathbf{x}) \\ &\text{subject to} \quad \mathbf{h}(\mathbf{x}) = \mathbf{0}, \quad \mathbf{x} \in \Omega \end{aligned} \tag{14.32}$$

This problem is clearly equivalent to the original problem (14.31), since combinations of the constraints adjoined to $f(\mathbf{x})$ do not affect the minimum point or the minimum value.

A typical step of an augmented Lagrangian method starts with a vector $\boldsymbol{\lambda}_k$. Then $\mathbf{x}(\boldsymbol{\lambda}_k)$ is found as the minimum point of

$$\text{minimize } f(\mathbf{x}) + \boldsymbol{\lambda}_k^T \mathbf{h}(\mathbf{x}) + \frac{1}{2} c |\mathbf{h}(\mathbf{x})|^2 \quad \text{subject to} \quad \mathbf{x} \in \Omega. \tag{14.33}$$

Next $\boldsymbol{\lambda}_k$ is updated to $\boldsymbol{\lambda}_{k+1}$. A standard method for the update is

$$\boldsymbol{\lambda}_{k+1} = \boldsymbol{\lambda}_k + c\mathbf{h}(\mathbf{x}(\boldsymbol{\lambda}_k)).$$

To motivate the adjustment procedure, consider $\Omega = E^n$ and the constrained problem (14.32) with $\boldsymbol{\lambda} = \boldsymbol{\lambda}_k$. The Lagrange multiplier corresponding to this problem is $\boldsymbol{\lambda}^* - \boldsymbol{\lambda}_k$, where $\boldsymbol{\lambda}^*$ is the Lagrange multiplier of (14.31). On the other hand since (14.33) is the penalty function corresponding to (14.32), it follows from the results of Sect. 13.3 that $c\mathbf{h}(\mathbf{x}(\boldsymbol{\lambda}_k))$ is approximately equal to the Lagrange multiplier of (14.32). Combining these two facts, we obtain $c\mathbf{h}(\mathbf{x}(\boldsymbol{\lambda}_k)) \simeq \boldsymbol{\lambda}^* - \boldsymbol{\lambda}_k$. Therefore, a good approximation to the unknown $\boldsymbol{\lambda}^*$ is $\boldsymbol{\lambda}_{k+1} = \boldsymbol{\lambda}_k + c\mathbf{h}(\mathbf{x}(\boldsymbol{\lambda}_k))$.

Although the main iteration in augmented Lagrangian methods is with respect to $\boldsymbol{\lambda}$, the penalty parameter c may also be adjusted during the process. As in ordinary penalty function methods, the sequence of c's is usually preselected; c is either held fixed, is increased toward a finite value, or tends (slowly) toward infinity. Since in this method it is not necessary for c to go to infinity, and in fact it may remain of relatively modest value, the ill-conditioning usually associated with the penalty function approach is mediated.

From the viewpoint of duality theory, the augmented Lagrangian is simply the standard Lagrangian for the problem

$$\begin{aligned} &\text{minimize} && f(\mathbf{x}) + \frac{1}{2}c|\mathbf{h}(\mathbf{x})|^2 \\ &\text{subject to} && \mathbf{h}(\mathbf{x}) = \mathbf{0}, \quad \mathbf{x} \in \Omega. \end{aligned} \tag{14.34}$$

This problem is equivalent to the original problem (14.31), since the addition of the term $\frac{1}{2}c|\mathbf{h}(\mathbf{x})|^2$ to the objective does not change the optimal value, the optimum solution point, nor the Lagrange multiplier. However, whereas the original Lagrangian may not be convex near the solution, and hence the standard duality method cannot be applied, the term $\frac{1}{2}c|\mathbf{h}(\mathbf{x})|^2$ tends to "convexify" the Lagrangian. For sufficiently large c, the Lagrangian will indeed be locally convex. Thus the duality method can be employed, and the corresponding dual problem can be solved by an iterative process in $\boldsymbol{\lambda}$. This viewpoint leads to the development of additional multiplier adjustment processes.

The Penalty Viewpoint

We begin our more detailed analysis of augmented Lagrangian methods by showing that if the penalty parameter c is sufficiently large, the augmented Lagrangian has a local minimum point near the true optimal point. This follows from the following simple lemma. (Again, we consider $\Omega = E^n$ for simplicity.)

Lemma. *Let* $\mathbf{A}$ *and* $\mathbf{B}$ *be* $n \times n$ *symmetric matrices. Suppose that* $\mathbf{B}$ *is positive semi-definite and that* $\mathbf{A}$ *is positive definite on the subspace* $\mathbf{Bx} = \mathbf{0}$*. Then there is a* c^* *such that for all* $c \geq c^*$ *the matrix* $\mathbf{A} + c\mathbf{B}$ *is positive definite.*

Proof. Suppose to the contrary that for every k there were an $\mathbf{x}_k$ with $|\mathbf{x}_k| = 1$ such that $\mathbf{x}_k^T(\mathbf{A} + k\mathbf{B})\mathbf{x}_k \leq 0$. The sequence $\{\mathbf{x}_k\}$ must have a convergent subsequence converging to a limit $\bar{\mathbf{x}}$. Now since $\mathbf{x}_k^T\mathbf{B}\mathbf{x}_k \geq 0$, it follows that $\bar{\mathbf{x}}^T\mathbf{B}\bar{\mathbf{x}} = 0$. It also follows that $\bar{\mathbf{x}}^T\mathbf{A}\bar{\mathbf{x}} \leq \mathbf{0}$. However, this contradicts the hypothesis of the lemma. ∎

This lemma applies directly to the Hessian of the augmented Lagrangian evaluated at the optimal solution pair $\mathbf{x}^*, \boldsymbol{\lambda}^*$. We assume as usual that the second-order sufficiency conditions for a constrained minimum hold at $\mathbf{x}^*, \boldsymbol{\lambda}^*$. The Hessian of the augmented Lagrangian evaluated at the optimal pair $\mathbf{x}^*, \boldsymbol{\lambda}^*$ is

$$\begin{aligned}\mathbf{L}_c(\mathbf{x}^*, \boldsymbol{\lambda}^*) &= \mathbf{F}(\mathbf{x}^*) + (\boldsymbol{\lambda}^*)^T\mathbf{H}(\mathbf{x}^*) + c\nabla\mathbf{h}(\mathbf{x}^*)^T\nabla\mathbf{h}(\mathbf{x}^*) \\ &= \mathbf{L}(\mathbf{x}^*) + c\nabla\mathbf{h}(\mathbf{x}^*)^T\nabla\mathbf{h}(\mathbf{x}^*).\end{aligned}$$

The first term, the Hessian of the normal Lagrangian, is positive definite on the subspace $\nabla\mathbf{h}(\mathbf{x}^*)\mathbf{x} = \mathbf{0}$. This corresponds to the matrix A in the lemma. The matrix $\nabla\mathbf{h}(\mathbf{x}^*)^T\nabla\mathbf{h}(\mathbf{x}^*)$ is positive semi-definite and corresponds to $\mathbf{B}$ in the lemma. It follows that there is a c^* such that for all $c > c^*$, $\mathbf{L}_c(\mathbf{x}^*, \boldsymbol{\lambda}^*)$ is positive definite. This leads directly to the first basic result concerning augmented Lagrangian.

Proposition 1. *Assume that the second-order sufficiency conditions for a local minimum are satisfied at $\mathbf{x}^*, \boldsymbol{\lambda}^*$. Then there is a c^* such that for all $c \geq c^*$, the augmented Lagrangian $l_c(\mathbf{x}, \boldsymbol{\lambda}^*)$ has a local minimum point at $\mathbf{x}^*$.*

By a continuity argument the result of the above proposition can be extended to a neighborhood around $\mathbf{x}^*, \boldsymbol{\lambda}^*$. That is, for any $\boldsymbol{\lambda}$ near $\boldsymbol{\lambda}^*$, the augmented Lagrangian has a unique local minimum point near $\mathbf{x}^*$. This correspondence defines a continuous function. If a value of $\boldsymbol{\lambda}$ can be found such that $\mathbf{h}(\mathbf{x}(\boldsymbol{\lambda})) = \mathbf{0}$, then that $\boldsymbol{\lambda}$ must in fact be $\boldsymbol{\lambda}^*$, since $\mathbf{x}(\boldsymbol{\lambda})$ satisfies the necessary conditions of the original problem. Therefore, the problem of determining the proper value of $\boldsymbol{\lambda}$ can be viewed as one of solving the equation $\mathbf{h}(\mathbf{x}(\boldsymbol{\lambda})) = \mathbf{0}$. For this purpose the iterative process

$$\boldsymbol{\lambda}_{k+1} = \boldsymbol{\lambda}_k + c\mathbf{h}(\mathbf{x}(\boldsymbol{\lambda}_k)),$$

is a method of successive approximation. This process will converge linearly in a neighborhood around $\boldsymbol{\lambda}^*$, although a rigorous proof is somewhat complex. We shall give more definite convergence results when we consider the duality viewpoint.

Example 1. Consider the simple quadratic problem studied in Sect. 13.8

$$\begin{aligned}&\text{minimize} \quad 2x^2 + 2xy + y^2 - 2y \\ &\text{subject to} \quad x = 0.\end{aligned}$$

The augmented Lagrangian for this problem is

$$l_c(x, y, \lambda) = 2x^2 + 2xy + y^2 - 2y + \lambda x + \frac{1}{2}cx^2.$$

The minimum of this can be found analytically to be $x = -(2 + \lambda)/(2 + c)$, $y = (4 + c + \lambda)/(2 + c)$. Since $h(x, y) = x$ in this example, it follows that the iterative process for λ_k is

$$\lambda_{k+1} = \lambda_k - \frac{c(2 + \lambda_k)}{2 + c}$$

or

$$\lambda_{k+1} = \left(\frac{2}{2 + c}\right)\lambda_k - \frac{2c}{2 + c}.$$

This converges to $\lambda = -2$ for any $c > 0$. The coefficient $2/(2 + c)$ governs the rate of convergence, and clearly, as c is increased the rate improves.

Geometric Interpretation

The augmented Lagrangian method can be interpreted geometrically in terms of the primal function in a manner analogous to that in Sects. 13.3 and 13.8 for the ordinary quadratic penalty function and the absolute-value penalty function. Consider again the primal function $\omega(\mathbf{y})$ defined as

$$\omega(\mathbf{y}) = \min\{f(\mathbf{x}) : \mathbf{h}(\mathbf{x}) = \mathbf{y}\},$$

where the minimum is understood to be taken locally near $\mathbf{x}^*$. We remind the reader that $\omega(\mathbf{0}) = f(\mathbf{x}^*)$ and that $\nabla\omega(\mathbf{0})^T = -\boldsymbol{\lambda}^*$. The minimum of the augmented Lagrangian at step k can be expressed in terms of the primal function as follows:

$$\begin{aligned} \min l_c(\mathbf{x}, \boldsymbol{\lambda}_k) &= \min_{\mathbf{x}}\{f(\mathbf{x}) + \boldsymbol{\lambda}_k^T\mathbf{h}(\mathbf{x}) + \frac{1}{2}c|\mathbf{h}(\mathbf{x})|^2\} \\ &= \min_{\mathbf{x},\mathbf{u}}\{f(\mathbf{x}) + \boldsymbol{\lambda}_k^T\mathbf{y} + \frac{1}{2}c|\mathbf{y}|^2 : \mathbf{h}(\mathbf{x}) = \mathbf{y}\} \\ &= \min_{\mathbf{u}}\{\omega(\mathbf{y}) + \boldsymbol{\lambda}_k^T\mathbf{y} + \frac{1}{2}c|\mathbf{y}|^2\}, \end{aligned} \tag{14.35}$$

where the minimization with respect to $\mathbf{y}$ is to be taken locally near $\mathbf{y} = \mathbf{0}$. This minimization is illustrated geometrically for the case of a single constraint in Fig. 14.5. The lower curve represents $\omega(\mathbf{y})$, and the upper curve represents $\omega(\mathbf{y}) + \frac{1}{2}c|\mathbf{y}|^2$. The minimum point $\mathbf{y}_k$ of (14.30) occurs at the point where this upper curve has slope equal to $-\lambda_k$. It is seen that for c sufficiently large this curve will be convex at $y = 0$. If λ_k is close to λ^*, it is clear that this minimum point will be close to 0; it will be exact if $\lambda_k = \lambda^*$.

The process for updating $\boldsymbol{\lambda}_k$ is also illustrated in Fig. 14.5. Note that in general, if $\mathbf{x}(\boldsymbol{\lambda}_k)$ minimizes $l_c(\mathbf{x}, \boldsymbol{\lambda}_k)$, then $\mathbf{y}_k = \mathbf{h}(\mathbf{x}(\boldsymbol{\lambda}_k))$ is the minimum point of $\omega(\mathbf{y}) + \boldsymbol{\lambda}_k^T\mathbf{y} + \frac{1}{2}c|\mathbf{y}|^2$. At that point we have as before

$$\nabla\omega(\mathbf{y}_k)^T + c\mathbf{y}_k = -\boldsymbol{\lambda}_k$$

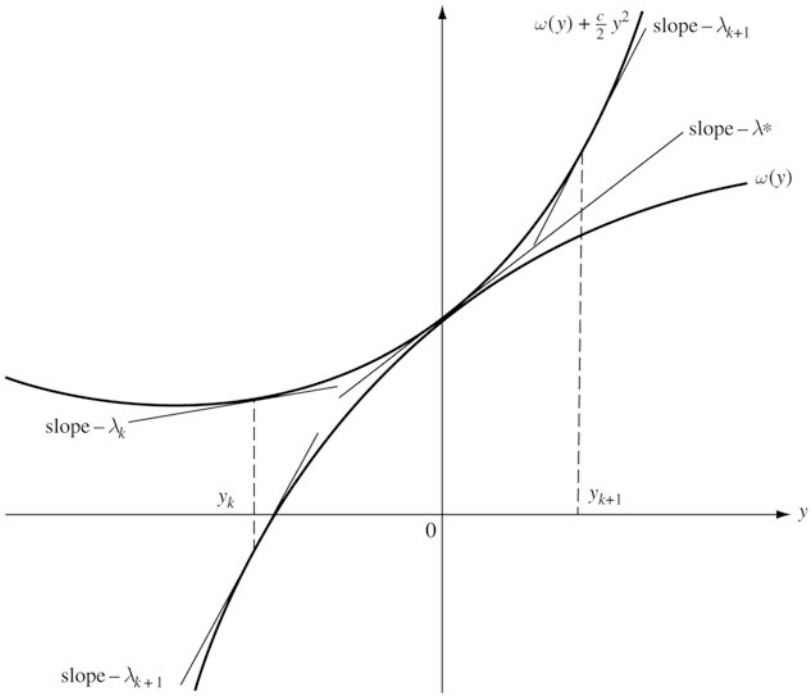

Fig. 14.5 Primal function and augmented Lagrangian

or equivalently,

$$\nabla\omega(\mathbf{y}_k)^T = -(\boldsymbol{\lambda}_k + c\mathbf{y}_k) = -(\boldsymbol{\lambda}_k + c\mathbf{h}(\mathbf{x}(\boldsymbol{\lambda}_k))).$$

It follows that for the next multiplier we have

$$\boldsymbol{\lambda}_{k+1} = \boldsymbol{\lambda}_k + c\mathbf{h}(\mathbf{x}(\boldsymbol{\lambda}_k)) = -\nabla\omega(\mathbf{y}_k)^T,$$

as shown in Fig. 14.5 for the one-dimensional case. In the figure the next point y_{k+1} is the point where $\omega(y) + \frac{1}{2}c|y|^2$ has slope $-\lambda_{k+1}$, which will yield a positive value of y_{k+1} in this case. It can be seen that if $\boldsymbol{\lambda}_k$ is sufficiently close to $\boldsymbol{\lambda}^*$, then $\boldsymbol{\lambda}_{k+1}$ will be even closer, and the iterative process will converge.

14.6 The Method of Multipliers

In the augmented Lagrangian method (the method of multipliers), the primary iteration is with respect to $\boldsymbol{\lambda}$, and therefore it is most natural to consider the method from the dual viewpoint. This is in fact the more powerful viewpoint and leads to improvements in the algorithm.

As we observed earlier, the constrained problem

$$\begin{aligned} &\text{minimize} \quad f(\mathbf{x}) \\ &\text{subject to} \quad \mathbf{h}(\mathbf{x}) = \mathbf{0}, \quad \mathbf{x} \in \Omega \end{aligned} \tag{14.36}$$

is *equivalent* to the problem

$$\begin{aligned} &\text{minimize} \quad f(\mathbf{x}) + \frac{1}{2}c|\mathbf{h}(\mathbf{x})|^2 \\ &\text{subject to} \quad \mathbf{h}(\mathbf{x}) = \mathbf{0}, \quad \mathbf{x} \in \Omega \end{aligned} \tag{14.37}$$

in the sense that the solution points, the optimal values, and the Lagrange multipliers are the same for both problems. However, as spelled out by Proposition 1 of the previous section, whereas problem (14.36) may not be locally convex, problem (14.37) is locally convex for sufficiently large c; specifically, the Hessian of the Lagrangian is positive definite at the solution pair $\mathbf{x}^*$, $\boldsymbol{\lambda}^*$. Thus local duality theory is applicable to problem (14.37) for sufficiently large c.

To apply the dual method to (14.37), we define the dual function

$$\phi(\boldsymbol{\lambda}) = \min\{f(\mathbf{x}) + \boldsymbol{\lambda}^T\mathbf{h}(\mathbf{x}) + \frac{1}{2}c|\mathbf{h}(\mathbf{x})|^2\} \tag{14.38}$$

in a region near $\mathbf{x}^*$, $\boldsymbol{\lambda}^*$. If $\mathbf{x}(\boldsymbol{\lambda})$ is the vector minimizing the right-hand side of (14.38), then as we have seen in Sect. 14.2, $\mathbf{h}(\mathbf{x}(\boldsymbol{\lambda}))$ is the gradient of ϕ. Thus the iterative process

$$\boldsymbol{\lambda}_{k+1} = \boldsymbol{\lambda}_k + c\mathbf{h}(\mathbf{x}(\boldsymbol{\lambda}_k))$$

used in the basic augmented Lagrangian method is seen to be *a steepest ascent iteration for maximizing the dual function* ϕ. It is a simple form of steepest ascent, using a constant stepsize c.

Although the stepsize c is a good choice (as will become even more evident later), it is clearly advantageous to apply the algorithmic principles of optimization developed previously by selecting the stepsize so that the new value of the dual function satisfies an ascent criterion. This can extend the range of convergence of the algorithm.

The rate of convergence of the optimal steepest ascent method (where the stepsize is selected to maximize ϕ in the gradient direction) is determined by the eigenvalues of the Hessian of ϕ. The Hessian of ϕ is found from (14.15) to be

$$\nabla\mathbf{h}(\mathbf{x}(\boldsymbol{\lambda}))[\mathbf{L}(\mathbf{x}(\boldsymbol{\lambda}),\ \boldsymbol{\lambda}) + c\nabla\mathbf{h}(\mathbf{x}(\boldsymbol{\lambda}))^T\nabla\mathbf{h}(\mathbf{x}(\boldsymbol{\lambda}))]^{-1}\nabla\mathbf{h}(\mathbf{x})^T. \tag{14.39}$$

The eigenvalues of this matrix at the solution point $\mathbf{x}^*$, $\boldsymbol{\lambda}^*$ determine the convergence rate of the method of steepest ascent.

To analyze the eigenvalues we make use of the matrix identity

$$c\mathbf{B}(\mathbf{A} + c\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T = \mathbf{I} - (\mathbf{I} + c\mathbf{B}\mathbf{A}^{-1}\mathbf{B}^T)^{-1},$$

which is a generalization of the Sherman-Morrison formula. (See Sect. 10.4.) It is easily seen from the above identity that the matrices $\mathbf{B}(\mathbf{A} + c\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T$ and $(\mathbf{B}\mathbf{A}^{-1}\mathbf{B}^T)$ have identical eigenvectors. One way to see this is to multiply both sides of the identity by $(\mathbf{I} + c\mathbf{B}\mathbf{A}^{-1}\mathbf{B}^T)$ on the right to obtain

$$c\mathbf{B}(\mathbf{A} + c\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T(\mathbf{I} + c\mathbf{B}\mathbf{A}^{-1}\mathbf{B}^T) = c\mathbf{B}\mathbf{A}^{-1}\mathbf{B}^T.$$

Suppose both sides are applied to an eigenvector $\mathbf{e}$ of $\mathbf{B}\mathbf{A}^{-1}\mathbf{B}^T$ having eigenvalue w. Then we obtain

$$c\mathbf{B}(\mathbf{A} + c\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T(1 + cw)\mathbf{e} = cw\mathbf{e}.$$

It follows that $\mathbf{e}$ is also an eigenvector of $\mathbf{B}(\mathbf{A} + c\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T$, and if v is the corresponding eigenvalue, the relation

$$cu(1 + cw) = cw$$

must hold. Therefore, the eigenvalues are related by

$$u = \frac{w}{1 + cw}. \tag{14.40}$$

The above relations apply directly to the Hessian (14.39) through the associations $\mathbf{A} = \mathbf{L}(\mathbf{x}^*, \boldsymbol{\lambda}^*)$ and $\mathbf{B} = \nabla\mathbf{h}(\mathbf{x}^*)$. Note that the matrix $\nabla\mathbf{h}(\mathbf{x}^*)\mathbf{L}(\mathbf{x}^*, \boldsymbol{\lambda}^*)^{-1}\nabla\mathbf{h}(\mathbf{x}^*)^T$, corresponding to $\mathbf{B}\mathbf{A}^{-1}\mathbf{B}^T$ above, is the Hessian of the dual function of the original problem (14.36). As shown in Sect. 14.3 the eigenvalues of this matrix determine the rate of convergence for the ordinary dual method. Let w and W be the smallest and largest eigenvalues of this matrix. From (14.40) it follows that the ratio of smallest to largest eigenvalues of the Hessian of the dual for the augmented problem is

$$\frac{\frac{1}{w} + c}{\frac{1}{w} + c}.$$

This shows explicitly how the rate of convergence of the multiplier method depends on c. As c goes to infinity, the ratio of eigenvalues goes to unity, implying arbitrarily fast convergence.

Other unconstrained optimization techniques may be applied to the maximization of the dual function defined by the augmented Lagrangian; conjugate gradient methods, Newton's method, and quasi-Newton methods can all be used. The use of Newton's method requires evaluation of the Hessian matrix (14.39). For some problems this may be feasible, but for others some sort of approximation is desirable. One approximation is obtained by noting that for large values of c, the Hessian (14.39) is approximately equal to $(1/c)\mathbf{I}$. Using this value for the Hessian and $\mathbf{h}(\mathbf{x}(\boldsymbol{\lambda}))$ for the gradient, we are led to the iterative scheme

$$\boldsymbol{\lambda}_{k+1} = \boldsymbol{\lambda}_k + c\mathbf{h}(\mathbf{x}(\boldsymbol{\lambda}_k)),$$

which is exactly the simple method of multipliers originally proposed.

We might summarize the above observations by the following statement relating primal and dual convergence rates. If a penalty term is incorporated into a problem, the condition number of the primal problem becomes increasingly poor as $c \to \infty$ but the condition number of the dual becomes increasingly good. To apply the dual method, however, an unconstrained penalty problem of poor condition number must be solved at each step.

Inequality Constraints

The advantage of augmented Lagrangian methods is mostly in dealing with equalities. But certain inequality constraints can be easily incorporated. Let us consider the problem with p inequality constraints:

$$\begin{aligned} &\text{minimize} \quad f(\mathbf{x}) \\ &\text{subject to} \quad \mathbf{g}(\mathbf{x}) \leq \mathbf{0}. \end{aligned} \tag{14.41}$$

We assume that this problem has a well-defined solution $\mathbf{x}^*$, which is a regular point of the constraints and which satisfies the second-order sufficiency conditions for a local minimum as specified in Sect. 11.8. This problem can be written as an equivalent problem with equality constraints:

$$\begin{aligned} &\text{minimize} \quad f(\mathbf{x}) \\ &\text{subject to} \quad \mathbf{g}(\mathbf{x}) + \mathbf{u} = \mathbf{0}, \ \mathbf{u} \geq \mathbf{0}. \end{aligned} \tag{14.42}$$

Through this conversion we can hope to simply apply the theory for equality constraints to problems with inequalities.

In order to do so we must insure that (14.42) satisfies the second-order sufficiency conditions of Sect. 11.5. These conditions will not hold unless we impose a *strict complementarity* assumption that $g_j(\mathbf{x}^*) = 0$ implies $\mu_j^* > 0$ as well as the usual second-order sufficiency conditions for the original problem (14.41). (See Exercise 10.)

With these assumptions we define the (partial) dual function corresponding to the augmented Lagrangian method as

$$\phi(\boldsymbol{\mu}) = \min_{\mathbf{u} \geq \mathbf{0},\ \mathbf{x}} \quad f(\mathbf{x}) + \boldsymbol{\mu}^T[\mathbf{g}(\mathbf{x}) + \mathbf{u}] + \frac{1}{2}c|\mathbf{g}(\mathbf{x}) + \mathbf{u}|^2. \tag{14.43}$$

The minimization with respect to $\mathbf{u}$ in (14.43) can be carried out analytically, and this will lead to a definition of the dual function that only involves minimization with respect to $\mathbf{x}$. The variable u_j enters the objective of the dual function only through the univariate quadratic expression

$$P_j = \mu_j[g_j(\mathbf{x}) + u_j] + \frac{1}{2}c[g_j(\mathbf{x}) + u_j]^2. \tag{14.44}$$

It is this expression that we must minimize with respect to $u_j \geq 0$. This is easily accomplished by differentiation: If $u_j > 0$, the derivative must vanish; if $u_j = 0$, the derivative must be nonnegative. The derivative is zero at $z_j = -g_j(\mathbf{x}) - \boldsymbol{\mu}_j/c$. Thus we obtain the solution

$$u_j = \begin{cases} -g_j(\mathbf{x}) - \frac{\mu_j}{c}, & \text{if } -g_j(\mathbf{x}) - \frac{\mu_j}{c} \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

or equivalently,

$$u_j = \max\left\{0,\ -g_j(\mathbf{x}) - \frac{\mu_j}{c}\right\}. \tag{14.45}$$

We now substitute this into (14.44) in order to obtain an explicit expression for the minimum of P_j.

For $u_j = 0$, we have

$$\begin{aligned} P_j &= \frac{1}{2c}\left(2\mu_j c g_j(\mathbf{x}) + c^2 g_j(\mathbf{x})^2\right) \\ &= \frac{1}{2c}\left([\mu_j + c g_j(\mathbf{x})]^2 - \mu_j^2\right). \end{aligned}$$

For $u_j = -g_j(\mathbf{x}) - \mu_j/c$ we have

$$P_j = -\mu_j^2/2c.$$

These can be combined into the formula

$$P_j = \frac{1}{2c}\left([\max\{0,\ \mu_j + c g_j(\mathbf{x})\}]^2 - \mu_j^2\right).$$

In view of the above, let us define the function of two scalar arguments t and $\boldsymbol{\mu}$:

$$P_c(t,\ \mu) = \frac{1}{2c}\left([\max\{0,\ \boldsymbol{\mu} + ct\}]^2 - \mu^2\right). \tag{14.46}$$

For a fixed $\mu > 0$, this function is shown in Fig. 14.6. Note that it is a smooth function with derivative with respect to t equal to μ at $t = 0$.

The dual function for the inequality problem can now be written as

$$\phi(\mu) = \min_{\mathbf{x}}\left(f(\mathbf{x}) + \sum_{j=1}^{p} P_c(g_j(\mathbf{x}),\ \boldsymbol{\mu}_j)\right). \tag{14.47}$$

Thus inequality problems can be treated by adjoining to $f(\mathbf{x})$ a special penalty function (that depends on $\boldsymbol{\mu}$). The Lagrange multiplier $\boldsymbol{\mu}$ can then be adjusted to maximize ϕ, just as in the case of equality constraints.

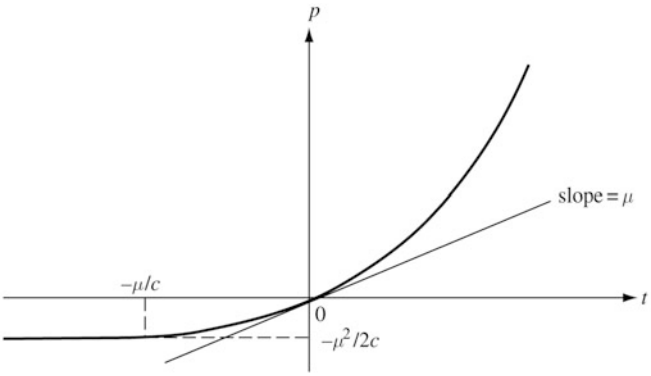

Fig. 14.6 Penalty function for inequality problem

14.7 The Alternating Direction Method of Multipliers

Consider the convex minimization model with linear constraints and an objective function which is the sum of two separable functions:

$$\begin{array}{ll} \text{minimize} & f_1(\mathbf{x}^1) + f_2(\mathbf{x}^2) \\ \text{subject to} & A_1\mathbf{x}^1 + A_2\mathbf{x}^2 = \mathbf{b}, \\ & \mathbf{x}^1 \in \Omega_1,\ \mathbf{x}^2 \in \Omega_2, \end{array} \tag{14.48}$$

where $A_i \in E^{m\times n_i}$ $(i = 1, 2)$, $\mathbf{b} \in E^m$, $\Omega_i \subset E^{n_i}$ $(i = 1, 2)$ are closed convex sets; and $f_i : E^{n_i} \to E$ $(i = 1, 2)$ are convex functions on Ω_i, respectively. Then, the augmented Lagrangian function for (14.48) would be

$$l_c(\mathbf{x}^1, \mathbf{x}^2, \boldsymbol{\lambda}) = f_1(\mathbf{x}^1) + f_2(\mathbf{x}^2) + \boldsymbol{\lambda}^T(A_1\mathbf{x}^1 + A_2\mathbf{x}^2 - \mathbf{b}) + \frac{c}{2}|A_1\mathbf{x}^1 + A_2\mathbf{x}^2 - \mathbf{b}|^2.$$

Throughout this section, we assume problem (14.48) has at least one optimal solution.

In contrast to the method of multipliers in the last section, the alternating direction method of multipliers (ADMM) is to (approximately) minimize $l_c(\mathbf{x}^1, \mathbf{x}^2, \boldsymbol{\lambda})$ in an alternative order:

$$\begin{aligned} \mathbf{x}^1_{k+1} &:= \arg\min_{\mathbf{x}^1\in\Omega_1} l_c(\mathbf{x}^1, \mathbf{x}^2_k, \boldsymbol{\lambda}_k), \\ \mathbf{x}^2_{k+1} &:= \arg\min_{\mathbf{x}^2\in\Omega_2} l_c(\mathbf{x}^1_{k+1}, \mathbf{x}^2, \boldsymbol{\lambda}_k), \\ \boldsymbol{\lambda}_{k+1} &:= \boldsymbol{\lambda}_k + c(A_1\mathbf{x}^1_{k+1} + A_2\mathbf{x}^2_{k+1} - \mathbf{b}). \end{aligned} \tag{14.49}$$

The idea is that each of the smaller minimization problems can be solved more efficiently or even in close forms for certain cases.

Convergence Speed Analysis

We present a convergence speed analysis of the ADMM. For simplicity, we shall let Ω_i be E^{n_i} and f_i be differentiable convex functions [the result is also valid for the ADMM applied to the aforementioned more general problem (14.48)]. Then, any optimal solution and multiplier $(\mathbf{x}_*^1, \mathbf{x}_*^2, \boldsymbol{\lambda}_*)$ satisfy

$$\nabla f_1(\mathbf{x}_*^1)^T + A_1^T \boldsymbol{\lambda}_* = \mathbf{0}, \quad \nabla f_2(\mathbf{x}_*^2)^T + A_2^T \boldsymbol{\lambda}_* = \mathbf{0}, \quad A_1\mathbf{x}_*^1 + A_2\mathbf{x}_*^2 - \mathbf{b} = \mathbf{0}, \tag{14.50}$$

and these conditions are also sufficient.

We first establish a key lemma.

Lemma 1. *Let* $\mathbf{d}_k^i = A_i(\mathbf{x}_k^i - \mathbf{x}_*^i)$, $i = 1, 2$, *and* $\mathbf{d}_k^\lambda = \boldsymbol{\lambda}_k - \boldsymbol{\lambda}_*$; *and* $\{\mathbf{x}_k^1, \mathbf{x}_k^2, \boldsymbol{\lambda}_k\}$ *be the sequence generated by ADMM* (14.49). *Then, it holds that*

$$c\left|A_2(\mathbf{x}_{k+1}^2 - \mathbf{x}_k^2)\right|^2 + \frac{1}{c}\left|\boldsymbol{\lambda}_{k+1} - \boldsymbol{\lambda}_k\right|^2 \le \left(c\left|A_2\mathbf{d}_k^2\right|^2 + \frac{1}{c}\left|\mathbf{d}_k^\lambda\right|^2\right) - \left(c\left|A_2\mathbf{d}_{k+1}^2\right|^2 + \frac{1}{c}\left|\mathbf{d}_{k+1}^\lambda\right|^2\right).$$

Proof. From the first-order optimality conditions of (14.49), we have

$$\begin{cases} \nabla f_1(\mathbf{x}_{k+1}^1)^T + A_1^T[\boldsymbol{\lambda}_k + c(A_1\mathbf{x}_{k+1}^1 + A_2\mathbf{x}_k^2 - \mathbf{b})] = \mathbf{0}, \\ \nabla f_2(\mathbf{x}_{k+1}^2)^T + A_2^T[\boldsymbol{\lambda}_k + c(A_1\mathbf{x}_{k+1}^1 + A_2\mathbf{x}_{k+1}^2 - \mathbf{b})] = \mathbf{0}, \\ \boldsymbol{\lambda}_{k+1} = \boldsymbol{\lambda}_k + c(A_1\mathbf{x}_{k+1}^1 + A_2\mathbf{x}_{k+1}^2 - \mathbf{b}). \end{cases} \tag{14.51}$$

Substituting the last equation into other equations in (14.51), we obtain

$$\begin{cases} \nabla f_1(\mathbf{x}_{k+1}^1)^T + A_1^T\boldsymbol{\lambda}_{k+1} = -cA_1^TA_2(\mathbf{x}_k^2 - \mathbf{x}_{k+1}^2), \\ \nabla f_2(\mathbf{x}_{k+1}^2)^T + A_2^T\boldsymbol{\lambda}_{k+1} = \mathbf{0}, \\ A_1\mathbf{x}_{k+1}^1 + A_2\mathbf{x}_{k+1}^2 - \mathbf{b} = \frac{1}{c}(\boldsymbol{\lambda}_{k+1} - \boldsymbol{\lambda}_k). \end{cases} \tag{14.52}$$

Moreover, the convexity of f_i, $i = 1, 2$, implies

$$(\nabla f_1(\mathbf{x}_{k+1}^1) - \nabla f_1(\mathbf{x}_*^1))(\mathbf{x}_{k+1}^1 - \mathbf{x}_*^1) \ge 0 \text{ and } (\nabla f_2(\mathbf{x}_{k+1}^2) - \nabla f_2(\mathbf{x}_*^2))(\mathbf{x}_{k+1}^2 - \mathbf{x}_*^2) \ge 0.$$

On the other hand, from (14.50) and (14.52),

$$\nabla f_1(\mathbf{x}_{k+1}^1)^T - \nabla f_1(\mathbf{x}_*^1)^T = \nabla f_1(\mathbf{x}_{k+1}^1) + A_1^T\boldsymbol{\lambda}_* = -A_1^T\mathbf{d}_{k+1}^\lambda - cA_1^TA_2(\mathbf{x}_k^2 - \mathbf{x}_{k+1}^2)$$

$$\nabla f_2(\mathbf{x}_{k+1}^2)^T - \nabla f_2(\mathbf{x}_*^2)^T = \nabla f_2(\mathbf{x}_{k+1}^2) + A_2^T\boldsymbol{\lambda}_* = -A_2^T\mathbf{d}_{k+1}^\lambda$$

and

$$\mathbf{0} = A_1\mathbf{x}_{k+1}^1 + A_2\mathbf{x}_{k+1}^2 - \mathbf{b} - \frac{1}{c}(\boldsymbol{\lambda}_{k+1} - \boldsymbol{\lambda}_k) = A_1\mathbf{d}_{k+1}^1 + A_2\mathbf{d}_2^{k+1} + \frac{1}{c}(\boldsymbol{\lambda}_k - \boldsymbol{\lambda}_{k+1}).$$

Thus,

$$
\begin{aligned}
0 \le & \begin{pmatrix} \mathbf{d}_{k+1}^1 \\ \mathbf{d}_2^{k+1} \\ \mathbf{d}_{k+1}^\lambda \end{pmatrix}^T \begin{pmatrix} \nabla f_1(\mathbf{x}_{k+1}^1)^T - \nabla f_1(\mathbf{x}_*^1)^T \\ \nabla f_2(\mathbf{x}_{k+1}^2)^T - \nabla f_2(\mathbf{x}_*^2)^T \\ \mathbf{0} \end{pmatrix} \\
= & \begin{pmatrix} \mathbf{d}_{k+1}^1 \\ \mathbf{d}_2^{k+1} \\ \mathbf{d}_{k+1}^\lambda \end{pmatrix}^T \begin{pmatrix} -A_1^T \mathbf{d}_{k+1}^\lambda - cA_1^T A_2(\mathbf{x}_k^2 - \mathbf{x}_{k+1}^2) \\ -A_2^T \mathbf{d}_{k+1}^\lambda \\ A_1 \mathbf{d}_{k+1}^1 + A_2 \mathbf{d}_2^{k+1} + \frac{1}{c}(\boldsymbol{\lambda}_k - \boldsymbol{\lambda}_{k+1}) \end{pmatrix} \\
= & \begin{pmatrix} \mathbf{d}_{k+1}^1 \\ \mathbf{d}_2^{k+1} \\ \mathbf{d}_{k+1}^\lambda \end{pmatrix}^T \left(\begin{bmatrix} -A_1^T \mathbf{d}_{k+1}^\lambda \\ -A_2^T \mathbf{d}_{k+1}^\lambda \\ A_1 \mathbf{d}_{k+1}^1 + A_2 \mathbf{d}_2^{k+1} \end{bmatrix} + \begin{bmatrix} -cA_1^T A_2(\mathbf{x}_k^2 - \mathbf{x}_{k+1}^2) \\ \mathbf{0} \\ \frac{1}{c}(\boldsymbol{\lambda}_k - \boldsymbol{\lambda}_{k+1}) \end{bmatrix} \right) \\
= & \begin{pmatrix} \mathbf{d}_{k+1}^1 \\ \mathbf{d}_2^{k+1} \\ \mathbf{d}_{k+1}^\lambda \end{pmatrix}^T \begin{pmatrix} -cA_1^T A_2(\mathbf{x}_k^2 - \mathbf{x}_{k+1}^2) \\ \mathbf{0} \\ \frac{1}{c}(\boldsymbol{\lambda}_k - \boldsymbol{\lambda}_{k+1}) \end{pmatrix} = \begin{pmatrix} -A_1 \mathbf{d}_{k+1}^1 \\ \mathbf{d}_{k+1}^\lambda \end{pmatrix}^T \begin{pmatrix} cA_2(\mathbf{x}_k^2 - \mathbf{x}_{k+1}^2) \\ \frac{1}{c}(\boldsymbol{\lambda}_k - \boldsymbol{\lambda}_{k+1}) \end{pmatrix}
\end{aligned} \tag{14.53}
$$

Again from $-A_1 \mathbf{d}_{k+1}^1 = \frac{1}{c}(\boldsymbol{\lambda}_{k+1} - \boldsymbol{\lambda}_k) + A_2 \mathbf{d}_{k+1}^2$, inequality (14.53) implies

$$
\begin{aligned}
0 \le & \begin{pmatrix} \frac{1}{c}(\boldsymbol{\lambda}_{k+1} - \boldsymbol{\lambda}_k) + A_2 \mathbf{d}_{k+1}^2 \\ \mathbf{d}_{k+1}^\lambda \end{pmatrix}^T \begin{pmatrix} cA_2(\mathbf{x}_k^2 - \mathbf{x}_{k+1}^2) \\ \frac{1}{c}(\boldsymbol{\lambda}_k - \boldsymbol{\lambda}_{k+1}) \end{pmatrix} \\
= & \begin{pmatrix} A_2 \mathbf{d}_{k+1}^2 \\ \mathbf{d}_{k+1}^\lambda \end{pmatrix}^T \begin{pmatrix} cA_2(\mathbf{x}_k^2 - \mathbf{x}_{k+1}^2) \\ \frac{1}{c}(\boldsymbol{\lambda}_k - \boldsymbol{\lambda}_{k+1}) \end{pmatrix} + (\boldsymbol{\lambda}_k - \boldsymbol{\lambda}_{k+1})^T A_2(\mathbf{x}_k^2 - \mathbf{x}_{k+1}^2)
\end{aligned}
$$

Since $\nabla f_2(\mathbf{x}_k^2) = -\boldsymbol{\lambda}_k^T A_2$ holds for every $k \ge 0$, it follows from the convexity of f_2 that

$$
(\boldsymbol{\lambda}_k - \boldsymbol{\lambda}_{k+1})^T A_2(\mathbf{x}_k^2 - \mathbf{x}_{k+1}^2) = -(\nabla f_2(\mathbf{x}_k^2) - \nabla f_2(\mathbf{x}_{k+1}^2))(\mathbf{x}_k^2 - \mathbf{x}_{k+1}^2) \le 0.
$$

Thus,

$$
\begin{pmatrix} A_2 \mathbf{d}_{k+1}^2 \\ \mathbf{d}_{k+1}^\lambda \end{pmatrix}^T \begin{pmatrix} cA_2(\mathbf{x}_k^2 - \mathbf{x}_{k+1}^2) \\ \frac{1}{c}(\boldsymbol{\lambda}_k - \boldsymbol{\lambda}_{k+1}) \end{pmatrix} \ge 0 \quad \text{or} \quad \begin{pmatrix} \sqrt{c} A_2 \mathbf{d}_{k+1}^2 \\ \frac{1}{\sqrt{c}} \mathbf{d}_{k+1}^\lambda \end{pmatrix}^T \begin{pmatrix} \sqrt{c} A_2(\mathbf{x}_{k+1}^2 - \mathbf{x}_k^2) \\ \frac{1}{\sqrt{c}}(\boldsymbol{\lambda}_{k+1} - \boldsymbol{\lambda}_k) \end{pmatrix} \le 0.
$$

Representing the left vector by $\mathbf{u}$ and the right one by $\mathbf{v}$ in the last inequality, we have

$$
0 \ge \mathbf{u}^T \mathbf{v} = \frac{1}{2}(|\mathbf{u}|^2 + |\mathbf{v}|^2 - |\mathbf{u} - \mathbf{v}|^2).
$$

Noting

$$
\mathbf{u} - \mathbf{v} = \begin{pmatrix} \sqrt{c} A_2 \mathbf{d}_{k+1}^2 \\ \frac{1}{\sqrt{c}} \mathbf{d}_{k+1}^\lambda \end{pmatrix} - \begin{pmatrix} \sqrt{c} A_2(\mathbf{x}_{k+1}^2 - \mathbf{x}_k^2) \\ \frac{1}{\sqrt{c}}(\boldsymbol{\lambda}_{k+1} - \boldsymbol{\lambda}_k) \end{pmatrix} = \begin{pmatrix} \sqrt{c} A_2 \mathbf{d}_k^2 \\ \frac{1}{\sqrt{c}} \mathbf{d}_k^\lambda \end{pmatrix},
$$

we obtain the desired result in Lemma 1. ∎

For simplicity, let $c = 1$ in the following. Taking the sum from iterate 0 to iterate k for the inequality in Lemma 1, we obtain

$$
\sum_{t=0}^{k} \left(\left| A_2(\mathbf{x}_{t+1}^2 - \mathbf{x}_t^2) \right|^2 + |\boldsymbol{\lambda}_{t+1} - \boldsymbol{\lambda}_t|^2 \right) \le \left| A_2 \mathbf{x}_0^2 - A_2 \mathbf{x}_*^2 \right|^2 + |\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}_*|^2 .
$$

Thus, we have

$$\min_{0\le t\le k}\left\{\left|A_2(\mathbf{x}_{t+1}^2-\mathbf{x}_t^2)\right|^2+|\boldsymbol{\lambda}_{t+1}-\boldsymbol{\lambda}_t|^2\right\}\le\frac{1}{k}\left(\left|A_2(\mathbf{x}_0^2-\mathbf{x}_*^2)\right|^2+|\boldsymbol{\lambda}_0-\boldsymbol{\lambda}_*|^2\right).$$

Therefore, from (14.52) we have

Theorem 1. *After k iterations of the ADMM method, there must be at least one iterate $0\le\bar{k}\le k$ such that*

$$\left\|\begin{pmatrix}\nabla f_1(\mathbf{x}_{\bar{k}+1}^1)^T+A_1^T\boldsymbol{\lambda}_{\bar{k}+1}\\ \nabla f_2(\mathbf{x}_{\bar{k}+1}^2)^T+A_2^T\boldsymbol{\lambda}_{\bar{k}+1}\\ A_1\mathbf{x}_{\bar{k}+1}^1+A_2\mathbf{x}_{\bar{k}+1}^2-\mathbf{b}\end{pmatrix}\right\|^2\le\frac{1+|A_1|}{k}\left(\left|A_2(\mathbf{x}_0^2-\mathbf{x}_*^2)\right|^2+|\boldsymbol{\lambda}_0-\boldsymbol{\lambda}_*|^2\right),$$

that is, $(\mathbf{x}_{\bar{k}+1}^1,\ \mathbf{x}_{\bar{k}+1}^2,\ \boldsymbol{\lambda}_{\bar{k}+1})$ has its optimality condition error square bounded by the quantity on the right-hand side that converges 0 arithmetically as $k\to\infty$.

The Three Block Extension

It is natural to consider the ADMM method for solving problems with more than two blocks:

$$\begin{aligned}&\text{minimize}\quad f_1(\mathbf{x}^1)+f_2(\mathbf{x}^2)+f_3(\mathbf{x}^3)\\ &\text{subject to}\quad A_1\mathbf{x}^1+A_2\mathbf{x}^2+A_3\mathbf{x}^3=\mathbf{b},\\ &\qquad\qquad \mathbf{x}^1\in\Omega_1,\ \mathbf{x}^2\in\Omega_2,\ \mathbf{x}^3\in\Omega_3,\end{aligned}\tag{14.54}$$

where $A_i\in E^{m\times n_i}$ $(i=1,2,3)$, $\mathbf{b}\in E^m$, $\Omega_i\subset E^{n_i}$ $(i=1,2,3)$ are closed convex sets; and $f_i:\ E^{n_i}\to E$ $(i=1,2,3)$ are convex functions on Ω_i, respectively. With the same philosophy as the ADMM to take advantage of the separable structure, one could consider the procedure

$$\begin{aligned}\mathbf{x}_{k+1}^1&:=\arg\min_{\mathbf{x}^1\in\Omega_1}l_c(\mathbf{x}^1,\mathbf{x}_k^2,\mathbf{x}_k^3,\boldsymbol{\lambda}_k),\\ \mathbf{x}_{k+1}^2&:=\arg\min_{\mathbf{x}^2\in\Omega_2}l_c(\mathbf{x}_{k+1}^1,\mathbf{x}^2,\mathbf{x}_k^3,\boldsymbol{\lambda}_k),\\ \mathbf{x}_{k+1}^3&:=\arg\min_{\mathbf{x}^3\in\Omega_3}l_c(\mathbf{x}_{k+1}^1,\mathbf{x}_{k+1}^2,\mathbf{x}^3,\boldsymbol{\lambda}_k),\\ \boldsymbol{\lambda}_{k+1}&:=\boldsymbol{\lambda}_k+c(A_1\mathbf{x}_{k+1}^1+A_2\mathbf{x}_{k+1}^2+A_3\mathbf{x}_{k+1}^3-\mathbf{b}),\end{aligned}\tag{14.55}$$

where the augmented Lagrangian function

$$l_c(\mathbf{x}^1,\mathbf{x}^2,\mathbf{x}^3,\boldsymbol{\lambda})=\sum_{i=1}^3 f_i(\mathbf{x}^i)+\boldsymbol{\lambda}^T(\sum_{i=1}^3 A_i\mathbf{x}^i-\mathbf{b})+\frac{c}{2}\Big|\sum_{i=1}^3 A_i\mathbf{x}^i-\mathbf{b}\Big|^2.$$

Unfortunately, unlike the convergence property for solving two-block problems, such a direct extension of ADMM not converge for problems with three blocks. Indeed, consider the following linear homogeneous equation with three variables

$$(A_1, A_2, A_3)\begin{pmatrix} x^1 \\ x^2 \\ x^3 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 2 \\ 1 & 2 & 2 \end{pmatrix}\begin{pmatrix} x^1 \\ x^2 \\ x^3 \end{pmatrix} = \mathbf{0}.$$

Let $c = 1$ and each block contain one variable. Then, simple calculation will show that the direct extension of ADMM (14.55) is divergent from any point in a subspace of E^3. Note that the convergence of ADMM (14.55) applied to solving the linear equations with a null objective is independent of the selection of the penalty parameter c. We conclude:

Theorem 2. *For the three-block convex minimization problem (14.54), the direct extension of ADMM (14.55) may not converge for any penalty parameter $c > 0$ starting from any point in a subspace.*

*14.8 *Cutting Plane Methods

Cutting plane methods are applied to problems having the general form

$$\begin{aligned} &\text{minimize} \quad \mathbf{c}^T\mathbf{x} \\ &\text{subject to} \quad \mathbf{x} \in S, \end{aligned} \tag{14.56}$$

where $S \subset E^n$ is a closed convex set. Problems that involve minimization of a convex function over a convex set, such as the problem

$$\begin{aligned} &\text{minimize} \quad f(\mathbf{y}) \\ &\text{subject to} \quad \mathbf{y} \in R, \end{aligned} \tag{14.57}$$

where $R \subset E^{n-1}$ is a convex set and f is a convex function, can be easily converted to the form (14.56) by writing (14.57) equivalently as

$$\begin{aligned} &\text{minimize} \quad r \\ &\text{subject to} \quad f(\mathbf{y}) - r \leqslant 0, \ \mathbf{y} \in R \end{aligned} \tag{14.58}$$

which, with $\mathbf{x} = (r, \mathbf{y}) \in E^n$, is a special case of (14.56).

General Form of Algorithm

The general form of a cutting-plane algorithm for problem (14.56) is as follows: Given a polytope $P_k \supset S$

Step 1. Minimize $\mathbf{c}^T\mathbf{x}$ over P_k obtaining a point $\mathbf{x}_k$ in P_k. If $\mathbf{x}_k \in S$, stop; $\mathbf{x}_k$ is optimal. Otherwise,

Step 2. Find a hyperplane H_k separating the point $\mathbf{x}_k$ from S, that is, find $\mathbf{a}_k \in E^n$, $b_k \in E^1$ such that $S \subset \{\mathbf{x} : \mathbf{a}_k^T \mathbf{x} \leqslant b_k\}$, $\mathbf{x}_k \in \{\mathbf{x} : \mathbf{a}_k^T \mathbf{x} > b_k\}$. Update P_k to obtain P_{k+1} including as a constraint $\mathbf{a}_k^T \mathbf{x} \leqslant b_k$.

The process is illustrated in Fig. 14.7.

Specific algorithms differ mainly in the manner in which the hyperplane that separates the current point $\mathbf{x}_k$ from the constraint set S is selected. This selection is, of course, the most important aspect of the algorithm, since it is the deepness of the cut associated with the separating hyperplane, the distance of the hyperplane from the current point, that governs how much improvement there is in the approximation to the constraint set, and hence how fast the method converges.

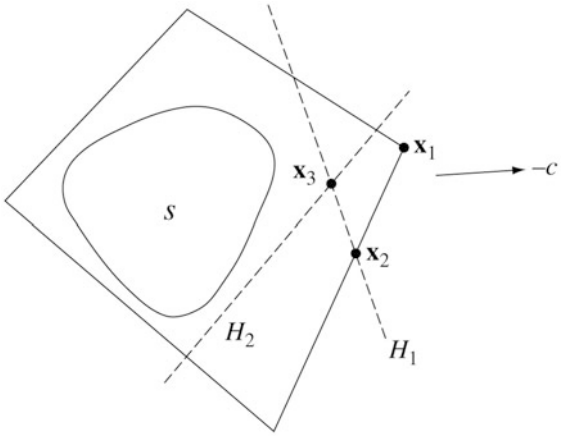

Fig. 14.7 Cutting plane method

Specific algorithms also differ somewhat with respect to the manner by which the polytope is updated once the new hyperplane is determined. The most straightforward procedure is to simply adjoin the linear inequality associated with that hyperplane to the ones determined previously. This yields the best possible updated approximation to the constraint set but tends to produce, after a large number of iterations, an unwieldy number of inequalities expressing the approximation. Thus, in some algorithms, older inequalities that are not binding at the current point are discarded from further consideration.

The general cutting plane algorithm can be regarded as an extended application of duality in linear programming, and although this viewpoint does not particularly aid in the analysis of the method, it reveals the basic interconnection between cutting plane and dual methods. The foundation of this viewpoint is the fact that S can be written as the intersection of all the half-spaces that contain it; thus

$$S = \{\mathbf{x} : \mathbf{a}_i^T \mathbf{x} \leq b_i,\ i \in I\},$$

where I is an (infinite) index set corresponding to all half-spaces containing S. With S viewed in this way problem (14.56) can be thought of as an (infinite) linear programming problem.

Corresponding to this linear program there is (at least formally) the dual problem

$$\begin{aligned} \text{maximize} \quad & \sum_{i \in I} \lambda_i b_i \\ \text{subject to} \quad & \sum_{i \in I} \lambda_i \mathbf{a}_i = \mathbf{c} \\ & \lambda_i \geqslant 0, \quad i \in I. \end{aligned} \tag{14.59}$$

Selecting a finite subset of I, say $\bar{I}$, and forming

$$P = \{\mathbf{x} : \mathbf{a}_i^T \mathbf{x} \leqslant b_i, \ i \in \bar{I}\}$$

gives a polytope that contains S. Minimizing $\mathbf{c}^T\mathbf{x}$ over this polytope yields a point and a corresponding subset of active constraints I_A. The dual problem with the additional restriction $\lambda_i = 0$ for $i \notin I_A$ will then have a feasible solution, but this solution will in general not be optimal. Thus, a solution to a polytope problem corresponds to a feasible but non-optimal solution to the dual. For this reason the cutting plane method can be regarded as working toward optimality of the (infinite dimensional) dual.

Kelley's Convex Cutting Plane Algorithm

The convex cutting plane method was developed to solve convex programming problems of the form

$$\begin{aligned} \text{minimize} \quad & f(\mathbf{x}) \\ \text{subject to} \quad & g_i(\mathbf{x}) \leqslant 0, \ i = 1, 2, \ \ldots, \ p, \end{aligned} \tag{14.60}$$

where $\mathbf{x} \in E^n$ and f and the g_i's are differentiable convex functions. As indicated in the last section, it is sufficient to consider the case where the objective function is linear; thus, we consider the problem

$$\begin{aligned} \text{minimize} \quad & \mathbf{c}^T\mathbf{x} \\ \text{subject to} \quad & \mathbf{g}(\mathbf{x}) \leqslant \mathbf{0} \end{aligned} \tag{14.61}$$

where $\mathbf{x} \in E^n$ and $\mathbf{g}(\mathbf{x}) \in E^p$ is convex and differentiable.

For $\mathbf{g}$ convex and differentiable we have the fundamental inequality

$$\mathbf{g}(\mathbf{x}) \geqslant \mathbf{g}(\mathbf{w}) + \nabla\mathbf{g}(\mathbf{w})(\mathbf{x} - \mathbf{w}) \tag{14.62}$$

for any $\mathbf{x}$, $\mathbf{w}$. We use this equation to determine the separating hyperplane. Specifically, the algorithm is as follows:

Let $S = \{\mathbf{x} : \mathbf{g}(\mathbf{x}) \leqslant \mathbf{0}\}$ and let P be an initial polytope containing S and such that $\mathbf{c}^T\mathbf{x}$ is bounded on P. Then

Step 1. Minimize $\mathbf{c}^T\mathbf{x}$ over P obtaining the point $\mathbf{x} = \mathbf{w}$. If $\mathbf{g}(\mathbf{w}) \leqslant \mathbf{0}$, stop; $\mathbf{w}$ is an optimal solution. Otherwise,

Step 2. Let i be an index maximizing $g_i(\mathbf{w})$. Clearly $g_i(\mathbf{w}) > 0$. Define the new approximating polytope to be the old one intersected with the half-space

$$\{\mathbf{x} : g_i(\mathbf{w}) + \nabla g_i(\mathbf{w})(\mathbf{x} - \mathbf{w}) \leqslant 0\}. \tag{14.63}$$

Return to Step 1.

The set defined by (14.63) is actually a half-space if $\nabla g_i(\mathbf{w}) \neq \mathbf{0}$. However, $\nabla g_i(\mathbf{w}) = \mathbf{0}$ would imply that $\mathbf{w}$ minimizes g_i which is impossible if S is nonempty. Furthermore, the half-space given by (14.63) contains S, since if $\mathbf{g}(\mathbf{x}) \leqslant \mathbf{0}$ then by (14.62) $g_i(\mathbf{w}) + \nabla g_i(\mathbf{w})(\mathbf{x} - \mathbf{w}) \leqslant g_i(\mathbf{x}) \leqslant \mathbf{0}$. The half-space does not contain the point $\mathbf{w}$ since $g_i(\mathbf{w}) > 0$. This method for selecting the separating hyperplane is illustrated in Fig. 14.8 for the one-dimensional case. Note that in one dimension, the procedure reduces to Newton's method.

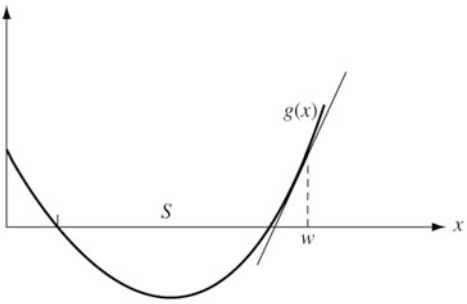

Fig. 14.8 Convex cutting plane

Calculation of the separating hyperplane is exceedingly simple in this algorithm, and hence the method really amounts to the solution of a series of linear programming problems. It should be noted that this algorithm, valid for any convex programming problem, does not involve any line searches. In that respect it is also similar to Newton's method applied to a convex function.

Convergence

Under fairly mild assumptions on the convex function, the convex cutting plane method is globally convergent. It is possible to apply the general convergence theorem to prove this, but somewhat easier, in this case, to prove it directly.

Theorem. *Let the convex functions g_i, $i = 1, 2, \ldots, p$ be continuously differentiable, and suppose the convex cutting plane algorithm generates the sequence of points $\{\mathbf{w}_k\}$. Any limit point of this sequence is a solution to problem (14.61).*

Proof. Suppose $\{\mathbf{w}_k\}$, $k \in \mathcal{K}$ is a subsequence of $\{\mathbf{w}_k\}$ converging to $\mathbf{w}$. By taking a further subsequence of this, if necessary, we may assume that the index i corresponding to Step 2 of the algorithm is fixed throughout the subsequence. Now if $k \in \mathcal{K}$, $k' \in \mathcal{K}$ and $k' > k$, then we must have

$$g_i(\mathbf{w}_k) + \nabla g_i(\mathbf{w}_k)(\mathbf{w}_{k'} - \mathbf{w}_k) \leqslant 0,$$

which implies that

$$g_i(\mathbf{w}_k) \leq |\nabla g_i(\mathbf{w}_k)||\mathbf{w}_{k'} - \mathbf{w}_k|. \tag{14.64}$$

Since $|\nabla g_i(\mathbf{w}_k)|$ is bounded with respect to $k \in \mathcal{K}$, the right-hand side of (14.64) goes to zero as k and k' go to infinity. The left-hand side goes to $g_i(\mathbf{w})$. Thus $g_i(\mathbf{w}) \leqslant 0$ and we see that $\mathbf{w}$ is feasible for problem (14.61).

If f^* is the optimal value of problem (14.61), we have $\mathbf{c}^T\mathbf{w}_k \leqslant f^*$ for each k since $\mathbf{w}_k$ is obtained by minimizing over a set containing S. Thus, by continuity, $\mathbf{c}^T\mathbf{w} \leqslant f^*$ and hence $\mathbf{w}$ is an optimal solution. ∎

As with most algorithms based on linear programming concepts, the rate of convergence of cutting plane algorithms has not yet been satisfactorily analyzed. Preliminary research shows that these algorithms converge arithmetically, that is, if $\mathbf{x}^*$ is optimal, then $|\mathbf{x}_k - \mathbf{x}^*|^2 \leqslant c/k$ for some constant c. This is an exceedingly poor type of convergence. This estimate, however, may not be the best possible and indeed there are indications that the convergence is actually geometric but with a ratio that goes to unity as the dimension of the problem increases.

Modifications

We now describe the supporting hyperplane algorithm (an alternative method for determining a cutting plane) and examine the possibility of dropping from consideration some old hyperplanes so that the linear programs do not grow too large. The convexity requirements are less severe for this algorithm. It is applicable to problems of the form

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

where $\mathbf{x} \in E^n$, $\mathbf{g}(\mathbf{x}) \in E^p$, the g_i's are continuously differentiable, and the constraint region S defined by the inequalities is convex. Note that convexity of the functions themselves is not required. We also assume the existence of a point interior to the constraint region, that is, we assume the existence of a point $\mathbf{y}$ such that $\mathbf{g}(\mathbf{y}) < \mathbf{0}$, and we assume that on the constraint boundary $g_i(\mathbf{x}) = 0$ implies $\nabla g_i(\mathbf{x}) \neq \mathbf{0}$. The algorithm is as follows:

Start with an initial polytope P containing S and such that $\mathbf{c}^T\mathbf{x}$ is bounded below on S. Then

Step 1. Determine $\mathbf{w} = \mathbf{x}$ to minimize $\mathbf{c}^T\mathbf{x}$ over P. If $\mathbf{w} \in S$, stop. Otherwise,

Step 2. Find the point $\mathbf{u}$ on the line joining $\mathbf{y}$ and $\mathbf{w}$ that lies on the boundary of S. Let i be an index for which $g_i(\mathbf{u}) = 0$ and define the half-space $H = \{\mathbf{x}: \nabla g_i(\mathbf{u})(\mathbf{x} - \mathbf{u}) \leqslant 0\}$. Update P by intersecting with H. Return to Step 1.

The algorithm is illustrated in Fig. 14.9.

The price paid for the generality of this method over the convex cutting plane method is that an interpolation along the line joining $\mathbf{y}$ and $\mathbf{w}$ must be executed to find the point $\mathbf{u}$. This is analogous to the line search for a minimum point required by most programming algorithms.

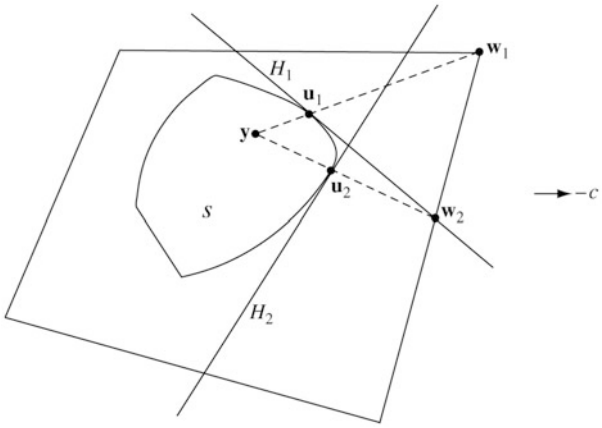

Fig. 14.9 Supporting hyperplane algorithm

Dropping Nonbinding Constraints

In all cutting plane algorithms nonbinding constraints can be dropped from the approximating set of linear inequalities so as to keep the complexity of the approximation manageable. Indeed, since n linearly independent hyperplanes determine a single point in E^n, the algorithm can be arranged, by discarding the nonbinding constraints at the end of each step, so that the polytope consists of exactly n linear inequalities at every stage.

Global convergence is not destroyed by this process, since the sequence of objective values will still be monotonically increasing. It is not known, however, what effect this has on the speed of convergence.

14.9 Exercises

1. (Linear programming) Use the global duality theorem to find the dual of the linear program

$$\begin{aligned} &\text{minimize} \quad \mathbf{c}^T\mathbf{x} \\ &\text{subject to} \quad \mathbf{A}\mathbf{x} = \mathbf{b},\ \mathbf{x} \geq \mathbf{0}. \end{aligned}$$

   Note that some of the regularity conditions may not be necessary for the linear case.
2. (Double dual) Show that the for a convex programming problem with a solution, the dual of the dual is in some sense the original problem.
3. (Non-convex?) Consider the problem

$$\begin{aligned} &\text{minimize} \quad xy \\ &\text{subject to} \quad x + y - 4 \geq 0 \\ &\qquad\qquad 1 \leq x \leq 5,\ 1 \leq y \leq 5. \end{aligned}$$

   Show that although the objective function is not convex, the primal function is convex. Find the optimal value and the Lagrange multiplier.
4. Find the global maximum of the dual function of Example 1, Sect. 14.2.
5. Show that the function ϕ defined for $\boldsymbol{\lambda},\ \boldsymbol{\mu},\ (\boldsymbol{\mu} \geqslant \mathbf{0})$, by $\phi(\boldsymbol{\lambda},\ \boldsymbol{\mu}) = \min_{\mathbf{x}}[f(\mathbf{x}) + \boldsymbol{\lambda}^T\mathbf{h}(\mathbf{x}) + \boldsymbol{\mu}^T\mathbf{g}(\mathbf{x})]$ is concave over any convex region where it is finite.
6. Prove that the dual canonical rate of convergence is not affected by a change of variables in $\mathbf{x}$.
7. Corresponding to the dual function (14.23):
   (a) Find its gradient.
   (b) Find its Hessian.
   (c) Verify that it has a local maximum at $\boldsymbol{\lambda}^*,\ \boldsymbol{\mu}^*$.
8. Find the Hessian of the dual function for a separable problem.
9. Find an explicit formula for the dual function for the entropy problem (Example 3, Sect. 11.4).
10. Consider the problems

$$\begin{aligned} &\text{minimize} \quad f(\mathbf{x}) \\ &\text{subject to} \quad g_i(\mathbf{x}) \leqslant 0,\ j = 1, 2, \ldots, p \end{aligned} \tag{14.65}$$

    and

$$\begin{aligned} &\text{minimize} \quad f(\mathbf{x}) \\ &\text{subject to} \quad g_i(\mathbf{x}) + z_j^2 = 0,\ j = 1, 2, \ldots, p. \end{aligned} \tag{14.66}$$

    (a) Let $\mathbf{x}^*,\ \mu_1^*,\ \mu_2^*,\ \ldots,\ \mu_p^*$ be a point and set of Lagrange multipliers that satisfy the first-order necessary conditions for (14.65). For $\mathbf{x}^*,\ \boldsymbol{\mu}^*$, write the second-order sufficiency conditions for (14.66).

(b) Show that in general they are not satisfied unless, in addition to satisfying the sufficiency conditions of Sect. 11.8, $g_j(\mathbf{x}^*)$ implies $\mu_j^* > 0$.

11. Establish global convergence for the supporting hyperplane algorithm.
12. Establish global convergence for an imperfect version of the supporting hyperplane algorithm that in interpolating to find the boundary point $\mathbf{u}$ actually finds a point somewhere on the segment joining $\mathbf{u}$ and $\frac{1}{2}\mathbf{u} + \frac{1}{2}\mathbf{w}$ and establishes a hyperplane there.
13. Prove that the convex cutting plane method is still globally convergent if it is modified by discarding from the definition of the polytope at each stage hyperplanes corresponding to inactive linear inequalities.

References

14.1 Global duality was developed in conjunction with the theory of Sect. 11.9, by Hurwicz [H14] and Slater [S7]. The theory was presented in this form in Luenberger [L8].

14.2–14.3 An important early differential form of duality was developed by Wolfe [W3]. The convex theory can be traced to the Legendre transformation used in the calculus of variations but it owes its main heritage to Fenchel [F3]. This line was further developed by Karlin [K1] and Hurwicz [H14]. Also see Luenberger [L8].

14.4 The solution of separable problems by dual methods in this manner was pioneered by Everett [E2].

14.5–14.6 The method of multipliers was originally suggested by Hestenes [H8] and from a different viewpoint by Powell [P7]. The relation to duality was presented briefly in Luenberger [L15]. The method for treating inequality constraints was devised by Rockafellar [R3]. For an excellent survey of multiplier methods see Bertsekas [B12].

14.7 The alternating direction method of multipliers was due to Gabay and Mercier[109] and Glowinski and Marrocco [102]; also see Fortin and Glowinski[96], Eckstein and Bertsekas [78] and Boyd et al. [41]. The convergence speed analysis was initially done by He and Yuan [124] and Monteiro and Svaiter [180]. The non-convergence examples of three blocks were constructed by Chen et al. [50].

14.8 Cutting plane methods were first introduced by Kelley [K3] who developed the convex cutting plane method. The supporting hyperplane algorithm was suggested by Veinott [V5]. To see how global convergence of cutting plane algorithms can be established from the general convergence theorem see Zangwill [Z2]. For some results on the convergence rates of cutting plane algorithms consult Topkis [T7], Eaves and Zangwill [E1], and Wolfe [W7].