5

Lagrange Multiplier Algorithms

Contents

In this chapter, we consider several computational methods for problems with equality and inequality constraints. All of these methods use some form of Lagrange multiplier estimates and typically provide in the limit not just stationary points of the original problem but also associated Lagrange multipliers. Additional methods using Lagrange multipliers will also be discussed in the next two chapters after the development of duality theory.

The methods of this chapter are based on one of the following two ideas:

(a) *Using a penalty or a barrier function.* Here a constrained problem is converted into a sequence of unconstrained problems, which involve an added high cost either for infeasibility or for approaching the boundary of the feasible region via its interior. These methods are discussed in Sections 5.1-5.3, and include interior point linear programming methods based on the logarithmic barrier function, augmented Lagrangian methods, and sequential quadratic programming.

(b) *Solving the necessary optimality conditions*, viewing them as a system of equations and/or inequalities in the problem variables and the associated Lagrange multipliers. These methods are first discussed in Section 5.1.2 in a specialized linear programming context, and later in Section 5.4. For nonlinear programming problems, they guarantee only local convergence in their pure form; that is, they converge only when a good solution estimate is initially available. However, their convergence region can be enlarged by using various schemes that involve penalty and barrier functions.

The methods based on these two ideas turn out to be quite interconnected, and as an indication of this, we note the derivations of optimality conditions using penalty and augmented Lagrangian techniques in Chapter 4. Generally, the methods of this chapter are particularly well-suited for nonlinear constraints, because, contrary to the feasible direction methods of Chapter 3, they do not involve projections or direction finding subproblems, which tend to become more difficult when the constraints are nonlinear. Still, however, some of the methods of this chapter are very well suited for linear and quadratic programming problems, thus illustrating the power of blending descent, penalty/barrier, and Lagrange multiplier ideas within a common algorithmic framework.

5.1 BARRIER AND INTERIOR POINT METHODS

Barrier methods apply to inequality constrained problems of the form

$$\begin{aligned} &\text{minimize} \;\; f(x) \\ &\text{subject to} \;\; x \in X, \qquad g_j(x) \le 0, \quad j = 1, \ldots, r, \end{aligned} \tag{5.1}$$

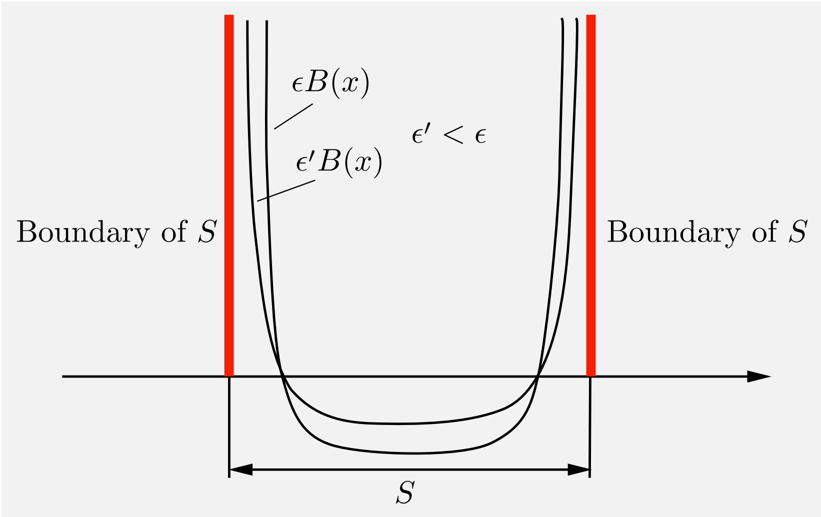

Figure 5.1.1. Form of a barrier function.

where X is a closed set, and $f : \Re^n \mapsto \Re$ and $g_j : \Re^n \mapsto \Re$ are given functions, with f continuous. The interior (relative to X) of the set defined by the inequality constraints is

$$S = \big\{x \in X \mid g_j(x) < 0,\, j = 1, \ldots, r\big\}.$$

We assume that S is nonempty and that any feasible point that is not in S can be approached arbitrarily closely by a vector from S; that is, given any feasible x and any $\delta > 0$, there exists $\tilde{x} \in S$ such that $\|\tilde{x} - x\| \leq \delta$. This property holds automatically if X and the constraint functions g_j are convex, as can be seen using the line segment principle [Prop. B.7(a) in Appendix B].

In barrier methods, we add to the cost a function $B(x)$ that is defined in the interior set S. This function, called the *barrier function*, is continuous and goes to ∞ as any one of the constraints $g_j(x)$ approaches 0 from negative values. The two most common examples of barrier functions are:

$$B(x) = -\sum_{j=1}^{r} \ln\big\{-g_j(x)\big\}, \qquad \text{logarithmic,}$$

$$B(x) = -\sum_{j=1}^{r} \frac{1}{g_j(x)}, \qquad \text{inverse.}$$

Note that both of these barrier functions are convex if all the constraint functions g_j are convex. Figure 5.1.1 illustrates the form of $B(x)$.

The barrier method is defined by introducing a parameter sequence $\{\epsilon^k\}$ with

$$0 < \epsilon^{k+1} < \epsilon^k, \quad k = 0, 1, \ldots, \qquad \epsilon^k \to 0.$$

It consists of finding

$$x^k \in \arg\min_{x\in S}\big\{f(x) + \epsilon^k B(x)\big\}, \qquad k = 0, 1, \ldots$$

Since the barrier function is defined only on the interior set S, the successive iterates of any method used for this minimization must be interior points. If $X = \Re^n$, one may use unconstrained methods such as Newton's method with the stepsize properly selected to ensure that all iterates lie in S; an initial interior point can be obtained as discussed in Section 3.2. Note that the barrier term $\epsilon^k B(x)$ goes to zero for all interior points $x \in S$ as $\epsilon^k \to 0$. Thus the barrier term becomes increasingly inconsequential as far as interior points are concerned, while progressively allowing x^k to get closer to the boundary of S (as it should if the solutions of the original constrained problem lie on the boundary of S). Figure 5.1.2 illustrates the convergence process, and the following proposition gives the main convergence result.

Proposition 5.1.1: Every limit point of a sequence $\{x^k\}$ generated by a barrier method is a global minimum of the original constrained problem (5.1).

Proof: Let $\{\bar{x}\}$ be the limit of a subsequence $\{x^k\}_{k\in K}$. If $\bar{x} \in S$, we have $\lim_{k\to\infty,\, k\in K} \epsilon^k B(x^k) = 0$, while if $\bar{x}$ lies on the boundary of S, we have by assumption $\lim_{k\to\infty,\, k\in K} B(x^k) = \infty$. In either case we obtain

$$\liminf_{k\to\infty} \epsilon^k B(x^k) \geq 0,$$

which implies that

$$\liminf_{k\to\infty,\, k\in K}\big\{f(x^k) + \epsilon^k B(x^k)\big\} = f(\bar{x}) + \liminf_{k\to\infty,\, k\in K}\big\{\epsilon^k B(x^k)\big\} \geq f(\bar{x}). \quad (5.2)$$

The vector $\bar{x}$ is a feasible point of the original problem (5.1), since $x^k \in S$ and X is a closed set. If $\bar{x}$ were not a global minimum, there would exist a feasible vector x^* such that $f(x^*) < f(\bar{x})$. Therefore, using the continuity of f and our assumption that x^* can be approached arbitrarily closely through the interior set S, there would also exist an interior point $\hat{x} \in S$ such that $f(\hat{x}) < f(\bar{x})$. We now have by the definition of x^k,

$$f(x^k) + \epsilon^k B(x^k) \leq f(\hat{x}) + \epsilon^k B(\hat{x}), \qquad k = 0, 1, \ldots,$$

which by taking the limit as $k \to \infty$ and $k \in K$, implies together with Eq. (5.2), that $f(\bar{x}) \leq f(\hat{x})$. This is a contradiction, thereby proving that $\bar{x}$ is a global minimum of the original problem. **Q.E.D.**

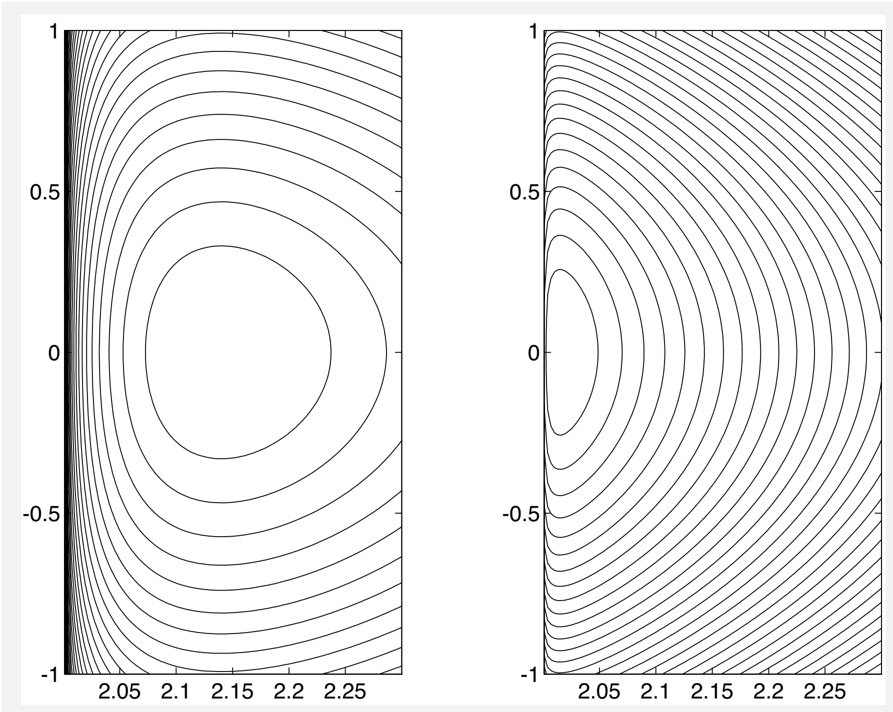

Figure 5.1.2. The convergence process of the barrier method for the problem

$$\begin{aligned} &\text{minimize} \quad f(x) = \tfrac{1}{2}\left(x_1^2 + x_2^2\right) \\ &\text{subject to} \quad 2 \le x_1, \end{aligned}$$

with optimal solution $x^* = (2, 0)$. For the case of the logarithmic barrier function $B(x) = -\ln(x_1 - 2)$, we have

$$x^k \in \arg\min_{x_1 > 2} \left\{ \tfrac{1}{2}\left(x_1^2 + x_2^2\right) - \epsilon^k \ln(x_1 - 2) \right\} = \left(1 + \sqrt{1 + \epsilon^k}\,, 0\right),$$

so as ϵ^k is decreased, the unconstrained minimum x^k approaches the constrained minimum $x^* = (2, 0)$. The figure shows the equal cost surfaces of

$$f(x) + \epsilon B(x)$$

for $\epsilon = 0.3$ (left side) and $\epsilon = 0.03$ (right side).

The logarithmic barrier function has been central to much research on methods that generate successive iterates lying in the interior set S. These methods are generically referred to as *interior point methods*, and have been extensively applied to linear and quadratic programming problems following the influential paper [Kar84]. We proceed to discuss the linear programming case in detail, using two different approaches based on the logarithmic barrier.

5.1.1 Path Following Methods for Linear Programming

In this section we consider the linear programming problem

$$\begin{aligned} &\text{minimize} \;\; c'x \\ &\text{subject to} \;\; Ax = b, \qquad x \geq 0, \end{aligned} \tag{LP}$$

where $c \in \Re^n$ and $b \in \Re^m$ are given vectors, and A is an $m \times n$ matrix of rank m. We will adapt the logarithmic barrier method to this problem. We assume that the problem has at least one optimal solution, and from the theory of Section 4.4.2, we have that the dual problem, given by

$$\begin{aligned} &\text{maximize} \;\; b'\lambda \\ &\text{subject to} \;\; A'\lambda \leq c, \end{aligned} \tag{DP}$$

also has an optimal solution. Furthermore, the optimal values of the primal and the dual problem are equal.

The method involves finding for various $\epsilon > 0$,

$$x(\epsilon) \in \arg\min_{x \in S} F_\epsilon(x), \tag{5.3}$$

where

$$F_\epsilon(x) = c'x - \epsilon \sum_{i=1}^{n} \ln x_i,$$

and S is the interior set

$$S = \{x \mid Ax = b,\, x > 0\},$$

where $x > 0$ means that all the coordinates of x are strictly positive. We assume that S is nonempty and bounded. Since $-\ln x_i$ grows to ∞ as $x_i \to 0$, this assumption can be used together with Weierstrass' theorem (Prop. A.8 in Appendix A) to show that there exists at least one global minimum of $F_\epsilon(x)$ over S, which must be unique because the function F_ϵ can be seen to be strictly convex. Therefore, for each $\epsilon > 0$, $x(\epsilon)$ is uniquely defined by Eq. (5.3).

The Central Path

For given A, b, and c, as ϵ is reduced towards 0, $x(\epsilon)$ follows a trajectory that is known as the *central path*. Figure 5.1.3 illustrates the central path for various values of the cost vector c. Note the following:

(a) For fixed A and b, the central paths corresponding to different cost vectors c start at the same vector x_∞. This is the unique minimizing point over S of

$$-\sum_{i=1}^{n} \ln x_i,$$

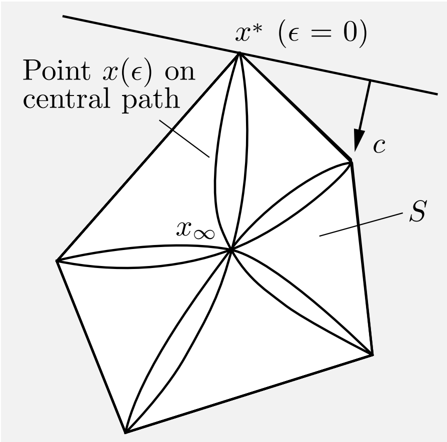

Figure 5.1.3. Central path trajectories $\big\{x(\epsilon) \mid 0 < \epsilon < \infty\big\}$ corresponding to ten different values of the cost vector c. All central paths start at the same vector, the *analytic center* x_∞, which corresponds to $\epsilon = \infty$,

$$x_\infty \in \arg\min_{x \in S} \left\{ -\sum_{i=1}^{n} \ln x_i \right\},$$

and end at optimal solutions of (LP).

corresponding to $\epsilon = \infty$, and is known as the *analytic center* of S.

(b) If c is such that (LP) has a unique optimal solution x^*, the central path ends at x^* [i.e., $\lim_{\epsilon \to 0} x(\epsilon) = x^*$]. This follows from Prop. 5.1.1, which implies that for every sequence $\{\epsilon^k\}$ with $\epsilon^k \to 0$, the corresponding sequence $\big\{x(\epsilon^k)\big\}$ converges to x^*.

(c) If c is such that (LP) has multiple optimal solutions, it can be shown that the central path ends at one of the optimal solutions [i.e., $\lim_{\epsilon \to 0} x(\epsilon)$ exists and is equal to some optimal solution of (LP)]. We will not prove this fact (see the end-of-chapter references).

Following Approximately the Central Path

The most straightforward way to implement the logarithmic barrier method is to use some iterative algorithm to minimize the function F_{ϵ^k} for each ϵ^k in a sequence $\{\epsilon^k\}$ with $\epsilon^k \downarrow 0$. This is equivalent to finding a sequence $\big\{x(\epsilon^k)\big\}$ of points on the central path. However, this approach is inefficient because it requires an infinite number of iterations to compute each point $x(\epsilon^k)$.

It turns out that a far more efficient approach is possible, whereby each minimization is done approximately through a few iterations (possibly only one) of the constrained version of Newton's method that was given in Section 3.3. For a fixed ϵ and a given $x \in S$, this method replaces x by

$$\tilde{x} = x + \alpha(\bar{x} - x),$$

where $\bar{x}$ is the pure Newton iterate defined as the optimal solution of the quadratic program in the vector z

$$\begin{aligned} &\text{minimize} \quad \nabla F_\epsilon(x)'(z - x) + \tfrac{1}{2}(z - x)'\nabla^2 F_\epsilon(x)(z - x) \\ &\text{subject to} \quad Az = b, \quad z \in \Re^n, \end{aligned}$$

and α is a stepsize selected by some rule. We have

$$\nabla F_\epsilon(x) = c - \epsilon x^{-1}, \qquad \nabla^2 F_\epsilon(x) = \epsilon X^{-2},$$

where x^{-1} denotes the vector with coordinates $(x_i)^{-1}$ and X denotes the diagonal matrix with the coordinates x_i along the diagonal:

$$X = \begin{pmatrix} x_1 & 0 & \cdots & 0 \\ 0 & x_2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & x_n \end{pmatrix}.$$

We can obtain an expression for $\bar{x}$ by using the formula for the projection on a linear manifold given in Example 3.1.5 of Section 3.1. We have that the pure Newton iterate is

$$\bar{x} = x - \epsilon^{-1} X^2 \left(c - \epsilon x^{-1} - A'\lambda\right),$$

where

$$\lambda = (AX^2A')^{-1} AX^2 \left(c - \epsilon x^{-1}\right).$$

These formulas can also be written as

$$\bar{x} = x - Xq(x, \epsilon), \tag{5.4}$$

where

$$q(x, \epsilon) = \frac{Xz}{\epsilon} - e, \tag{5.5}$$

with e and z being the vectors

$$e = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}, \qquad z = c - A'\lambda, \tag{5.6}$$

and

$$\lambda = (AX^2A')^{-1} AX\big(Xc - \epsilon e\big). \tag{5.7}$$

Based on Eq. (5.4), we have $q(x, \epsilon) = X^{-1}(\bar{x} - x)$, so we may view the vector $q(x, \epsilon)$ as a transformed version of the Newton increment $(x - \bar{x})$ using the transformation matrix X^{-1} that maps the vector x into the vector e. Since $\bar{x}$ is the Newton step approximation to $x(\epsilon)$, we can consider $\big\|q(x, \epsilon)\big\|$ as a *measure of proximity* of the current point x to the point $x(\epsilon)$ on the central path. In particular, it can be seen that we have $q(x, \epsilon) = 0$ if and only if $x = x(\epsilon)$.

The key result to be shown shortly is that for convergence of the logarithmic barrier method, it is sufficient to stop the minimization of F_{ϵ^k} and decrease ϵ^k to ϵ^{k+1} once the current iterate x^k satisfies

$$\big\|q(x^k, \epsilon^k)\big\| < 1.$$

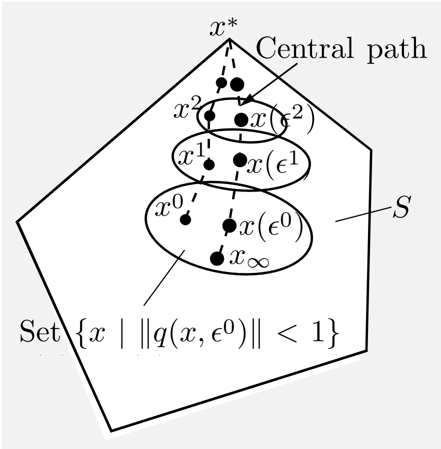

Figure 5.1.4. Following approximately the central path. For each ϵ^k, it is sufficient to carry out the minimization of F_{ϵ^k} up to where $\big\|q(x,\epsilon^k)\big\| < 1$.

Another way to phrase this result is that if a sequence of pairs $\big\{(x^k,\epsilon^k)\big\}$ satisfies

$$\big\|q(x^k,\epsilon^k)\big\| < 1, \qquad 0 < \epsilon^{k+1} < \epsilon^k, \quad k = 0,1,\ldots, \qquad \epsilon^k \to 0,$$

then every limit point of $\{x^k\}$ is an optimal solution of (LP); see Fig. 5.1.4. The following proposition establishes this result.

Proposition 5.1.2: If $x > 0$, $Ax = b$, and $\|q(x,\epsilon)\| < 1$, then

$$c'x - f^* \le c'x - b'\lambda \le \epsilon\big(n + \big\|q(x,\epsilon)\big\|\sqrt{n}\big) \le \epsilon\big(n + \sqrt{n}\big), \tag{5.8}$$

where λ is given by Eq. (5.7), and f^* is the optimal value of (LP), i.e.,

$$f^* = \min_{Ay=b,\, y\ge 0} c'y.$$

Proof: Using the definition (5.5)-(5.7) of q, we can write the hypothesis $\big\|q(x,\epsilon)\big\| < 1$ as

$$\left\|\frac{X(c - A'\lambda)}{\epsilon} - e\right\| < 1. \tag{5.9}$$

Thus the coordinates of $\big(X(c - A'\lambda)/\epsilon\big) - e$ must lie between -1 and 1, implying that the coordinates of $X(c - A'\lambda)$ are positive. Since the diagonal elements of X are positive, it follows that the coordinates of $c - A'\lambda$ are also positive. Hence $c \ge A'\lambda$, and for any optimal solution x^* of (LP), we obtain (using the fact $x^* \ge 0$)

$$f^* = c'x^* \ge \lambda' A x^* = \lambda' b. \tag{5.10}$$

On the other hand, since $||e|| = \sqrt{n}$, we have using Eq. (5.9),

$$e'\left(\frac{X(c - A'\lambda)}{\epsilon} - e\right) \le ||e|| \left\|\frac{X(c - A'\lambda)}{\epsilon} - e\right\| = \sqrt{n}\,\big\|q(x,\epsilon)\big\| \le \sqrt{n}, \tag{5.11}$$

and by using also Eq. (5.10),

$$e'\left(\frac{X(c - A'\lambda)}{\epsilon} - e\right) = \frac{x'(c - A'\lambda)}{\epsilon} - n = \frac{c'x - b'\lambda}{\epsilon} - n \ge \frac{c'x - f^*}{\epsilon} - n. \tag{5.12}$$

By combining Eqs. (5.11) and (5.12), the result follows. **Q.E.D.**

Note that from Eq. (5.8), $c'x - b'\lambda$ provides a readily computable upper bound to the (unknown) cost error $c'x - f^*$. What is happening here is that x and λ are feasible solutions to the primal and dual problems (LP) and (DP), respectively, and the common optimal value f^* lies between the corresponding primal and dual costs $c'x$ and $b'\lambda$.

Path-Following by Using Newton's Method

Since in order to implement the termination criterion $\big\|q(x,\epsilon)\big\| < 1$, we must calculate the pure Newton iterate $\bar{x} = x - Xq(x,\epsilon)$, it is natural to use a convergent version of Newton's method for approximate minimization of F_ϵ. This method replaces x by

$$\tilde{x} = x + \alpha(\bar{x} - x),$$

where α is a stepsize selected by the minimization rule or the Armijo rule (with unit initial stepsize) over the range of positive stepsizes such that $\tilde{x}$ is an interior point. We expect that for x sufficiently close to $x(\epsilon)$, the stepsize α can be taken equal to 1, so that the pure form of the method is used and a quadratic rate of convergence is obtained. The following proposition shows that the "termination set" $\big\{x \mid \|q(x,\epsilon)\| < 1\big\}$ is part of the region of quadratic convergence of the pure form of Newton's method.

Proposition 5.1.3: If $x > 0$, $Ax = b$, and $\big\|q(x,\epsilon)\big\| < 1$, then the pure Newton iterate $\bar{x} = x - Xq(x,\epsilon)$ is an interior point, i.e., $\bar{x} \in S$. Furthermore, we have $\big\|q(\bar{x},\epsilon)\big\| < 1$ and in fact

$$\big\|q(\bar{x},\epsilon)\big\| \le \big\|q(x,\epsilon)\big\|^2. \tag{5.13}$$

Proof: Let us define

$$p = Xz/\epsilon = X(c - A'\lambda)/\epsilon,$$

so that $q(x,\epsilon) = p-e$ [cf. Eqs. (5.4) and (5.5)]. Since $\|p-e\| < 1$, we see that the coordinates of p satisfy $0 < p_i < 2$ for all i. We have $\bar{x} = x - X(p-e)$, so that

$$\bar{x}_i = (2 - p_i)x_i > 0$$

for all i, and since also $A\bar{x} = b$, it follows that $\bar{x}$ is an interior point.

It can be shown (Exercise 5.1.3) that the vector $\bar{\lambda}$ corresponding to $\bar{x}$ in the manner of Eq. (5.7) satisfies

$$\bar{\lambda} \in \arg\min_{\xi \in \Re^m} \left\| \frac{\bar{X}(c - A'\xi)}{\epsilon} - e \right\|,$$

where $\bar{X}$ is the diagonal matrix with $\bar{x}_i$ along the diagonal. Hence,

$$\big\|q(\bar{x},\epsilon)\big\| = \left\| \frac{\bar{X}(c - A'\bar{\lambda})}{\epsilon} - e \right\| \le \left\| \frac{\bar{X}(c - A'\lambda)}{\epsilon} - e \right\| = \|\bar{X}X^{-1}p - e\|.$$

Since $\bar{x} = 2x - Xp$, we have

$$\bar{X}X^{-1}p = (2X - XP)X^{-1}p = 2p - Pp,$$

where P is the diagonal matrix with p_i along the diagonal. The last two relations yield

$$\begin{aligned} \big\|q(\bar{x},\epsilon)\big\|^2 &\le \|2p - Pp - e\|^2 \le \sum_{i=1}^n \big(2p_i - p_i^2 - 1\big)^2 = \sum_{i=1}^n (p_i - 1)^4 \\ &\le \left(\sum_{i=1}^n (p_i - 1)^2 \right)^2 = \|p - e\|^4 = \big\|q(x,\epsilon)\big\|^4. \end{aligned}$$

This proves the result. **Q.E.D.**

The preceding proposition shows that if $\big\|q(x,\epsilon)\big\|$ is substantially less than 1, then a single pure Newton step, changing x to $\bar{x}$, reduces $\big\|q(x,\epsilon)\big\|$ by a substantial factor [cf. Eq. (5.13)]. Thus, we expect that if $\bar{\epsilon}$ is not much smaller than ϵ and $\big\|q(x,\epsilon)\big\|$ is substantially less than 1, then $\big\|q(\bar{x},\bar{\epsilon})\big\|$ will also be substantially less than 1. This means that by carefully selecting the ϵ-reduction factor $\epsilon^{k+1}/\epsilon^k$ in combination with an appropriately small termination tolerance for the first minimization of F_ϵ ($k = 0$), we can execute all the subsequent approximate minimizations of F_{ϵ^k} ($k \ge 1$) in a single pure Newton step; see Fig. 5.1.5. One possibility is, given ϵ^k and x^k such that

$$\big\|q(x^k,\epsilon^k)\big\| \le 1,$$

to obtain x^{k+1} by a single Newton step and then to select ϵ^{k+1} so that $\big\|q(x^{k+1},\epsilon^{k+1})\big\|$ is minimized. This minimization can be done in closed

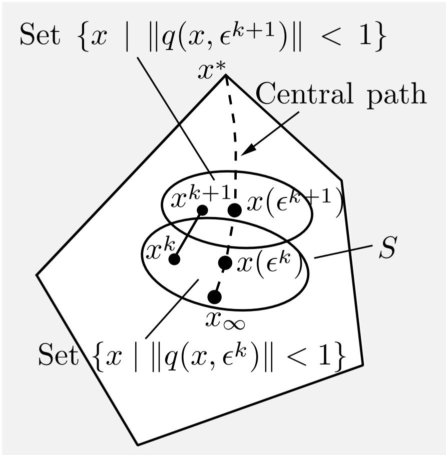

Figure 5.1.5. Following approximately the central path by using a single Newton step for each ϵ^k. If ϵ^k is close to ϵ^{k+1} and x^k is close to the central path, one expects that x^{k+1} obtained from x^k by a single pure Newton step will also be close to the central path.

form because $\big\|q(x,\epsilon)\big\|$ is quadratic in $1/\epsilon$ [cf. Eq. (5.5)]. Another possibility is shown in the next proposition.

Proposition 5.1.4: Suppose that $x > 0$, $Ax = b$, and that $\|q(x,\epsilon)\| \leq \gamma$ for some $\gamma < 1$. For any $\delta \in (0, n^{1/2})$, let $\bar\epsilon = (1 - \delta n^{-1/2})\epsilon$. Then

$$\big\|q(\bar x,\bar\epsilon)\big\| \leq \frac{\gamma^2 + \delta}{1 - \delta n^{-1/2}}.$$

In particular, if

$$\delta \leq \frac{\gamma(1-\gamma)}{1+\gamma}, \tag{5.14}$$

we have $\|q(\bar x,\bar\epsilon)\| \leq \gamma$.

Proof: Let $\theta = \delta n^{-1/2}$. We have using Eq. (5.5)

$$q(\bar x,\bar\epsilon) = \frac{\bar X z}{\bar\epsilon} - e = \frac{\bar X z}{(1-\theta)\epsilon} - e = \frac{q(\bar x,\epsilon)+e}{1-\theta} - e = \frac{1}{1-\theta}\big(q(\bar x,\epsilon) + \theta e\big).$$

Thus, using also Eq. (5.13),

$$\begin{aligned}
\big\|q(\bar x,\bar\epsilon)\big\| &\leq \frac{1}{1-\theta}\big(\big\|q(\bar x,\epsilon)\big\| + \theta\|e\|\big)\\
&= \frac{1}{1-\theta}\big(\big\|q(\bar x,\epsilon)\big\| + \theta n^{1/2}\big)\\
&\leq \frac{1}{1-\theta}\big(\big\|q(x,\epsilon)\big\|^2 + \delta\big)\\
&\leq \frac{\gamma^2+\delta}{1-\theta}.
\end{aligned}$$

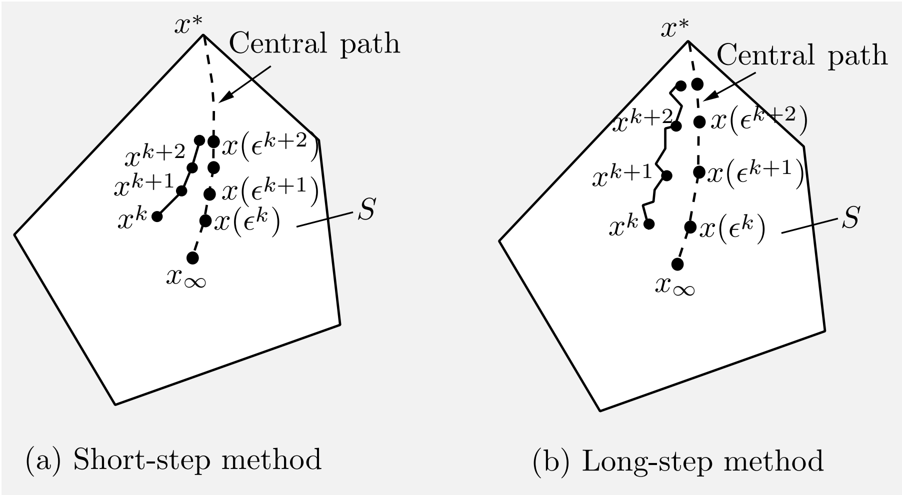

Figure 5.1.6. Following approximately the central path by decreasing ϵ^k slowly as in (a) or quickly as in (b). In (a) a single Newton step is required in each approximate minimization at the expense of a large number of approximate minimizations.

Finally, Eq. (5.14) can be written as $(\gamma^2+\delta)/(1-\delta) \le \gamma$, which, in combination with the relation just proved, implies that $\|q(\bar{x},\bar{\epsilon})\| \le \gamma$. **Q.E.D.**

Note that in the preceding proposition one can maintain x very close to the central path ($\gamma << 1$) provided one takes δ to be very small [cf. Eq. (5.14)], or equivalently, one uses an ϵ-reduction factor $1-\delta n^{-1/2}$ that is very close to 1. Unfortunately, even when γ is close to 1, in order to guarantee the single-step attainment of the tolerance $\|q(x,\epsilon)\| < \gamma$, it is still necessary to decrease ϵ very slowly. In particular, since we must take $\delta < 1$ in order for $\|q(\bar{x},\bar{\epsilon})\| < \gamma$ [cf. Eq. (5.14)], the reduction factor $\bar{\epsilon}/\epsilon$ must exceed $1-n^{-1/2}$, which is very close to 1. This means that, even though each approximate minimization after the first will require a single Newton step, a very large number of approximate minimizations will be needed to attain an acceptable accuracy. Thus, it may be more efficient in practice to decrease ϵ^k at a faster rate, while accepting the possibility of multiple Newton steps before switching from ϵ^k to ϵ^{k+1}, as illustrated in Fig. 5.1.6.

The preceding results form the basis for worst-case estimates of the number of Newton iterations required to reduce the error $c'x^k - f^*$ below some tolerance, where x^k is obtained by approximate minimization of F_{ϵ^k} using different termination criteria and reduction factors $\epsilon^{k+1}/\epsilon^k$. Exercise 5.1.5 provides a sample of this type of analysis. Many researchers consider a low estimate of number of iterations a good indicator of algorithmic performance. We note, however, that the worst-case estimates that have been obtained for interior point methods are so unrealistically high that they are entirely meaningless if taken literally.

One may hope that these estimates are meaningful in a comparative sense, i.e., the practical performance of two algorithms would compare consistently with the corresponding worst-case estimates of required numbers of iterations. Unfortunately, this does not turn out to be true in practice. In particular, we note that the lowest estimates of the required number of iterations have been obtained for the so-called *short-step methods*, where ϵ^k is reduced very slowly so that the corresponding approximate minimization can be done in a single Newton step (cf. Prop. 5.1.4). The best practical performance, however, has been obtained with the so-called *long-step methods*, where ϵ^k is reduced at a much faster rate. Thus one should view worst-case analyses of the required number of iterations of interior point methods with some skepticism; they may be primarily considered as an analytical vehicle for understanding better the corresponding methods.

Quadratic and Convex Programming

The logarithmic barrier method in conjunction with Newton's method can also be fruitfully applied to the convex programming problem

$$\begin{aligned} &\text{minimize} \quad f(x) \\ &\text{subject to} \quad Ax = b, \qquad x \geq 0, \end{aligned}$$

where $f : \Re^n \mapsto \Re$ is a convex function. The implementation of the method benefits from the extensive experience that has been accumulated from the linear programming case. For the special case of the quadratic programming problem

$$\begin{aligned} &\text{minimize} \quad c'x + \tfrac{1}{2}x'Qx \\ &\text{subject to} \quad Ax = b, \qquad x \geq 0, \end{aligned}$$

with Q positive semidefinite, the performance and the analysis of the method are similar to that for linear programs. We refer to the end-of-chapter references for a detailed treatment.

5.1.2 Primal-Dual Methods for Linear Programming

We will now discuss an alternative interior point method for solving the linear program

$$\begin{aligned} &\text{minimize} \quad c'x \\ &\text{subject to} \quad Ax = b, \qquad x \geq 0, \end{aligned} \tag{LP}$$

and its dual problem,

$$\begin{aligned} &\text{maximize} \quad b'\lambda \\ &\text{subject to} \quad A'\lambda \leq c. \end{aligned} \tag{DP}$$

Here as in the preceding section, $c \in \Re^n$ and $b \in \Re^m$ are given vectors, and A is an $m \times n$ matrix of rank m.

The logarithmic barrier method of the preceding section involves the sequential minimization of

$$F_{\epsilon^k}(x) = c'x - \epsilon^k \sum_{i=1}^{n} \ln x_i,$$

where S is the interior set

$$S = \{x \mid Ax = b,\ x > 0\},$$

and $\{\epsilon^k\}$ is a sequence that decreases to 0. This minimization is done approximately, using one or more Newton iterations.

We will now consider a related approach: applying Newton's method for solving the system of optimality conditions of the problem of minimizing $F_{\epsilon^k}(\cdot)$ over S. The salient features of this approach are:

(a) Only one Newton iteration is carried out for each value of ϵ^k.

(b) The Newton iterations generate a sequence of primal and dual solution pairs (x^k, λ^k), corresponding to a sequence of barrier parameters ϵ^k that converge to 0.

(c) For every k, the pair (x^k, λ^k) is such that x^k is an interior point of the positive orthant, i.e., $x^k > 0$, while λ^k is an interior point of the dual feasible region, i.e.,

$$c - A'\lambda^k > 0.$$

(However, x^k need not be primal-feasible, i.e., it need not satisfy the equation $Ax = b$ as it does in the path-following approach of Section 5.1.1.)

(d) Global convergence is enforced by using as merit function the expression

$$P^k = x^{k\prime} z^k + \|Ax^k - b\|, \tag{5.15}$$

where z^k is the vector

$$z^k = c - A'\lambda^k.$$

The expression (5.15) consists of two nonnegative terms: the first term is $x^{k\prime} z^k$, which is positive (since $x^k > 0$ and $z^k > 0$) and can be written as

$$x^{k\prime} z^k = x^{k\prime}(c - A'\lambda^k) = c'x^k - b'\lambda^k + (b - Ax^k)'\lambda^k.$$

Thus when x^k is primal-feasible ($Ax^k = b$), $x^{k\prime} z^k$ is equal to the duality gap, that is, the difference between the primal and the dual costs, $c'x^k - b'\lambda^k$. The second term is the norm of the primal constraint violation $\|Ax^k - b\|$. In the method to be described, neither of the terms $x^{k\prime}z^k$ and $\|Ax^k - b\|$ may increase at each iteration, so that $P^{k+1} \le P^k$ (and typically $P^{k+1} < P^k$) for all k. If we can show that $P^k \to 0$, then asymptotically both the duality gap and the primal constraint violation will be driven to zero. Thus every limit point of $\{(x^k, \lambda^k)\}$ will be a pair of primal and dual optimal solutions, in view of the duality relation

$$\min_{Ax=b,\, x\ge 0} c'x = \max_{A'\lambda \le c} b'\lambda,$$

shown in Section 4.4.2.

Let us write the necessary and sufficient conditions for (x, λ) to be a (global) minimum-Lagrange multiplier pair for the problem of minimizing the barrier function $F_\epsilon(x)$ subject to $Ax = b$. They are

$$c - \epsilon x^{-1} - A'\lambda = 0, \qquad Ax = b, \tag{5.16}$$

where x^{-1} denotes the vector with coordinates $(x_i)^{-1}$. Let z be the vector

$$z = c - A'\lambda,$$

and note that λ is dual feasible if and only if $z \ge 0$.

Using the vector z, we can write the first condition of Eq. (5.16) as $z - \epsilon x^{-1} = 0$ or, equivalently, $XZe = \epsilon e$, where X and Z are the diagonal matrices with the coordinates of x and z, respectively, along the diagonal, and e is the vector with unit coordinates:

$$X = \begin{pmatrix} x_1 & 0 & \cdots & 0 \\ 0 & x_2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & x_n \end{pmatrix}, \quad Z = \begin{pmatrix} z_1 & 0 & \cdots & 0 \\ 0 & z_2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & z_n \end{pmatrix}, \quad e = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}.$$

Thus the optimality conditions (5.16) can be written in the equivalent form

$$XZe = \epsilon e, \tag{5.17}$$

$$Ax = b, \tag{5.18}$$

$$z + A'\lambda = c. \tag{5.19}$$

Given (x, λ, z) satisfying $z + A'\lambda = c$, and such that $x > 0$ and $z > 0$, a Newton iteration for solving this system is

$$x(\alpha, \epsilon) = x + \alpha \Delta x, \tag{5.20}$$

$$\lambda(\alpha, \epsilon) = \lambda + \alpha \Delta \lambda,$$

$$z(\alpha, \epsilon) = z + \alpha \Delta z,$$

where α is a stepsize such that $0 < \alpha \leq 1$ and

$$x(\alpha, \epsilon) > 0, \qquad z(\alpha, \epsilon) > 0,$$

and the pure Newton step $(\Delta x, \Delta\lambda, \Delta z)$ solves the linearized version of the system (5.17)-(5.19)

$$X\Delta z + Z\Delta x = -v, \tag{5.21}$$

$$A\Delta x = b - Ax, \tag{5.22}$$

$$\Delta z + A'\Delta\lambda = 0, \tag{5.23}$$

with v defined by

$$v = XZe - \epsilon e. \tag{5.24}$$

After a straightforward calculation, it can be verified that the solution of the linearized system (5.21)-(5.23) can be written as

$$\Delta\lambda = \left(AZ^{-1}XA'\right)^{-1}\left(AZ^{-1}v + b - Ax\right), \tag{5.25}$$

$$\Delta z = -A'\Delta\lambda, \tag{5.26}$$

$$\Delta x = -Z^{-1}v - Z^{-1}X\Delta z.$$

Note that $\lambda(\alpha, \epsilon)$ is dual feasible, since from Eq. (5.23) and the condition $z + A'\lambda = c$, we see that $z(\alpha, \epsilon) + A'\lambda(\alpha, \epsilon) = c$. Note also that if $\alpha = 1$, i.e., a pure Newton step is used, $x(\alpha, \epsilon)$ is primal feasible, since from Eq. (5.22) we have $A(x + \Delta x) = b$.

Merit Function Improvement

We will now evaluate the changes in the constraint violation and the merit function induced by the Newton iteration. By using Eqs. (5.20)and (5.22), the new constraint violation is given by

$$Ax(\alpha, \epsilon) - b = Ax + \alpha A\Delta x - b = Ax + \alpha(b - Ax) - b = (1 - \alpha)(Ax - b). \tag{5.27}$$

Thus, since $0 < \alpha \leq 1$, the new norm of constraint violation $\|Ax(\alpha, \epsilon) - b\|$ is always no larger than the old one. Furthermore, if x is primal-feasible $(Ax = b)$, the new iterate $x(\alpha, \epsilon)$ is also primal-feasible.

The inner product

$$g = x'z \tag{5.28}$$

after the iteration becomes

$$\begin{aligned} g(\alpha, \epsilon) &= x(\alpha, \epsilon)'z(\alpha, \epsilon) \\ &= (x + \alpha\Delta x)'(z + \alpha\Delta z) \\ &= x'z + \alpha(x'\Delta z + z'\Delta x) + \alpha^2\Delta x'\Delta z. \end{aligned} \tag{5.29}$$

From Eqs. (5.22) and (5.26) we have

$$\Delta x' \Delta z = (Ax - b)' \Delta\lambda,$$

while by premultiplying Eq. (5.21) with e' and using the definition (5.24) for v, we obtain

$$x' \Delta z + z' \Delta x = -e'v = n\epsilon - x'z.$$

By substituting the last two relations in Eq. (5.29) and by using also the expression (5.28) for g, we see that

$$g(\alpha, \epsilon) = g - \alpha(g - n\epsilon) + \alpha^2 (Ax - b)' \Delta\lambda. \tag{5.30}$$

Let us now denote by P and $P(\alpha, \epsilon)$ the value of the merit function (5.15) before and after the iteration, respectively. We have by using the expressions (5.27) and (5.30),

$$\begin{aligned} P(\alpha, \epsilon) &= g(\alpha, \epsilon) + \big\|Ax(\alpha, \epsilon) - b\big\| \\ &= g - \alpha(g - n\epsilon) + \alpha^2 (Ax - b)' \Delta\lambda + (1 - \alpha)\|Ax - b\|, \end{aligned}$$

or

$$P(\alpha, \epsilon) = P - \alpha\big(g - n\epsilon + \|Ax - b\|\big) + \alpha^2 (Ax - b)' \Delta\lambda.$$

Thus if ϵ is chosen to satisfy

$$\epsilon < \frac{g}{n}$$

and α is chosen to be small enough so that the second order term $\alpha^2 (Ax - b)' \Delta\lambda$ is dominated by the first order term $\alpha(g - n\epsilon)$, the merit function will be improved as a result of the iteration.

A General Class of Primal-Dual Algorithms

Let us consider now the general class of algorithms of the form

$$x^{k+1} = x(\alpha^k, \epsilon^k), \qquad \lambda^{k+1} = \lambda(\alpha^k, \epsilon^k), \qquad z^{k+1} = z(\alpha^k, \epsilon^k),$$

where α^k and ϵ^k are positive scalars such that

$$x^{k+1} > 0, \qquad z^{k+1} > 0, \qquad \epsilon^k < \frac{g^k}{n},$$

where g^k is the inner product

$$g^k = x^{k\prime} z^k + (Ax^k - b)' \lambda^k,$$

and α^k is such that the merit function P^k is reduced. Initially we must have $x^0 > 0$, and $z^0 = c - A'\lambda^0 > 0$ (such a point can often be easily found; otherwise an appropriate reformulation of the problem is necessary for which we refer to the specialized literature). These methods have been called *primal-dual*. As we have seen, they are really Newton-like methods for approximate solution of the system of optimality conditions (5.16), supplemented by a stepsize procedure that guarantees that the merit function P^k is improved at each iteration.

It can be shown that it is possible to choose α^k and ϵ^k so that the merit function is not only reduced at each iteration, but also converges to zero. Furthermore, with suitable choices of α^k and ϵ^k, algorithms with good theoretical properties, such as polynomial complexity and superlinear convergence, can be derived. The main convergence analysis ideas rely on a primal-dual version of the central path discussed in Section 5.1.1, and some of the associated path following concepts. We refer to the research monograph [Wri97a] for a detailed discussion.

With properly chosen sequences α^k and ϵ^k, and appropriate implementation, the practical performance of the primal-dual methods has been shown to be excellent. The choice

$$\epsilon^k = \frac{g^k}{n^2},$$

leading to the relation

$$g^{k+1} = (1 - \alpha^k + \alpha^k/n)g^k$$

for feasible x^k, has been suggested as a good practical rule. Usually, when x^k has already become feasible, α^k is chosen as $\theta\tilde{\alpha}^k$, where θ is a factor very close to 1 (say 0.999), and $\tilde{\alpha}^k$ is the maximum stepsize α that guarantees that $x(\alpha, \epsilon^k) \geq 0$ and $z(\alpha, \epsilon^k) \geq 0$

$$\tilde{\alpha}^k = \min\left\{ \min_{i=1,\ldots,n} \left\{ \frac{x_i^k}{-\Delta x_i} \;\Big|\; \Delta x_i < 0 \right\}, \min_{i=1,\ldots,n} \left\{ \frac{z_i^k}{-\Delta z_i} \;\Big|\; \Delta z_i < 0 \right\} \right\}.$$

When x^k is not feasible, the choice of α^k must also be such that the merit function is improved. In some works, a different stepsize for the x update than for the (λ, z) update has been suggested. The stepsize for the x update is near the maximum stepsize α that guarantees $x(\alpha, \epsilon^k) \geq 0$, and the stepsize for the (λ, z) update is near the maximum stepsize α that guarantees $z(\alpha, \epsilon^k) \geq 0$. There are a number of additional practical issues related to implementation, for which we refer to the specialized literature.

Predictor-Corrector Variants

We mentioned briefly in Section 1.4.3 the variation of Newton's method where the Hessian is evaluated periodically every $p > 1$ iterations in order to economize in iteration overhead. When $p = 2$ and the problem is to solve the system $g(x) = 0$, where $g : \Re^n \mapsto \Re^n$, this variation of Newton's method takes the form

$$\hat{x}^k = x^k - \big(\nabla g(x^k)'\big)^{-1} g(x^k), \tag{5.31}$$

$$x^{k+1} = \hat{x}^k - \big(\nabla g(x^k)'\big)^{-1} g(\hat{x}^k). \tag{5.32}$$

Thus, given x^k, this iteration performs a regular Newton step to obtain $\hat{x}^k$, and then an approximate Newton step from $\hat{x}^k$, using, however, the already available Jacobian inverse $\big(\nabla g(x^k)'\big)^{-1}$. It can be shown that if $x^k \to x^*$, the order of convergence of the error $\|x^k - x^*\|$ is cubic, i.e.,

$$\limsup_{k\to\infty} \frac{\|x^{k+1} - x^*\|}{\|x^k - x^*\|^3} < \infty,$$

under the same assumptions that the ordinary Newton's method ($p = 1$) attains a quadratic order of convergence; see [OrR70], p. 315. Thus, the price for the 50% saving in Jacobian evaluations and inversions is a small degradation of the convergence rate over the ordinary Newton's method (which attains a quartic order of convergence when two successive ordinary Newton steps are counted as one).

Two-step Newton methods such as the iteration (5.31), (5.32), when applied to the system of optimality conditions (5.17)-(5.19) for the linear program (LP) are known as *predictor-corrector* methods (the name comes from their similarity with predictor-corrector methods for solving differential equations). They operate as follows:

Given (x, z, λ) with

$$x > 0, \qquad z = c - A'\lambda > 0,$$

the *predictor iteration* [cf. Eq. (5.31)], solves for $\big(\Delta\hat{x}, \Delta\hat{z}, \Delta\hat{\lambda}\big)$ the system

$$X\Delta\hat{z} + Z\Delta\hat{x} = -\hat{v}, \tag{5.33}$$

$$A\Delta\hat{x} = b - Ax, \tag{5.34}$$

$$\Delta\hat{z} + A'\Delta\hat{\lambda} = 0, \tag{5.35}$$

with $\hat{v}$ defined by

$$\hat{v} = XZe - \hat{\epsilon}e, \tag{5.36}$$

[cf. Eqs. (5.21)-(5.24)].

The *corrector iteration* [cf. Eq. (5.32)], solves for $\big(\Delta\bar{x}, \Delta\bar{z}, \Delta\bar{\lambda}\big)$ the system

$$X\Delta\bar{z} + Z\Delta\bar{x} = -\bar{v}, \tag{5.37}$$

$$A\Delta\bar{x} = b - A(x + \Delta\hat{x}), \tag{5.38}$$

$$\Delta\bar{z} + A'\Delta\bar{\lambda} = 0, \tag{5.39}$$

with $\bar{v}$ defined by

$$\bar{v} = (X + \Delta\hat{X})(Z + \Delta\hat{Z})e - \bar{\epsilon}e, \tag{5.40}$$

where $\Delta\hat{X}$ and $\Delta\hat{Z}$ are the diagonal matrices with the components of $\Delta\hat{x}$ and $\Delta\hat{z}$ along the diagonal, respectively. Here $\hat{\epsilon}$ and $\bar{\epsilon}$ are the barrier parameters corresponding to the two iterations.

The *composite Newton direction* is

$$\Delta x = \Delta\hat{x} + \Delta\bar{x},$$

$$\Delta z = \Delta\hat{z} + \Delta\bar{z},$$

$$\Delta\lambda = \Delta\hat{\lambda} + \Delta\bar{\lambda},$$

and the corresponding iteration is

$$x(\alpha, \epsilon) = x + \alpha\Delta x,$$

$$\lambda(\alpha, \epsilon) = \lambda + \alpha\Delta\lambda,$$

$$z(\alpha, \epsilon) = z + \alpha\Delta z,$$

where α is a stepsize such that $0 < \alpha \leq 1$ and

$$x(\alpha, \epsilon) > 0, \qquad z(\alpha, \epsilon) > 0.$$

Adding Eqs. (5.33)-(5.35)and Eqs. (5.37)-(5.39), we obtain

$$X(\Delta\hat{z} + \Delta\bar{z})z + Z(\Delta\hat{x} + \Delta\bar{x}) = -\hat{v} - \bar{v}, \tag{5.41}$$

$$A(\Delta\hat{x} + \Delta\bar{x})x = b - Ax + b - A(x + \Delta\hat{x}), \tag{5.42}$$

$$\Delta\hat{z} + \Delta\bar{z} + A'(\Delta\hat{\lambda} + \Delta\bar{\lambda}) = 0, \tag{5.43}$$

We now use the fact

$$b - A(x + \Delta\hat{x}) = 0$$

[cf. Eq. (5.34)], and we also use Eqs. (5.40) and (5.33) to write

$$\begin{aligned}
\bar{v} &= (X + \Delta\hat{X})(Z + \Delta\hat{Z})e - \bar{\epsilon}e \\
&= XZe + \Delta\hat{X}Ze + X\Delta\hat{Z}e + \Delta\hat{X}\Delta\hat{Z}e - \bar{\epsilon}e \\
&= XZe + Z\Delta\hat{x} + X\Delta\hat{z} + \Delta\hat{X}\Delta\hat{Z}e - \bar{\epsilon}e \\
&= XZe - \hat{v} + \Delta\hat{X}\Delta\hat{Z}e - \bar{\epsilon}e.
\end{aligned}$$

Substituting in Eqs. (5.41)-(5.43), we see that the composite Newton direction

$$(\Delta x, \Delta z, \Delta\lambda) = (\Delta\hat{x} + \Delta\bar{x}, \Delta\hat{z} + \Delta\bar{z}, \Delta\hat{\lambda} + \Delta\bar{\lambda})$$

is obtained by solving the following system of equations:

$$X\Delta z + Z\Delta x = -XZe - \Delta\hat{X}\Delta\hat{Z}e + \bar{\epsilon}e, \tag{5.44}$$

$$A\Delta x = b - Ax, \tag{5.45}$$

$$\Delta z + A'\Delta\lambda = 0. \tag{5.46}$$

To implement the predictor-corrector method, we need to solve the system (5.33)-(5.36) for some value of $\hat{\epsilon}$ to obtain $\big(\Delta\hat{X}, \Delta\hat{Z}\big)$, and then to solve the system (5.44)-(5.46) for some value of $\bar{\epsilon}$ to obtain $\big(\Delta x, \Delta z, \Delta\lambda\big)$. It is important to note here that most of the work needed for the first system, namely the factorization of the matrix

$$AZ^{-1}XA'$$

in Eq. (5.25), need not be repeated when solving the second system, so that solving both systems requires relatively little extra work over solving the first one.

In an implementation that has proved successful in practice, $\hat{\epsilon}$ is taken to be zero. Furthermore, $\bar{\epsilon}$ is chosen on the basis of the solution of the first system according to the formula

$$\bar{\epsilon} = \left(\frac{\hat{g}}{x'z}\right)^2 \frac{\hat{g}}{n},$$

where $\hat{g}$ is the duality gap that would result from a feasibility-restricted primal-dual step given by

$$\hat{g} = (x + \alpha_P\Delta\hat{x})'(z + \alpha_D\Delta\hat{z}),$$

where

$$\alpha_P = \theta \min_{i=1,\ldots,n}\left\{\frac{x_i^k}{-\Delta x_i} \;\Big|\; \Delta x_i < 0\right\},$$

$$\alpha_D = \theta \min_{i=1,\ldots,n}\left\{\frac{z_i^k}{-\Delta z_i} \;\Big|\; \Delta z_i < 0\right\}.$$

and θ is a factor very close to 1 (say 0.999). We refer to the specialized literature for further details [Meh92], [LMS92], [Wri97a], [Ye97].

E X E R C I S E S

5.1.1

Consider the linear program

$$\begin{aligned} &\text{minimize} \;\; x_1 + 2x_2 + 3x_3 \\ &\text{subject to} \;\; x_1 + x_2 + x_3 = 1, \qquad x \geq 0. \end{aligned}$$

(a) Sketch on paper the central path. Write a computer program to implement a short-step and a long-step path-following method based on Newton's method for this problem. Compare the number of Newton steps for a given solution accuracy for the starting points $x^0 = (.8, .15, .05)$ and $x^0 = (.1, .2, .7)$.

(b) Write a computer program to implement a primal-dual interior point method and its predictor-corrector variant, and solve the problem for $\lambda^0 = 0$ and x^0 as in part (a).

5.1.2

Given x, show how to find an $\epsilon > 0$ that minimizes $\big\|q(x,\epsilon)\big\|$ [cf. Eq. (5.5)]. How would you use this idea to accelerate convergence in a short-step path-following method?

5.1.3

Show that the vector λ of Eq. (5.7) satisfies

$$\lambda \in \arg \min_{\xi \in \Re^m} \left\| \frac{X(c - A'\xi)}{\epsilon} - e \right\|.$$

5.1.4

Let $\delta = \big\|q(x,\epsilon)\big\|$, where $q(x,\epsilon)$ is the scaled Newton step defined by Eq. (5.5), and assume that $\delta < 1$. Show that

$$\big\|X^{-1}\big(x - x(\epsilon)\big)\big\| \leq \frac{\delta}{1-\delta},$$

$$F_\epsilon(x) - F_\epsilon\big(x(\epsilon)\big) \leq \frac{\delta^2}{1-\delta^2},$$

$$\big|c'x - c'x(\epsilon)\big| \leq \frac{\delta(1+\delta)\epsilon}{1-\delta}\sqrt{n}.$$

5.1.5 (Complexity of the Short-Step Method [Tse89])

The purpose of this exercise is to show that the number of iterations required by the short-step logarithmic barrier method to achieve a given accuracy is proportional to $\sqrt{n}$. Consider the linear programming problem (LP), a vector $x^0 \in S$, and a sequence $\{\epsilon^k\}$ such that $\left\|q(x^0, \epsilon^0)\right\| \leq 1/2$ and $\epsilon^{k+1} = (1-\theta)\epsilon^k$, where $\theta = 1/(6n^{1/2})$ (cf. Prop. 5.1.4). Let x^{k+1} be generated from x^k by a pure Newton step aimed at minimizing F_{ϵ^k}. For a given integer r, let $\bar{k}$ be the smallest integer k such that $-\ln(n\epsilon^k) \geq r$ and let $r^0 = -\ln(n\epsilon^0)$. Show that

$$\bar{k} \leq 6(r - r^0)\sqrt{n}$$

and

$$c'x^{\bar{k}} - f^* \leq \frac{3}{2}e^{-r}.$$

Note: We have assumed here that a vector x^0 with $\left\|q(x^0, \epsilon^0)\right\| \leq 1/2$ is available. It is possible to show that such a point can be found in a number of Newton steps that is proportional to $\sqrt{n}$.

5.1.6 (The Dual Problem as an Equality Constrained Problem)

Consider the dual problem

$$\begin{aligned} &\text{maximize} \ \ b'\lambda \\ &\text{subject to} \ \ A'\lambda \leq c, \end{aligned} \tag{DP}$$

and its equivalent version

$$\begin{aligned} &\text{maximize} \ \ b'\lambda \\ &\text{subject to} \ \ A'\lambda + z = c, \quad z \geq 0, \end{aligned}$$

that involves the vector of additional variables z. Let P_A be the matrix that projects a vector x onto the nullspace of the matrix A, and note that using the analysis of Example 3.1.5 in Section 3.1, we have

$$P_A = I - A'(AA')^{-1}A.$$

Show that the dual linear program (DP) is equivalent to the linear program

$$\begin{aligned} &\text{minimize} \ \ \bar{x}'z \\ &\text{subject to} \ \ P_A z = P_A c, \quad z \geq 0, \end{aligned} \tag{5.47}$$

where $\bar{x}$ is any primal feasible vector, i.e., $A\bar{x} = b$, $\bar{x} \geq 0$.

5.1.7 (Dual Central Path)

Consider the dual problem (DP). Using its equivalent reformulation (5.47) of Exercise 5.1.6, it is seen that the appropriate definition of the central path of the dual problem is

$$z(\epsilon) \in \arg \min_{P_A z = P_A c,\, z>0} \left\{ \bar{x}' z - \sum_{i=1}^{n} \ln z_i \right\},$$

where $\bar{x}$ is any primal feasible vector. Show that the primal and dual central paths are related by

$$z(\epsilon) = \epsilon x(\epsilon)^{-1},$$

and that the corresponding duality gap satisfies

$$c' x(\epsilon) - b' \lambda(\epsilon) = n\epsilon,$$

where $\lambda(\epsilon)$ is any vector such that $A'\lambda(\epsilon) = c - z(\epsilon)$.

5.2 PENALTY AND AUGMENTED LAGRANGIAN METHODS

The basic idea in penalty methods is to eliminate some or all of the constraints and add to the cost function a penalty term that prescribes a high cost to infeasible points. Associated with these methods is a penalty parameter c that determines the severity of the penalty and as a consequence, the extent to which the resulting unconstrained problem approximates the original constrained problem. As c takes higher values, the approximation becomes increasingly accurate. We focus attention primarily on the popular quadratic penalty function. Some other penalty functions, including the exponential, are discussed in Section 5.2.5.

Consider first the equality constrained problem

$$\begin{aligned} &\text{minimize} \quad f(x) \\ &\text{subject to} \quad h(x) = 0, \qquad x \in X, \end{aligned} \tag{5.48}$$

where $f : \Re^n \mapsto \Re$, $h : \Re^n \mapsto \Re^m$ are given functions, and X is a given subset of $\Re^n$. Much of our analysis in this section will focus on the case where $X = \Re^n$, and x^* together with a Lagrange multiplier vector λ^* satisfies the sufficient optimality conditions of Prop. 4.2.1. At the center of our development is the *augmented Lagrangian function* $L_c : \Re^n \times \Re^m \mapsto \Re$, introduced in Section 4.2 and given by

$$L_c(x, \lambda) = f(x) + \lambda' h(x) + \frac{c}{2} \|h(x)\|^2,$$

where c is a positive penalty parameter.

There are two mechanisms by which unconstrained minimization of $L_c(\cdot, \lambda)$ can yield points close to x^*:

(a) *By taking λ close to λ^*.* Indeed, as shown in Section 4.2.1, if c is higher than a certain threshold, then for some $\gamma > 0$ and $\epsilon > 0$ we have

$$L_c(x, \lambda^*) \geq L_c(x^*, \lambda^*) + \frac{\gamma}{2}\|x - x^*\|^2, \qquad \forall\, x \text{ with } \|x - x^*\| < \epsilon,$$

so that x^* is a strict unconstrained local minimum of the augmented Lagrangian $L_c(\cdot, \lambda^*)$ corresponding to λ^*. This suggests that if λ is close to λ^*, a good approximation to x^* can be found by unconstrained minimization of $L_c(\cdot, \lambda)$.

(b) *By taking c large.* Indeed for high c, there is high cost for infeasibility, so the unconstrained minima of $L_c(\cdot, \lambda)$ will be nearly feasible. Since $L_c(x, \lambda) = f(x)$ for feasible x, we expect that $L_c(x, \lambda) \approx f(x)$ for nearly feasible x. Therefore, we can also expect to obtain a good approximation to x^* by unconstrained minimization of $L_c(\cdot, \lambda)$ when c is large.

Example 5.2.1

Consider the two-dimensional problem

$$\begin{aligned} &\text{minimize} \quad f(x) = \tfrac{1}{2}(x_1^2 + x_2^2) \\ &\text{subject to} \quad x_1 = 1, \end{aligned}$$

with optimal solution $x^* = (1, 0)$ and corresponding Lagrange multiplier $\lambda^* = -1$. The augmented Lagrangian is

$$L_c(x, \lambda) = \tfrac{1}{2}(x_1^2 + x_2^2) + \lambda(x_1 - 1) + \tfrac{c}{2}(x_1 - 1)^2,$$

and by setting its gradient to zero we can verify that its unique unconstrained minimum $x(\lambda, c)$ has coordinates given by

$$x_1(\lambda, c) = \frac{c - \lambda}{c + 1}, \qquad x_2(\lambda, c) = 0. \tag{5.49}$$

Thus, we have for all $c > 0$,

$$\lim_{\lambda \to \lambda^*} x_1(\lambda, c) = x_1(-1, c) = 1 = x_1^*, \qquad \lim_{\lambda \to \lambda^*} x_2(\lambda^*, c) = 0 = x_2^*,$$

showing that as λ is chosen close to λ^*, the unconstrained minimum of $L_c(\cdot, \lambda)$ approaches the constrained minimum (see Fig. 5.2.1).

Using Eq. (5.49), we also have for all λ,

$$\lim_{c \to \infty} x_1(\lambda, c) = 1 = x_1^*, \qquad \lim_{c \to \infty} x_2(\lambda, c) = 0 = x_2^*,$$

showing that as c increases, the unconstrained minimum of $L_c(\cdot, \lambda)$ approaches the constrained minimum (see Fig. 5.2.2).

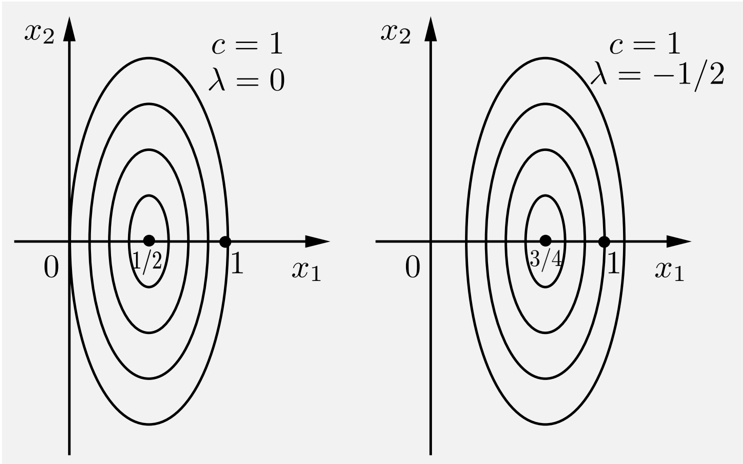

Figure 5.2.1. Equal cost surfaces of the augmented Lagrangian

$$L_c(x, \lambda) = \tfrac{1}{2}(x_1^2 + x_2^2) + \lambda(x_1 - 1) + \tfrac{c}{2}(x_1 - 1)^2,$$

of Example 5.2.1, for $c = 1$ and two different values of λ. The unconstrained minimum of $L_c(\cdot, \lambda)$ approaches the constrained minimum $x^* = (1, 0)$ as $\lambda \to \lambda^* = -1$.

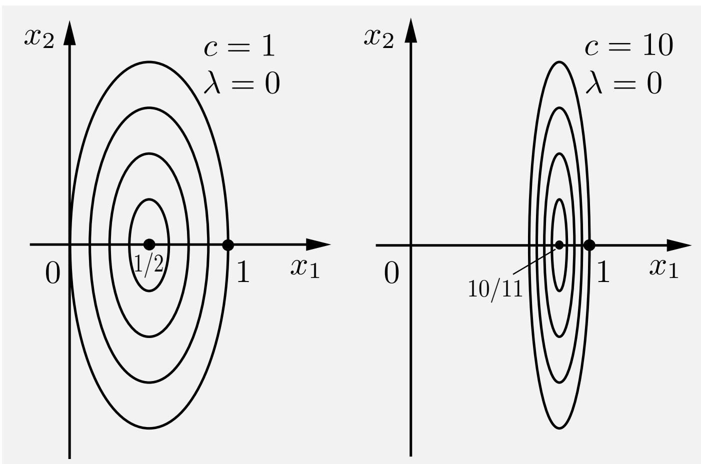

Figure 5.2.2. Equal cost surfaces of the augmented Lagrangian

$$L_c(x, \lambda) = \tfrac{1}{2}(x_1^2 + x_2^2) + \lambda(x_1 - 1) + \tfrac{c}{2}(x_1 - 1)^2,$$

of Example 5.2.1, for $\lambda = 0$ and two different values of c. The unconstrained minimum of $L_c(\cdot, \lambda)$ approaches the constrained minimum $x^* = (1, 0)$ as $c \to \infty$.

5.2.1 The Quadratic Penalty Function Method

The quadratic penalty function method is motivated by the preceding considerations. It consists of solving a sequence of problems of the form

$$\begin{aligned} &\text{minimize} \;\; L_{c^k}(x, \lambda^k) \\ &\text{subject to} \;\; x \in X, \end{aligned}$$

where $\{\lambda^k\}$ is a sequence in $\Re^m$ and $\{c^k\}$ is a penalty parameter sequence.

In the original version of the penalty method introduced in the early 60s, the multipliers λ^k were taken to be equal to zero. The idea of using λ^k that are "good" approximations to a Lagrange multiplier vector was not known at that time. In our development here, we allow λ^k to change in the course of the algorithm, but for the moment we give no rule for updating λ^k. Thus the method depends for its validity on increasing c^k to ∞, and applies even if the problem has no Lagrange multiplier. The following proposition is the basic convergence result.

Proposition 5.2.1: Assume that f and h are continuous functions, that X is a closed set, and that the constraint set $\big\{x \in X \mid h(x) = 0\big\}$ is nonempty. For $k = 0, 1, \ldots$, let x^k be a global minimum of the problem

$$\begin{aligned} &\text{minimize} \;\; L_{c^k}(x, \lambda^k) \\ &\text{subject to} \;\; x \in X, \end{aligned}$$

where $\{\lambda^k\}$ is bounded, $0 < c^k < c^{k+1}$ for all k, and $c^k \to \infty$. Then every limit point of the sequence $\{x^k\}$ is a global minimum of the original problem (5.48).

Proof: Let $\bar{x}$ be a limit point of $\{x^k\}$. We have by definition of x^k

$$L_{c^k}(x^k, \lambda^k) \le L_{c^k}(x, \lambda^k), \qquad \forall\, x \in X. \tag{5.50}$$

Let f^* denote the optimal value of the original problem (5.48). We have

$$\begin{aligned} f^* &= \inf_{h(x)=0,\, x \in X} f(x) \\ &= \inf_{h(x)=0,\, x \in X} \left\{ f(x) + \lambda^{k\prime} h(x) + \frac{c^k}{2} \big\| h(x) \big\|^2 \right\} \\ &= \inf_{h(x)=0,\, x \in X} L_{c^k}(x, \lambda^k). \end{aligned}$$

Hence, by taking the infimum of the right-hand side of Eq. (5.50) over $x \in X$, $h(x) = 0$, we obtain

$$L_{c^k}(x^k, \lambda^k) = f(x^k) + \lambda^{k\prime} h(x^k) + \frac{c^k}{2} \big\| h(x^k) \big\|^2 \le f^*.$$

The sequence $\{\lambda^k\}$ is bounded and hence it has a limit point $\bar{\lambda}$. Without loss of generality, we may assume that $\lambda^k \to \bar{\lambda}$. By taking upper limit in the relation above and by using the continuity of f and h, we obtain

$$f(\bar{x}) + \bar{\lambda}' h(\bar{x}) + \limsup_{k\to\infty} \frac{c^k}{2} \big\|h(x^k)\big\|^2 \leq f^*. \tag{5.51}$$

Since $\big\|h(x^k)\big\|^2 \geq 0$ and $c^k \to \infty$, it follows that $h(x^k) \to 0$ and

$$h(\bar{x}) = 0, \tag{5.52}$$

for otherwise the left-hand side of Eq. (5.51) would equal ∞, while $f^* < \infty$ (since the constraint set is assumed nonempty). Since X is a closed set, we also obtain that $\bar{x} \in X$. Hence, $\bar{x}$ is feasible, and since from Eqs. (5.51) and (5.52) we have $f(\bar{x}) \leq f^*$, it follows that $\bar{x}$ is optimal. **Q.E.D.**

Lagrange Multiplier Estimates – Inexact Minimization

The preceding convergence result assumes implicitly that the minimum of the augmented Lagrangian is found exactly. On the other hand, unconstrained minimization methods are usually terminated when the cost gradient is sufficiently small, but not necessarily zero. In particular, when $X = \Re^n$, and f and h are differentiable, the algorithm for solving the unconstrained problem

$$\begin{aligned} &\text{minimize} \quad L_{c^k}(x, \lambda^k) \\ &\text{subject to} \quad x \in \Re^n, \end{aligned}$$

will typically be terminated at a point x^k satisfying

$$\big\|\nabla_x L_{c^k}(x^k, \lambda^k)\big\| \leq \epsilon^k,$$

where ϵ^k is some small scalar. We address this situation in the next proposition, where we show in addition that we can usually obtain a Lagrange multiplier vector as a by-product of the computation.

Proposition 5.2.2: Assume that $X = \Re^n$, and f and h are continuously differentiable. For $k = 0, 1, \ldots$, let x^k satisfy

$$\big\|\nabla_x L_{c^k}(x^k, \lambda^k)\big\| \leq \epsilon^k,$$

where $\{\lambda^k\}$ is bounded, and $\{\epsilon^k\}$ and $\{c^k\}$ satisfy

$$0 < c^k < c^{k+1}, \quad \forall\, k, \qquad c^k \to \infty,$$

$$0 \leq \epsilon^k, \quad \forall\, k, \qquad \epsilon^k \to 0.$$

Assume that a subsequence $\{x^k\}_K$ converges to a vector x^* such that $\nabla h(x^*)$ has rank m. Then

$$\big\{\lambda^k + c^k h(x^k)\big\}_K \to \lambda^*,$$

where λ^* is a vector satisfying, together with x^*, the first order necessary conditions

$$\nabla f(x^*) + \nabla h(x^*)\lambda^* = 0, \qquad h(x^*) = 0.$$

Proof: Without loss of generality we assume that the entire sequence $\{x^k\}$ converges to x^*. Define for all k

$$\tilde{\lambda}^k = \lambda^k + c^k h(x^k).$$

We have

$$\nabla_x L_{c^k}(x^k, \lambda^k) = \nabla f(x^k) + \nabla h(x^k)\big(\lambda^k + c^k h(x^k)\big) = \nabla f(x^k) + \nabla h(x^k)\tilde{\lambda}^k. \tag{5.53}$$

Since $\nabla h(x^*)$ has rank m, $\nabla h(x^k)$ has rank m for all k that are sufficiently large. Without loss of generality, we assume that $\nabla h(x^k)$ has rank m for all k. Then, by multiplying Eq. (5.53) with

$$\big(\nabla h(x^k)'\nabla h(x^k)\big)^{-1}\nabla h(x^k)',$$

we obtain

$$\tilde{\lambda}^k = \big(\nabla h(x^k)'\nabla h(x^k)\big)^{-1}\nabla h(x^k)'\big(\nabla_x L_{c^k}(x^k, \lambda^k) - \nabla f(x^k)\big). \tag{5.54}$$

The hypothesis implies that $\nabla_x L_{c^k}(x^k, \lambda^k) \to 0$, so Eq. (5.54) yields

$$\tilde{\lambda}^k \to \lambda^*,$$

where

$$\lambda^* = -\big(\nabla h(x^*)'\nabla h(x^*)\big)^{-1}\nabla h(x^*)'\nabla f(x^*).$$

Using again the fact $\nabla_x L_{c^k}(x^k, \lambda^k) \to 0$ and Eq. (5.53), we see that

$$\nabla f(x^*) + \nabla h(x^*)\lambda^* = 0.$$

Since $\{\lambda^k\}$ is bounded and $\lambda^k + c^k h(x^k) \to \lambda^*$, it follows that $\big\{c^k h(x^k)\big\}$ is bounded. Since $c^k \to \infty$, we must have $h(x^k) \to 0$ and we conclude that $h(x^*) = 0$. **Q.E.D.**

Practical Behavior – Ill-Conditioning

Let us now consider the practical behavior of the quadratic penalty method, assuming that $X = \Re^n$, and f and h are continuously differentiable. Suppose that the kth unconstrained minimization of $L_{c^k}(x, \lambda^k)$ is terminated when

$$\big\|\nabla_x L_{c^k}(x^k, \lambda^k)\big\| \leq \epsilon^k, \tag{5.55}$$

where $\epsilon^k \to 0$. There are three possibilities:

(a) The method breaks down because an x^k satisfying the condition (5.55) cannot be found.

(b) A sequence $\{x^k\}$ satisfying the condition (5.55) for all k is obtained, but it either has no limit points, or for each of its limit points x^* the matrix $\nabla h(x^*)$ has linearly dependent columns.

(c) A sequence $\{x^k\}$ satisfying the condition (5.55) for all k is found and it has a limit point x^* such that $\nabla h(x^*)$ has rank m. Then, by Prop. 5.2.2, x^* together with λ^* [the corresponding limit point of $\big\{\lambda^k + c^k h(x^k)\big\}$] satisfies the first order necessary conditions for optimality.

Possibility (a) usually occurs when $L_{c^k}(\cdot, \lambda^k)$ is unbounded below as discussed following Prop. 5.2.1.

Possibility (b) usually occurs when $L_{c^k}(\cdot, \lambda^k)$ is bounded below, but the original problem has no feasible solution. Typically then the penalty term dominates as $k \to \infty$, and the method usually converges to an infeasible vector x^*, which is a stationary point of the function $\big\|h(x)\big\|^2$. This means that

$$\nabla h(x^*)h(x^*) = \tfrac{1}{2}\nabla\big\{\|h(x^*)\|^2\big\} = 0,$$

implying that $\nabla h(x^*)$ has linearly dependent columns.

Possibility (c) is the normal case, where the unconstrained minimization algorithm terminates successfully for each k and $\{x^k\}$ converges to a feasible vector, which is also regular. It is of course possible (although unusual in practice) that $\{x^k\}$ converges to a local minimum x^*, which is not regular. Then, if there is no Lagrange multiplier vector corresponding to x^*, the sequence $\big\{\lambda^k + c^k h(x^k)\big\}$ diverges and has no limit point.

Extensive practical experience shows that the penalty function method is on the whole quite reliable and usually converges to at least a local minimum of the original problem. Whenever it fails, this is usually due to the increasing difficulty of the minimization

$$\begin{aligned} &\text{minimize} \ \ L_{c^k}(x, \lambda^k) \\ &\text{subject to} \ \ x \in X, \end{aligned}$$

as $c^k \to \infty$. In particular, let us assume that $X = \Re^n$, and f and h are twice differentiable. Then, according to the convergence rate analysis of Section

1.3, the degree of difficulty depends on the ratio of largest to smallest eigenvalue (the condition number) of the Hessian matrix $\nabla^2_{xx}L_{c^k}(x^k,\lambda^k)$, and this ratio tends to increase with c^k. We illustrate this by means of an example. A proof is outlined in Exercise 5.2.8; see also [Ber82a], p. 102.

Example 5.2.2

Consider the problem of Example 5.2.1:

$$\begin{aligned} &\text{minimize} \quad f(x) = \tfrac{1}{2}(x_1^2 + x_2^2) \\ &\text{subject to} \quad x_1 = 1. \end{aligned}$$

The augmented Lagrangian is

$$L_c(x,\lambda) = \tfrac{1}{2}(x_1^2 + x_2^2) + \lambda(x_1 - 1) + \tfrac{c}{2}(x_1 - 1)^2,$$

and its Hessian is

$$\nabla^2_{xx}L_c(x,\lambda) = I + c\begin{pmatrix} 1 \\ 0 \end{pmatrix}\begin{pmatrix} 1 & 0 \end{pmatrix} = \begin{pmatrix} 1+c & 0 \\ 0 & 1 \end{pmatrix}.$$

The ratio of largest to smallest eigenvalue of the Hessian is $1+c$ and tends to ∞ as $c \to \infty$. The associated ill-conditioning can also be observed from the narrow level sets of the augmented Lagrangian for large c in Fig. 5.2.2.

To overcome ill-conditioning, it is recommended to use a Newton-like method for minimizing $L_{c^k}(\cdot,\lambda^k)$, as well as double precision arithmetic to deal with roundoff errors. It is common to adopt as a starting point for minimizing $L_{c^k}(\cdot,\lambda^k)$ the last point x^{k-1} of the previous minimization. In order, however, for x^{k-1} to be near a minimizing point of $L_{c^k}(\cdot,\lambda^k)$, it is necessary that c^k is close to c^{k-1}. This in turn requires that the rate of increase of the penalty parameter c^k should be relatively small. There is a basic tradeoff here. If c^k is increased at a fast rate, then $\{x^k\}$ converges faster, but the likelihood of ill-conditioning is greater. Usually, a sequence $\{c^k\}$ generated by $c^{k+1} = \beta c^k$ with β in the range $[4, 10]$ works well if a Newton-like method is used for minimizing $L_{c^k}(\cdot,\lambda^k)$; otherwise, a smaller value of β may be needed. Some trial and error may be needed to choose the initial penalty parameter c^0, since there is no safe guideline on how to determine this value. For an indication of this, note that if the problem functions f and h are scaled by multiplication with a scalar $s > 0$, then c^0 should be divided by s to maintain the same condition number for the Hessian of the augmented Lagrangian.

Inequality Constraints

The simplest way to treat inequality constraints in the context of the quadratic penalty method, is to convert them to equality constraints by using squared additional variables. We have already used this device in our discussion of optimality conditions for inequality constraints in Section 4.3.2.

Consider the problem

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & h_1(x) = 0, \ldots, h_m(x) = 0, \\ & g_1(x) \leq 0, \ldots, g_r(x) \leq 0. \end{array} \tag{5.56}$$

As discussed in Section 4.3.2, we can convert this problem to the equality constrained problem

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & h_1(x) = 0, \ldots, h_m(x) = 0, \\ & g_1(x) + z_1^2 = 0, \ldots, g_r(x) + z_r^2 = 0, \end{array} \tag{5.57}$$

where $z_1, \ldots, z_r$ are additional variables. The quadratic penalty method for this problem involves unconstrained minimizations of the form

$$\begin{aligned} \min_{x,z} \bar{L}_c(x, z, \lambda, \mu) = {} & f(x) + \lambda' h(x) + \frac{c}{2} \big\| h(x) \big\|^2 \\ & + \sum_{j=1}^{r} \Big\{ \mu_j \big(g_j(x) + z_j^2 \big) + \frac{c}{2} \big| g_j(x) + z_j^2 \big|^2 \Big\}, \end{aligned}$$

for various values of λ, μ, and c. This type of minimization can be done by first minimizing $\bar{L}_c(x, z, \lambda, \mu)$ with respect to z, obtaining

$$L_c(x, \lambda, \mu) = \min_{z} \bar{L}_c(x, z, \lambda, \mu),$$

and then by minimizing $L_c(x, \lambda, \mu)$ with respect to x. A key observation is that *the first minimization with respect to z can be carried out in closed form for each fixed x*, thereby yielding a closed form expression for $L_c(x, \lambda, \mu)$.

Indeed, we have

$$\begin{aligned} \min_{z} \bar{L}_c(x, z, \lambda, \mu) = {} & f(x) + \lambda' h(x) + \frac{c}{2} \big\| h(x) \big\|^2 \\ & + \sum_{j=1}^{r} \min_{z_j} \Big\{ \mu_j \big(g_j(x) + z_j^2 \big) + \frac{c}{2} \big| g_j(x) + z_j^2 \big|^2 \Big\}, \end{aligned} \tag{5.58}$$

and the minimization with respect to z_j in the last term is equivalent to

$$\min_{u_j \geq 0} \Big\{ \mu_j \big(g_j(x) + u_j \big) + \frac{c}{2} \big| g_j(x) + u_j \big|^2 \Big\}.$$

The function in braces above is quadratic in u_j. Its constrained minimum is $u_j^* = \max\{0, \hat{u}_j\}$, where $\hat{u}_j$ is the unconstrained minimum at which the derivative, $\mu_j + c\big(g_j(x) + \hat{u}_j\big)$, is zero. Thus,

$$u_j^* = \max\left\{0, -\left(\frac{\mu_j}{c} + g_j(x)\right)\right\}.$$

Denoting

$$g_j^+(x, \mu, c) = \max\left\{g_j(x), -\frac{\mu_j}{c}\right\}, \tag{5.59}$$

we have $g_j(x) + u_j^* = g_j^+(x, \mu, c)$. Substituting this expression in Eq. (5.58), we obtain a closed form expression for $L_c(x, \lambda, \mu) = \min_z \bar{L}_c(x, z, \lambda, \mu)$ given by

$$\begin{aligned} L_c(x, \lambda, \mu) = {} & f(x) + \lambda' h(x) + \frac{c}{2}\big\|h(x)\big\|^2 \\ & + \sum_{j=1}^r \left\{\mu_j g_j^+(x, \mu, c) + \frac{c}{2}\left(g_j^+(x, \mu, c)\right)^2\right\}. \end{aligned} \tag{5.60}$$

After some calculation, left for the reader, we can also write this expression as

$$\begin{aligned} L_c(x, \lambda, \mu) = {} & f(x) + \lambda' h(x) + \frac{c}{2}\big\|h(x)\big\|^2 \\ & + \frac{1}{2c}\sum_{j=1}^r \left\{\big(\max\{0, \mu_j + c g_j(x)\}\big)^2 - \mu_j^2\right\}, \end{aligned} \tag{5.61}$$

and we can view it as the augmented Lagrangian function for the inequality constrained problem (5.56).

Note that the penalty term

$$\frac{1}{2c}\big(\max\{0, \mu_j + c g_j(x)\}\big)^2 - \mu_j^2$$

corresponding to the jth inequality constraint in Eq. (5.61) is continuously differentiable in x if g_j is continuously differentiable (see Fig. 5.2.3). However, its Hessian matrix is discontinuous for all x such that $g_j(x) = -\mu_j/c$; this may cause some difficulties in the minimization of $L_c(x, \lambda, \mu)$ and motivates alternative augmented Lagrangian methods for inequality constraints (see Section 5.2.5).

To summarize, the quadratic penalty method for the inequality constrained problem (5.56) consists of a sequence of minimizations of the form

$$\begin{aligned} &\text{minimize} \;\; L_{c^k}(x, \lambda^k, \mu^k) \\ &\text{subject to} \;\; x \in X, \end{aligned}$$

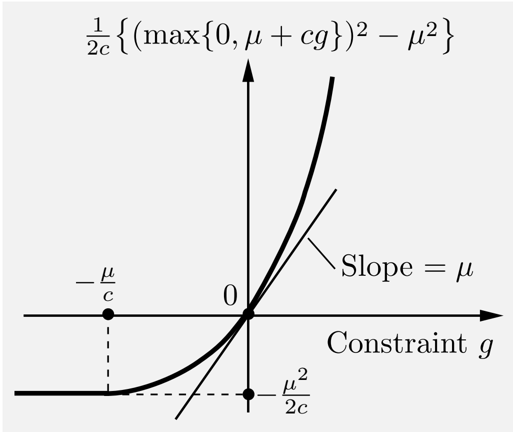

Figure 5.2.3. Form of the quadratic penalty function for inequality constraints.

where $L_c(x, \lambda^k, \mu^k)$ is given by Eq. (5.60) or Eq. (5.61), $\{\lambda^k\}$ and $\{\mu^k\}$ are sequences in $\Re^m$ and $\Re^r$, respectively, and $\{c^k\}$ is a positive penalty parameter sequence. Since this method is equivalent to the equality-constrained method applied to the corresponding equality-constrained problem (5.57), our convergence result of Prop. 5.2.1 applies with the obvious modifications.

Furthermore, if $X = \Re^n$, f, h, and g are continuously differentiable, and the generated sequence $\{x^k\}$ converges to a local minimum x^* which is also regular, application of Prop. 5.2.2 to the equivalent equality constrained problem (5.57) shows that the sequences

$$\big\{\lambda_i^k + c^k h_i(x^k)\big\}, \qquad \max\big\{0, \mu_j^k + c^k g_j(x^k)\big\} \tag{5.62}$$

converge to the corresponding Lagrange multipliers λ_i^* and μ_j^* [for the jth inequality constraint, the Lagrange multiplier estimate is

$$\mu_j^k + c^k g_j^+(x^k, \mu^k, c^k),$$

which is equal to $\max\big\{0, \mu_j^k + c^k g_j(x^k)\big\}$ in view of the form (5.59) for g_j^+].

5.2.2 Multiplier Methods – Main Ideas

Let us return to the case where $X = \Re^n$ and the problem has only equality constraints,

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

We mentioned earlier that optimal solutions of this problem can be well approximated by unconstrained minima of the augmented Lagrangian $L_c(\cdot, \lambda)$ under two types of circumstances:

(a) The vector λ is close to a Lagrange multiplier.

(b) The penalty parameter c is large.

We analyzed in the previous subsection the quadratic penalty method consisting of unconstrained minimization of $L_{c^k}(\cdot, \lambda^k)$ for a sequence $c^k \to \infty$. No assumptions on the sequence $\{\lambda^k\}$ were made other than boundedness. Still, we found that, under minimal assumptions on f and h (continuity), the method has satisfactory convergence properties (Prop. 5.2.1).

We now consider intelligent ways to update λ^k so that it tends to a Lagrange multiplier. We will see that under some reasonable assumptions, this approach is workable even if c^k is not increased to ∞, thereby alleviating much of the difficulty with ill-conditioning. Furthermore, even when c^k is increased to ∞, the rate of convergence is significantly enhanced by using good updating schemes for λ^k.

The Method of Multipliers

A first update formula for λ^k in the quadratic penalty method is

$$\lambda^{k+1} = \lambda^k + c^k h(x^k). \tag{5.63}$$

The rationale is provided by Prop. 5.2.2, which shows that, if the generated sequence $\{x^k\}$ converges to a local minimum x^* that is regular, then $\big\{\lambda^k + c^k h(x^k)\big\}$ converges to the corresponding Lagrange multiplier λ^*.

The quadratic penalty method with the preceding update formula for λ^k is known as the *method of multipliers* (also called *augmented Lagrangian method*). There are a number of interesting convergence and rate of convergence results regarding this method, which will be given shortly. We first illustrate the method with some examples.

Example 5.2.3 (A Convex Problem)

Consider again the problem of Examples 5.2.1 and 5.2.2:

$$\begin{aligned} &\text{minimize} \ \ f(x) = \tfrac{1}{2}(x_1^2 + x_2^2) \\ &\text{subject to} \ \ x_1 = 1, \end{aligned}$$

with optimal solution $x^* = (1, 0)$ and Lagrange multiplier $\lambda^* = -1$. The augmented Lagrangian is

$$L_c(x, \lambda) = \tfrac{1}{2}(x_1^2 + x_2^2) + \lambda(x_1 - 1) + \tfrac{c}{2}(x_1 - 1)^2.$$

The vectors x^k generated by the method of multipliers minimize $L_{c^k}(\cdot, \lambda^k)$ and are given by

$$x^k = \left(\frac{c^k - \lambda^k}{c^k + 1}, 0\right).$$

Using this expression, the multiplier updating formula (5.63) can be written as

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

or by introducing the Lagrange multiplier $\lambda^* = -1$,

$$\lambda^{k+1} - \lambda^* = \frac{\lambda^k - \lambda^*}{c^k + 1}.$$

From this formula, it can be seen that

(a) $\lambda^k \to \lambda^* = -1$ and $x^k \to x^* = (1, 0)$ for every nondecreasing sequence $\{c^k\}$ [since the scalar $1/(c^k+1)$ multiplying $\lambda^k - \lambda^*$ in the above formula is always less than one].

(b) The convergence rate becomes faster as c^k becomes larger; in fact the error sequence $\{|\lambda^k - \lambda^*|\}$ converges superlinearly if $c^k \to \infty$.

Note that it is not necessary to increase c^k to ∞, although doing so results in a better convergence rate.

Example 5.2.4 (A Nonconvex Problem)

Consider the problem

$$\begin{aligned} &\text{minimize} \;\; \tfrac{1}{2}(-x_1^2 + x_2^2) \\ &\text{subject to} \;\; x_1 = 1 \end{aligned}$$

with optimal solution $x^* = (1, 0)$ and Lagrange multiplier $\lambda^* = 1$. The augmented Lagrangian is given by

$$L_c(x, \lambda) = \tfrac{1}{2}(-x_1^2 + x_2^2) + \lambda(x_1 - 1) + \tfrac{c}{2}(x_1 - 1)^2.$$

The vector x^k minimizing $L_{c^k}(x, \lambda^k)$ is given by

$$x^k = \left(\frac{c^k - \lambda^k}{c^k - 1}, 0 \right). \tag{5.64}$$

For this formula to be correct, however, it is necessary that $c^k > 1$; for $c^k < 1$ the augmented Lagrangian has no minimum, and the same is true for $c^k = 1$ unless $\lambda^k = 1$. The multiplier updating formula (5.63) can be written using Eq. (5.64) as

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

or by introducing the Lagrange multiplier $\lambda^* = 1$,

$$\lambda^{k+1} - \lambda^* = -\frac{\lambda^k - \lambda^*}{c^k - 1}. \tag{5.65}$$

From this iteration, it can be seen that similar conclusions to those of the preceding example can be drawn. In particular, it is not necessary to increase c^k to ∞ to obtain convergence, although doing so results in a better convergence rate. However, there is a difference: whereas in the preceding example, convergence was guaranteed for all positive sequences $\{c^k\}$, in the present example, the minimizing points exist only if $c^k > 1$. Here, c^k plays a convexification role: once it exceeds the threshold value of 1 the penalty term convexifies the augmented Lagrangian, thus compensating for the non-convexity of the cost function. Moreover, it is seen from Eq. (5.65) that to obtain convergence, the penalty parameter c^k must eventually exceed 2, so that the scalar

$$\frac{-1}{c^k - 1}$$

multiplying λ^k has absolute value less than one. The need for c^k to exceed twice the value of the convexification threshold is a fundamental characteristic of multiplier methods when applied to nonconvex problems, as we will see shortly.

Geometric Interpretation of the Method of Multipliers

The conclusions from the preceding two examples hold in considerable generality. We first provide a geometric interpretation. Assume that f and h are twice differentiable and let x^* be a local minimum of f over $h(x) = 0$. Assume also that x^* is regular and together with its associated Lagrange multiplier vector λ^* satisfies the second order sufficiency conditions of Prop. 4.2.1. Then the assumptions of the sensitivity theorem (Prop. 4.2.2) are satisfied and we can consider the *primal function*

$$p(u) = \min_{h(x)=u} f(x),$$

defined for u in an open sphere centered at $u = 0$. [The minimization above is understood to be local in an open sphere within which x^* is the unique local minimum of f over $h(x) = 0$ (cf. Prop. 4.2.2).] Note that we have

$$p(0) = f(x^*), \qquad \nabla p(0) = -\lambda^*,$$

(cf. Prop. 4.2.2). The primal function is illustrated in Fig. 5.2.4.

We can break down the minimization of $L_c(\cdot, \lambda)$ into two stages, first minimizing over all x such that $h(x) = u$ with u fixed, and then minimizing over all u. Thus,

$$\begin{aligned} \min_x L_c(x, \lambda) &= \min_u \min_{h(x)=u} \left\{ f(x) + \lambda' h(x) + \frac{c}{2} \|h(x)\|^2 \right\} \\ &= \min_u \left\{ p(u) + \lambda' u + \frac{c}{2} \|u\|^2 \right\}, \end{aligned}$$

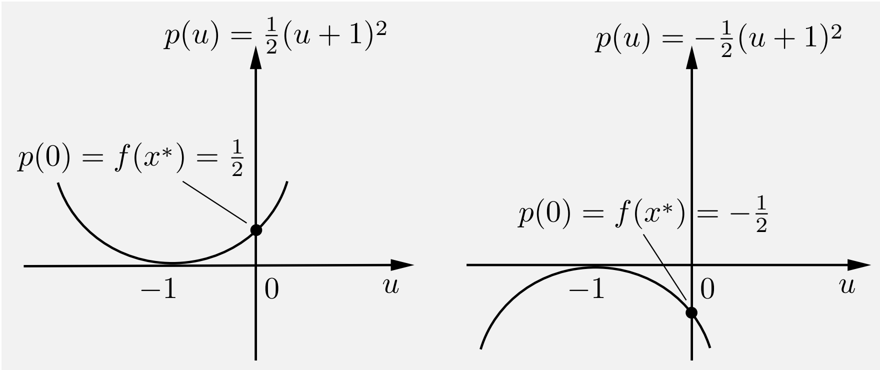

Figure 5.2.4. Illustration of the primal function. In (a) we show the primal function

$$p(u) = \min_{x_1 - 1 = u} \tfrac{1}{2}(x_1^2 + x_2^2)$$

for the problem of Example 5.2.3. In (b) we show the primal function

$$p(u) = \min_{x_1 - 1 = u} \tfrac{1}{2}(-x_1^2 + x_2^2)$$

for the problem of Example 5.2.4. The latter primal function is not convex because the cost function is not convex on the subspace that is orthogonal to the constraint set (this observation can be generalized).

where the minimization above is understood to be local in a neighborhood of $u = 0$. This minimization can be interpreted as shown in Fig. 5.2.5. The minimum is attained at the point $u(\lambda, c)$ for which the gradient of $p(u) + \lambda' u + \frac{c}{2}\|u\|^2$ is zero, or, equivalently,

$$\nabla \left\{ p(u) + \frac{c}{2}\|u\|^2 \right\} \Big|_{u=u(\lambda,c)} = -\lambda.$$

Thus, the minimizing point $u(\lambda, c)$ is obtained as shown in Fig. 5.2.5. We also have

$$\min_x L_c(x, \lambda) - \lambda' u(\lambda, c) = p\big(u(\lambda, c)\big) + \frac{c}{2}\big\|u(\lambda, c)\big\|^2,$$

so the tangent hyperplane to the graph of $p(u) + \frac{c}{2}\|u\|^2$ at $u(\lambda, c)$ (which has "slope" $-\lambda$) intersects the vertical axis at the value $\min_x L_c(x, \lambda)$ as shown in Fig. 5.2.5. It can be seen that if c is sufficiently large, then the function

$$p(u) + \lambda' u + \frac{c}{2}\|u\|^2$$

is convex in a neighborhood of the origin. Furthermore, for λ close to λ^* and large c, the value $\min_x L_c(x, \lambda)$ is close to $p(0) = f(x^*)$.

Figure 5.2.6 provides a geometric interpretation of the multiplier iteration

$$\lambda^{k+1} = \lambda^k + c^k h(x^k).$$

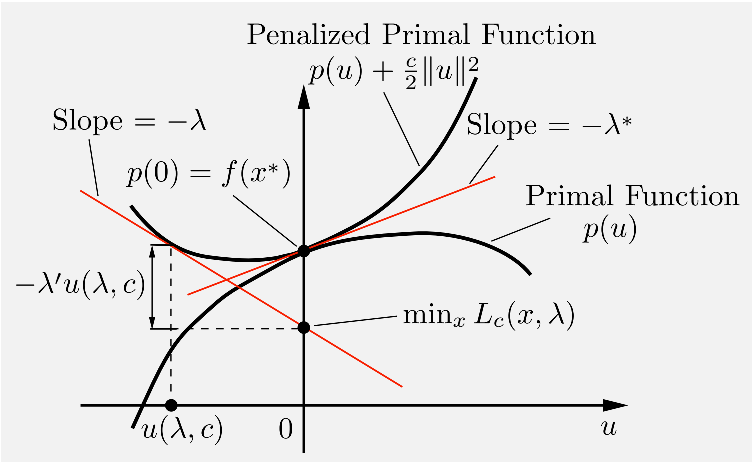

Figure 5.2.5. Geometric interpretation of minimization of the augmented Lagrangian. The value $\min_x L_c(x,\lambda)$ is the point where the tangent hyperplane to the graph of $p(u)+\frac{c}{2}\|u\|^2$ at $u(\lambda,c)$ (which has "slope" $-\lambda$) intersects the vertical axis. This point is close to $p(0)=f(x^*)$ if either λ is close to λ^* or c is large (or both).

To understand this figure, note that if x^k minimizes $L_{c^k}(\cdot,\lambda^k)$, then by the preceding analysis the vector u^k given by $u^k = h(x^k)$ minimizes $p(u) + {\lambda^k}' u + \frac{c^k}{2}\|u\|^2$. Hence,

$$\nabla\left\{p(u)+\frac{c^k}{2}\|u\|^2\right\}\Big|_{u=u^k} = -\lambda^k,$$

and

$$\nabla p(u^k) = -(\lambda^k + c^k u^k) = -\big(\lambda^k + c^k h(x^k)\big).$$

It follows that the next multiplier λ^{k+1} is

$$\lambda^{k+1} = \lambda^k + c^k h(x^k) = -\nabla p(u^k),$$

as shown in Fig. 5.2.6. The figure shows that if λ^k is sufficiently close to λ^* and/or c^k is sufficiently large, the next multiplier λ^{k+1} will be closer to λ^* than λ^k is. Furthermore, c^k need not be increased to ∞ in order to obtain convergence; it is sufficient that c^k eventually exceeds some threshold level. The figure also shows that if $p(u)$ is linear, convergence to λ^* will be achieved in one iteration.

In summary, the geometric interpretation of the method of multipliers just presented suggests the following:

(a) If c is large enough so that $cI + \nabla^2 p(0)$ is positive definite, then the "penalized primal function"

$$p(u)+\frac{c}{2}\|u\|^2$$

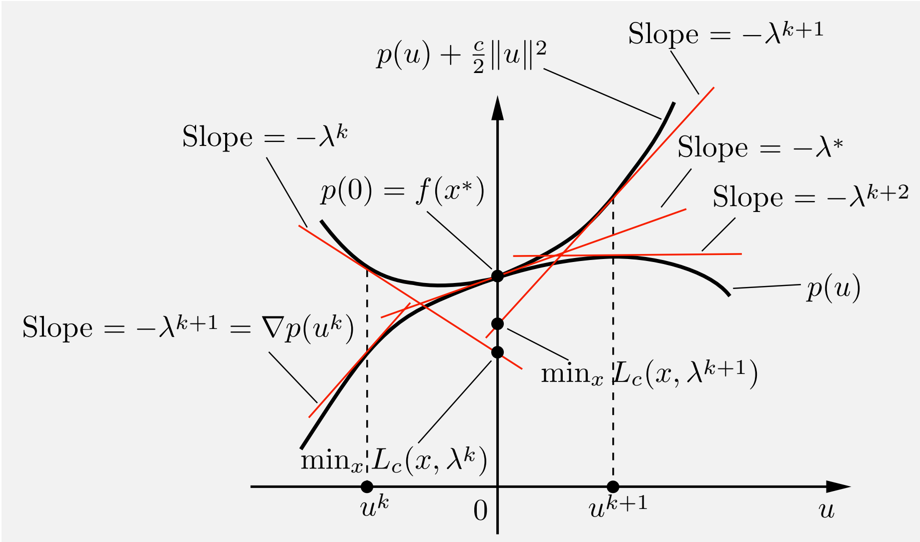

Figure 5.2.6. Geometric interpretation of the first order multiplier iteration. The figure shows the process of obtaining λ^{k+1} from λ^k, assuming a constant penalty parameter $c^k = c^{k+1} = c$.

is convex within a sphere centered at $u = 0$. Furthermore, a local minimum of the augmented Lagrangian $L_c(x, \lambda)$ that is near x^* exists if λ is close enough to λ^*. The reason is that $\nabla^2_{xx} L_c(x^*, \lambda^*)$ is positive definite if and only if $cI + \nabla^2 p(0)$ is positive definite, a fact the reader may wish to verify as an exercise.

(b) If c^k is sufficiently large [the threshold can be shown to be twice the value of c needed to make $cI + \nabla^2 p(0)$ positive definite; see Exercise 5.2.4], then

$$\|\lambda^{k+1} - \lambda^*\| \leq \|\lambda^k - \lambda^*\|$$

and $\lambda^k \to \lambda^*$.

(c) Convergence can be obtained even if c^k is not increased to ∞.

(d) As c^k is increased, the rate of convergence of λ^k improves.

(e) If $\nabla^2 p(0) = 0$, the convergence is very fast.

These conclusions will be formalized in Section 5.2.3.

Computational Aspects – Choice of Parameters

In addition to addressing the problem of ill-conditioning, an important practical question in the method of multipliers is how to select the initial multiplier λ^0 and the penalty parameter sequence. Clearly, in view of the interpretations given earlier, any prior knowledge should be exploited to select λ^0 as close as possible to λ^*. The main considerations to be kept in mind for selecting the penalty parameter sequence are the following:

(a) c^k should eventually become larger than the threshold level necessary to bring to bear the positive features of the multiplier iteration.

(b) The initial parameter c^0 should not be too large to the point where it causes ill-conditioning at the first unconstrained minimization.

(c) c^k should not be increased too fast to the point where too much ill-conditioning is forced upon the unconstrained minimization routine too early.

(d) c^k should not be increased too slowly, at least in the early minimizations, to the extent that the multiplier iteration has poor convergence rate.

A good practical scheme is to choose a moderate value c^0 (if necessary by preliminary experimentation), and then increase c^k via the equation $c^{k+1} = \beta c^k$, where β is a scalar with $\beta > 1$. In this way, the threshold level for multiplier convergence will eventually be exceeded. If the method used for augmented Lagrangian minimization is powerful, such as a Newton-like method, fairly large values of β (say $\beta \in [5, 10]$) are recommended; otherwise, smaller values of β may be necessary, depending on the method's ability to deal with ill-conditioning.

Another reasonable parameter adjustment scheme is to increase c^k by multiplication with a factor $\beta > 1$ only if the constraint violation as measured by $\big\|h\big(x(\lambda^k, c^k)\big)\big\|$ is not decreased by a factor $\gamma < 1$ over the previous minimization; i.e.,

$$c^{k+1} = \begin{cases} \beta c^k & \text{if } \big\|h(x^k)\big\| > \gamma \big\|h(x^{k-1})\big\|, \\ c^k & \text{if } \big\|h(x^k)\big\| \le \gamma \big\|h(x^{k-1})\big\|. \end{cases}$$

A choice such as $\gamma = 0.25$ is typically recommended.

Still another possibility is an adaptive scheme that uses a different penalty parameter c_i^k for each constraint $h_i(x) = 0$, and increases by a certain factor the penalty parameters of the constraints that are violated most. For example, increase c_i^k if the constraint violation as measured by $\big|h_i(x^k)\big|$ is not decreased by a certain factor over $\big|h_i(x^{k-1})\big|$.

Inexact Minimization of the Augmented Lagrangian

In practice the minimization of $L_{c^k}(x, \lambda^k)$ is typically terminated early. For example, it may be terminated at a point x^k satisfying

$$\big\|\nabla_x L_{c^k}(x^k, \lambda^k)\big\| \le \epsilon^k,$$

where $\{\epsilon^k\}$ is a positive sequence converging to zero. Then it is still appropriate to use the multiplier update

$$\lambda^{k+1} = \lambda^k + c^k h(x^k),$$

although in theory, some of the linear convergence rate results to be given shortly will not hold any more. This deficiency does not seem to be important in practice, but can also be corrected by using the alternative termination criterion

$$\big\|\nabla_x L_{c^k}(x^k, \lambda^k)\big\| \le \min\big\{\epsilon^k, \gamma^k \|h(x^k)\|\big\},$$

where $\{\epsilon^k\}$ and $\{\gamma^k\}$ are positive sequences converging to zero; for an analysis see the author's works [Ber75b], [Ber76a], and [Ber82a], Section 2.5.

In Section 5.4, we will see that with certain safeguards, it is possible to terminate the minimization of the augmented Lagrangian after a few Newton steps (possibly only one), and follow it by a second order multiplier update of the type that will be discussed later in this section. Such algorithmic strategies give rise to some of the most effective methods using Lagrange multipliers.

Inequality Constraints

To treat inequality constraints $g_j(x) \le 0$ in the context of the method of multipliers, we convert them into equality constraints $g_j(x) + z_j^2 = 0$, using the additional variables z_j [cf. problems (5.56) and (5.57)]. In particular, the multiplier update formulas are

$$\lambda^{k+1} = \lambda^k + c^k h(x^k),$$

$$\mu_j^{k+1} = \max\big\{0, \mu_j^k + c^k g_j(x^k)\big\}, \qquad j = 1, \ldots, r,$$

[cf. Eq. (5.62)], where x^k minimizes the augmented Lagrangian

$$\begin{aligned} L_{c^k}(x, \lambda^k, \mu^k) = f(x) + \lambda^{k\prime} h(x) + \frac{c}{2}\big\|h(x)\big\|^2 \\ + \frac{1}{2c^k} \sum_{j=1}^{r} \Big\{ \big(\max\{0, \mu_j^k + c^k g_j(x)\}\big)^2 - (\mu_j^k)^2 \Big\}. \end{aligned}$$

Many problems encountered in practice involve two-sided constraints of the form

$$\alpha_j \le g_j(x) \le \beta_j,$$

where α_j and β_j are given scalars. Each two-sided constraint could of course be separated into two one-sided constraints. This would require, however, the assignment of two multipliers per two-sided constraint, and is somewhat wasteful, since we know that at a solution at least one of the two multipliers will be zero. It turns out that there is an alternative approach that requires only *one multiplier per two-sided constraint* (see Exercise 5.2.7).

Partial Elimination of Constraints

In the preceding multiplier algorithms, all the equality and inequality constraints are eliminated by means of a penalty. In some cases it is convenient to eliminate only part of the constraints, while retaining the remaining constraints explicitly. A typical example is a problem of the form

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

While, in addition to $h(x) = 0$, it is possible to eliminate by means of a penalty the bound constraints $x \geq 0$, it is often desirable to handle these constraints explicitly by a gradient projection or two-metric projection method of the type discussed in Sections 3.3 and 3.4, respectively.

More generally, a method of multipliers with partial elimination of constraints for the problem

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

consists of finding x^k that solves the problem

$$\begin{aligned} &\text{minimize} \;\; f(x) + \lambda^{k\prime} h(x) + \frac{c}{2}\big\|h(x)\big\|^2 \\ &\text{subject to} \;\; g(x) \leq 0, \end{aligned}$$

followed by the multiplier iteration

$$\lambda^{k+1} = \lambda^k + c^k h(x^k).$$

In fact it is not essential that just the equality constraints are eliminated by means of a penalty above. Any mixture of equality and inequality constraints can be eliminated by means of a penalty and a multiplier, while the remaining constraints can be explicitly retained. For a detailed treatment of partial elimination of constraints, we refer to [Ber77], [Ber82a], Section 2.4, [Dun91a], and [Dun93b].

5.2.3 Convergence Analysis of Multiplier Methods

We now discuss the convergence properties of multiplier methods and substantiate the conclusions derived informally earlier. We focus attention throughout on the equality constrained problem

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

and on a particular local minimum x^*. We assume that x^* *is regular and together with a Lagrange multiplier vector* λ^* *satisfies the second order sufficiency conditions of Prop. 4.2.1*. In view of our earlier treatment of inequality constraints by conversion to equalities, our analysis readily carries over to the case of mixed equality and inequality constraints, under the second order sufficiency conditions of Prop. 4.2.1.

The convergence results described in this section can be strengthened considerably under additional convexity assumptions on the problem. This is discussed further in Section 7.3; see also [Ber82a], Chapter 5.

Error Bounds for Local Minima of the Augmented Lagrangian

A first basic issue is whether local minima x^k of the augmented Lagrangian $L_{c^k}(\cdot, \lambda^k)$ exist, so that the method itself is well-defined. We have shown that for the local minimum-Lagrange multiplier pair (x^*, λ^*) there exist scalars $\bar{c} > 0$, $\gamma > 0$, and $\epsilon > 0$, such that

$$L_c(x, \lambda^*) \geq L_c(x^*, \lambda^*) + \frac{\gamma}{2}\|x - x^*\|^2, \qquad \forall\, x \text{ with } \|x - x^*\| < \epsilon, \text{ and } c \geq \bar{c},$$

(cf. the discussion following Lemma 3.2.1). It is thus reasonable to infer that if λ is close to λ^*, there should exist a local minimum of $L_c(\cdot, \lambda)$ close to x^* for every $c \geq \bar{c}$. More precisely, for a fixed $c \geq \bar{c}$, we can show this by considering the system of equations

$$\nabla_x L_c(x, \lambda) = \nabla f(x) + \nabla h(x)\big(\lambda + ch(x)\big) = 0,$$

and by using the implicit function theorem in a neighborhood of (x^*, λ^*). [This can be done because the Jacobian of the system with respect to x is $\nabla_x^2 L_c(x^*, \lambda^*)$, and is positive definite since $c \geq \bar{c}$.] Thus, for λ sufficiently close to λ^*, there is an unconstrained local minimum $x(\lambda, c)$ of $L_c(\cdot, \lambda)$, which is defined via the equation

$$\nabla f\big(x(\lambda, c)\big) + \nabla h\big(x(\lambda, c)\big)\big(\lambda + ch(x(\lambda, c))\big) = 0.$$

A closer examination of the preceding argument shows that for application of the implicit function theorem it is not essential that λ be close to λ^* but rather that the vector $\lambda + ch\big(x(\lambda, c)\big)$ be close to λ^*. Proposition 5.2.2 indicates that for any λ, if c is sufficiently large and $x(\lambda, c)$ minimizes $L_c(x, \lambda)$, the vector $\lambda + ch\big(x(\lambda, c)\big)$ is close to λ^*. This suggests that *there should exist a local minimum of* $L_c(\cdot, \lambda)$ *close to* x^* *even for* λ *that are far from* λ^*, *provided* c *is sufficiently large*. This can indeed be shown. In fact it turns out that for existence of the local minimum $x(\lambda, c)$, what is really important is that the ratio $\|\lambda - \lambda^*\|/c$ be sufficiently small. However, proving this simultaneously for the entire range of values $c \in [\bar{c}, \infty)$ is not easy. The following proposition, due to [Ber82a], provides a precise mathematical statement of the existence result, together with some error estimates that quantify the rate of convergence. The proof requires the introduction of the variables $\tilde{\lambda} = \lambda + ch(x)$, and $t = (\lambda - \lambda^*)/c$, together with the system of equations $\nabla_x L_0(x, \tilde{\lambda}) = 0$ and an analysis based on a more refined form of the implicit function theorem than the one given in Appendix A.

Proposition 5.2.3: Let $\bar{c}$ be a positive scalar such that

$$\nabla^2_{xx} L_{\bar{c}}(x^*, \lambda^*) > 0.$$

There exist positive scalars δ, ϵ, and M such that:

(a) For all (λ, c) in the set $D \subset \Re^{m+1}$ defined by

$$D = \big\{(\lambda, c) \mid \|\lambda - \lambda^*\| < \delta c,\ \bar{c} \le c\big\}, \tag{5.66}$$

the problem

$$\begin{aligned} &\text{minimize} \quad L_c(x, \lambda) \\ &\text{subject to} \quad \|x - x^*\| < \epsilon \end{aligned}$$

has a unique solution denoted $x(\lambda, c)$. The function $x(\cdot, \cdot)$ is continuously differentiable in the interior of D, and for all $(\lambda, c) \in D$, we have

$$\big\|x(\lambda, c) - x^*\big\| \le M \frac{\|\lambda - \lambda^*\|}{c}.$$

(b) For all $(\lambda, c) \in D$, we have

$$\big\|\tilde{\lambda}(\lambda, c) - \lambda^*\big\| \le M \frac{\|\lambda - \lambda^*\|}{c},$$

where

$$\tilde{\lambda}(\lambda, c) = \lambda + ch\big(x(\lambda, c)\big).$$

(c) For all $(\lambda, c) \in D$, the matrix $\nabla^2_{xx} L_c\big(x(\lambda, c), \lambda\big)$ is positive definite and the matrix $\nabla h\big(x(\lambda, c)\big)$ has rank m.

Proof: See [Ber82a], p. 108.

Figure 5.2.7 shows the set D of pairs (λ, c) within which the conclusions of Prop. 5.2.3 are valid [cf. Eq. (5.66)]. It can be seen that, for any λ, there exists a c_λ such that (λ, c) belongs to D for every $c \ge c_\lambda$. The estimate δc on the allowable distance of λ from λ^* grows linearly with c

[compare with Eq. (5.66)]. In particular problems, the actual allowable distance may grow at a higher than linear rate, and in fact it is possible that for every λ and $c > 0$ there exists a unique global minimum of $L_c(\cdot, \lambda)$. (Take for instance the scalar problem $\min\{x^2 \mid x = 0\}$.) However, it is shown by example in [Ber82a], p. 111, that the estimate of a linear order of growth cannot be improved.

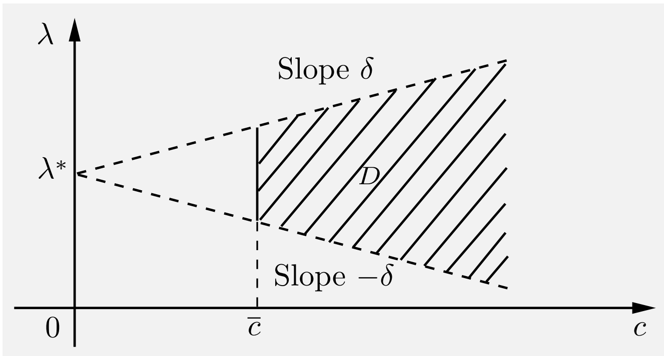

Figure 5.2.7. Illustration of the set

$$D = \left\{(\lambda, c) \mid \|\lambda - \lambda^*\| < \delta c,\, \bar{c} \leq c\right\}$$

within which the conclusions of Prop. 5.2.3 are valid.

Convergence and Rate of Convergence

Proposition 5.2.3 yields both a convergence and a convergence rate result for the multiplier iteration

$$\lambda^{k+1} = \lambda^k + c^k h(x^k).$$

It shows that if the generated sequence $\{\lambda^k\}$ is bounded [this can be enforced if necessary by leaving λ^k unchanged if $\lambda^k + c^k h(x^k)$ does not belong to a prespecified bounded open set known to contain λ^*], the penalty parameter c^k is sufficiently large after a certain index [so that $(\lambda^k, c^k) \in D$], and after that index, minimization of $L_{c^k}(\cdot, \lambda^k)$ yields the local minimum $x^k = x(\lambda^k, c^k)$ closest to x^*, then we obtain $x^k \to x^*$, $\lambda^k \to \lambda^*$. Furthermore, the rate of convergence of the error sequences $\left\{\|x^k - x^*\|\right\}$ and $\left\{\|\lambda^k - \lambda^*\|\right\}$ is linear, and it is superlinear if $c^k \to \infty$.

It is possible to conduct a more refined convergence and rate of convergence analysis that supplements Prop. 5.2.3. This analysis quantifies the threshold level of the penalty parameter for convergence to occur and gives a precise estimate of the linear convergence rate. We refer to the book [Ber82a] for an extensive discussion; see also Exercise 5.2.4.

5.2.4 Duality and Second Order Multiplier Methods

Let $\bar{c}$, δ, and ϵ be as in Prop. 5.2.3, and define for (λ, c) in the set

$$D = \big\{(\lambda, c) \mid \|\lambda - \lambda^*\| < \delta c,\ \bar{c} \le c\big\}$$

the *dual function* q_c by

$$q_c(\lambda) = \min_{\|x - x^*\| < \epsilon} L_c(x, \lambda) = L_c\big(x(\lambda, c), \lambda\big). \tag{5.67}$$

Since $x(\cdot, c)$ is continuously differentiable (Prop. 5.2.3), the same is true for q_c.

Calling q_c a dual function is not inconsistent with the duality theory already formulated in Section 4.4 and further developed in Chapter 6. Indeed q_c is the dual function for the problem

$$\begin{aligned} &\text{minimize} \ \ f(x) + \frac{c}{2}\big\|h(x)\big\|^2 \\ &\text{subject to} \ \ \|x - x^*\| < \epsilon, \quad h(x) = 0, \end{aligned}$$

which for $c \ge \bar{c}$, has x^* as its unique optimal solution and λ^* as the corresponding Lagrange multiplier.

We compute the gradient of q_c with respect to λ. From Eq. (5.67), we have

$$\nabla q_c(\lambda) = \nabla_\lambda x(\lambda, c) \nabla_x L_c\big(x(\lambda, c), \lambda\big) + h\big(x(\lambda, c)\big).$$

Since $\nabla_x L_c\big(x(\lambda, c), \lambda\big) = 0$, we obtain

$$\nabla q_c(\lambda) = h\big(x(\lambda, c)\big), \tag{5.68}$$

and since $x(\cdot, c)$ is continuously differentiable, the same is true for ∇q_c.

Next we compute the Hessian $\nabla^2 q_c$. Differentiating ∇q_c, as given by Eq. (5.68), with respect to λ, we obtain

$$\nabla^2 q_c(\lambda) = \nabla_\lambda x(\lambda, c) \nabla h\big(x(\lambda, c)\big). \tag{5.69}$$

We also have, for all (λ, c) in the set D,

$$\nabla_x L_c\big(x(\lambda, c), \lambda\big) = 0.$$

Differentiating with respect to λ, we obtain

$$\nabla_\lambda x(\lambda, c) \nabla^2_{xx} L_c\big(x(\lambda, c), \lambda\big) + \nabla^2_{\lambda x} L_c\big(x(\lambda, c), \lambda\big) = 0,$$

and since

$$\nabla^2_{\lambda x} L_c\big(x(\lambda, c), \lambda\big) = \nabla h\big(x(\lambda, c)\big)',$$

it follows that

$$\nabla_\lambda x(\lambda, c) = -\nabla h\big(x(\lambda, c)\big)'\big(\nabla^2_{xx} L_c\big(x(\lambda, c), \lambda\big)\big)^{-1}.$$

Substitution in Eq. (5.69) yields the formula

$$\nabla^2 q_c(\lambda) = -\nabla h\big(x(\lambda, c)\big)'\big(\nabla^2_{xx} L_c\big(x(\lambda, c), \lambda\big)\big)^{-1}\nabla h\big(x(\lambda, c)\big). \tag{5.70}$$

Since $\nabla^2_{xx} L_c\big(x(\lambda, c), \lambda\big)$ is positive definite and $\nabla h\big(x(\lambda, c)\big)$ has rank m for $(\lambda, c) \in D$ (cf. Prop. 5.2.3), it follows from Eq. (5.70) that $\nabla^2 q_c(\lambda)$ is negative definite for all $(\lambda, c) \in D$, so that q_c is concave within the set $\big\{\lambda \mid \|\lambda - \lambda^*\| \le \delta c\big\}$. Furthermore, using Eq. (5.68), we have, for all $c \ge \bar{c}$,

$$\nabla q_c(\lambda^*) = h\big(x(\lambda^*, c)\big) = h(x^*) = 0.$$

Thus, *for every $c \ge \bar{c}$, λ^* maximizes $q_c(\lambda)$ over the set $\big\{\lambda \mid \|\lambda - \lambda^*\| < \delta c\big\}$.* Also in view of Eq. (5.68), the multiplier update formula can be written as

$$\lambda^{k+1} = \lambda^k + c^k \nabla q_{c^k}(\lambda^k), \tag{5.71}$$

so *it is a steepest ascent iteration for maximizing q_{c^k}*. When $c^k = c$ for all k, then Eq. (5.71) is the constant stepsize steepest ascent method

$$\lambda^{k+1} = \lambda^k + c\nabla q_c(\lambda^k)$$

for maximizing q_c.

The Second Order Method of Multipliers

In view of the interpretation of the multiplier iteration as a steepest ascent method, it is natural to consider the Newton-like iteration

$$\lambda^{k+1} = \lambda^k - \big(\nabla^2 q_{c^k}(\lambda^k)\big)^{-1}\nabla q_{c^k}(\lambda^k),$$

for maximizing the dual function. In view of the gradient and Hessian formulas (5.68) and (5.70), this iteration can be written as

$$\lambda^{k+1} = \lambda^k + (B^k)^{-1} h(x^k),$$

where

$$B^k = \nabla h\big(x^k\big)'\big(\nabla^2_{xx} L_{c^k}(x^k, \lambda^k)\big)^{-1}\nabla h\big(x^k\big)$$

and x^k minimizes $L_{c^k}(\cdot, \lambda^k)$.

An alternative form, which turns out to be more appropriate when the minimization of the augmented Lagrangian is inexact is given by

$$\lambda^{k+1} = \lambda^k + (B^k)^{-1}\big(h(x^k) - \nabla h(x^k)'\big(\nabla^2_{xx} L_{c^k}(x^k, \lambda^k)\big)^{-1}\nabla_x L_{c^k}(x^k, \lambda^k)\big). \tag{5.72}$$

When the augmented Lagrangian is minimized exactly, $\nabla_x L_{c^k}(x^k, \lambda^k) = 0$, and the two forms are equivalent.

To provide motivation for iteration (5.72), let us consider Newton's method for solving the system of necessary conditions

$$\nabla_x L_c(x, \lambda) = \nabla f(x) + \nabla h(x)\big(\lambda + ch(x)\big) = 0, \qquad h(x) = 0.$$

In this method, we linearize the above system around the current iterate (x^k, λ^k), and we obtain the next iterate (x^{k+1}, λ^{k+1}) from the solution of the linearized system

$$\begin{pmatrix} \nabla^2_{xx} L_{c^k}(x^k, \lambda^k) & \nabla h(x^k) \\ \nabla h(x^k)' & 0 \end{pmatrix} \begin{pmatrix} x^{k+1} - x^k \\ \lambda^{k+1} - \lambda^k \end{pmatrix} = -\begin{pmatrix} \nabla_x L_{c^k}(x^k, \lambda^k) \\ h(x^k) \end{pmatrix}.$$

It is straightforward to verify (a derivation will be given in Section 5.4.2) that λ^{k+1} is given by Eq. (5.72), while

$$x^{k+1} = x^k - \big(\nabla^2_{xx} L_{c^k}(x^k, \lambda^k)\big)^{-1} \nabla_x L_{c^k}\big(x^k, \lambda^{k+1}\big).$$

This justifies the use of the extra term

$$\nabla h(x^k)' \big(\nabla^2_{xx} L_{c^k}(x^k, \lambda^k)\big)^{-1} \nabla_x L_{c^k}(x^k, \lambda^k)$$

in Eq. (5.72) when the minimization of the augmented Lagrangian is inexact.

5.2.5 Nonquadratic Augmented Lagrangians - The Exponential Method of Multipliers

One of the drawbacks of the method of multipliers when applied to inequality constrained problems is that the corresponding augmented Lagrangian function is not twice differentiable even if the cost and constraint functions are. As a result, serious difficulties can arise when Newton-like methods are used to minimize the augmented Lagrangian, particularly for polyhedral-type problems. This motivates alternative twice differentiable augmented Lagrangians to handle inequality constraints, which we now describe.

Consider the problem

$$\begin{aligned} &\text{minimize} \quad f(x) \\ &\text{subject to} \quad g_1(x) \le 0, \ldots, g_r(x) \le 0. \end{aligned}$$

We introduce a method of multipliers characterized by a twice differentiable penalty function $\psi : \Re \mapsto \Re$ with the following properties:

(i) $\nabla^2 \psi(t) > 0$ for all $t \in \Re$,

(ii) $\psi(0) = 0$, $\nabla \psi(0) = 1$,

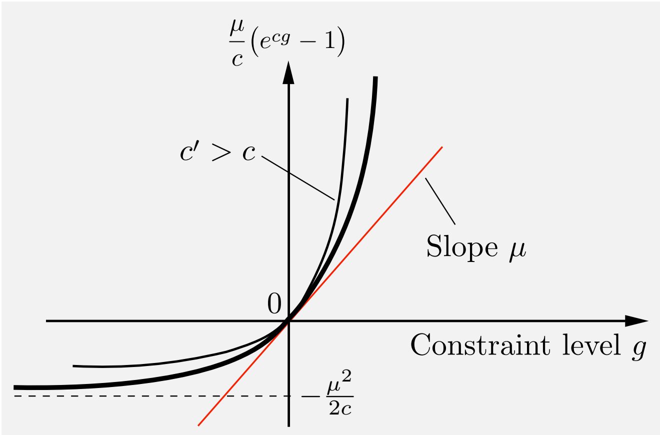

Figure 5.2.8. The penalty term of the exponential method of multipliers. The slope at 0 is μ, regardless of the value of x.

(iii) $\lim_{t\to-\infty}\psi(t) > -\infty$,

(iv) $\lim_{t\to-\infty}\nabla\psi(t) = 0$ and $\lim_{t\to\infty}\nabla\psi(t) = \infty$.

A simple and interesting special case is the *exponential penalty function*

$$\psi(t) = e^t - 1,$$

(see Fig. 5.2.8).

The method consists of the sequence of unconstrained minimizations

$$x^k \in \arg\min_{x\in\Re^n}\left\{f(x) + \sum_{j=1}^m \frac{\mu_j^k}{c_j^k}\psi\big(c_j^k g_j(x)\big)\right\},$$

followed by the multiplier iterations

$$\mu_j^{k+1} = \mu_j^k \nabla\psi\big(c_j^k g_j(x^k)\big), \qquad j = 1,\ldots,r. \tag{5.73}$$

Here $\{c_j^k\}$ is a positive penalty parameter sequence for each j, and the initial multipliers μ_j^0 are arbitrary positive numbers.

Note that for fixed $\mu_j^k > 0$, the "penalty" term

$$\frac{\mu_j^k}{c_j^k}\psi\big(c_j^k g_j(x)\big)$$

tends (as $c_j^k \to \infty$) to ∞ for all infeasible x $[g_j(x) > 0]$ and to zero for all feasible x $[g_j(x) \le 0]$. To see this, note that by convexity of ψ, we have $\psi(ct) \ge \psi(ct/2) + (ct/2)\nabla\psi(ct/2)$, from which we obtain

$$\frac{1}{c}\psi(ct) \ge \frac{1}{c}\psi(ct/2) + \frac{t}{2}\nabla\psi(ct/2).$$

Thus if $t > 0$, the assumptions $\psi(ct/2) > 0$ and $\lim_{\tau\to\infty} \nabla\psi(\tau) = \infty$ imply that $\lim_{c\to\infty}(1/c)\psi(ct) = \infty$. Also, if $t < 0$, we have $\inf_{c>0} \psi(ct) = \inf_{\tau<0} \psi(\tau) > -\infty$, so that $\lim_{c\to\infty}(1/c)\psi(ct) = 0$.

On the other hand, for fixed c_j^k, as $\mu_j^k \to 0$ (which is expected to occur if the jth constraint is inactive at the optimum), the penalty term goes to zero for all x, feasible or infeasible. This is contrary to what happens in the quadratic penalty and augmented Lagrangian methods, and turns out to be a complicating factor in the analysis.

For the exponential penalty function $\psi(t) = e^t - 1$, the multiplier iteration (5.73) takes the form

$$\mu_j^{k+1} = \mu_j^k e^{c_j^k g_j(x^k)}, \qquad j = 1, \ldots, r.$$

Another interesting method, known as the *modified barrier method*, is based on the following version of the logarithmic barrier function

$$\psi(t) = -\ln(1-t),$$

for which the multiplier iteration (5.73) takes the form

$$\mu_j^{k+1} = \frac{\mu_j^k}{1 - c_j^k g_j(x^k)}, \qquad j = 1, \ldots, r.$$

This method is not really a special case of the generic method (5.73) because the penalty function ψ is defined only on the set $(-\infty, 1)$, but it shares the same qualitative characteristics as the generic method.

Practical Implementation

Two practical points regarding the exponential method are worth mentioning. The first is that the exponential terms in the augmented Lagrangian function can easily become very large with an attendant computer overflow. [The modified barrier method has a similar and even more serious difficulty: it tends to ∞ as $c_j^k g_j(x^k) \to 1$.] One way to deal with this disadvantage is to define $\psi(t)$ as the exponential $e^t - 1$ only for t in an interval $(-\infty, A]$, where A is such that e^A is within the floating point range of the computer; outside the interval $(-\infty, A]$, $\psi(t)$ can be defined as any function such that the properties required of ψ, including twice differentiability, are maintained over the entire real line. For example ψ can be a quadratic function with parameters chosen so that $\nabla^2\psi$ is continuous at the splice point A.

The second point is that it makes sense to introduce a different penalty parameter c_j^k for the jth constraint and to let c_j^k depend on the current values of the corresponding multiplier μ_j^k via

$$c_j^k = \frac{w^k}{\mu_j^k}, \qquad j = 1, \ldots, r, \tag{5.74}$$

where $\{w^k\}$ is a positive scalar sequence with $w^k \leq w^{k+1}$ for all k. The reason can be seen by using the series expansion of the exponential term to write

$$\frac{\mu_j}{c}\left(e^{cg_j(x)} - 1\right) = \frac{\mu_j}{c}\left(cg_j(x) + \frac{c^2}{2}\big(g_j(x)\big)^2 + \frac{c^3}{3!}\big(g_j(x)\big)^3 + \cdots\right).$$

If the jth constraint is active at the eventual limit, the terms of order higher than quadratic can be neglected and we can write approximately

$$\frac{\mu_j}{c}\left(e^{cg_j(x)} - 1\right) \approx \mu_j g_j(x) + \frac{c\mu_j}{2} g_j(x)^2.$$

Thus the exponential augmented Lagrangian term becomes similar to the quadratic term, except that the role of the penalty parameter is played by the product $c\mu_j$. This motivates the use of selection rules such as Eq. (5.74). A similar penalty selection rationale applies also to other penalty functions in the class.

Generally the convergence analysis of the exponential method of multipliers and other methods in the class, under second order sufficiency conditions, turns out to be not much more difficult than for the quadratic method (see the references). The convergence results available are as powerful as for the quadratic method. The practical performances of the exponential and the quadratic method of multipliers are roughly comparable for problems where second order differentiability of the augmented Lagrangian function turns out to be of no concern. The exponential method has an edge for problems where the lack of second order differentiability in the quadratic method causes difficulties.

EXERCISES

5.2.1

Consider the problem

$$\begin{aligned} &\text{minimize} \ \ f(x) = \tfrac{1}{2}\left(x_1^2 - x_2^2\right) - 3x_2 \\ &\text{subject to} \ \ x_2 = 0. \end{aligned}$$

(a) Calculate the optimal solution and the Lagrange multiplier.

(b) For $k = 0, 1, 2$ and $c^k = 10^{k+1}$ calculate and compare the iterates of the quadratic penalty method with $\lambda^k = 0$ for all k and the method of multipliers with $\lambda^0 = 0$.

(c) Draw the figure that interprets geometrically the two methods (cf. Figs. 5.2.5 and 5.2.6) for this problem, and plot the iterates of the two methods on this figure for $k = 0, 1, 2$.

(d) Suppose that c is taken to be constant in the method of multipliers. For what values of c would the augmented Lagrangian have a minimum and for what values of c would the method converge?

5.2.2

Consider the problem

$$\begin{aligned} &\text{minimize} \;\; f(x) = \tfrac{1}{2}\big(x_1^2 + |x_2|^\rho\big) + 2x_2 \\ &\text{subject to} \;\; x_2 = 0, \end{aligned}$$

where $\rho > 1$.

(a) Calculate the optimal solution and the Lagrange multiplier.

(b) Write a computer program to calculate the iterates of the multiplier method with $\lambda^0 = 0$, and $c^k = 1$ for all k. Confirm computationally that the rate of convergence is sublinear if $\rho = 1.5$, linear if $\rho = 2$, and superlinear if $\rho = 3$.

(c) Give a heuristic argument why the rate of convergence is sublinear if $\rho < 2$, linear if $\rho = 2$, and superlinear if $\rho > 2$. What happens in the limit where $\rho = 1$?

5.2.3

Consider the problem of Exercise 5.2.1. Verify that the second order method of multipliers converges in one iteration provided c is sufficiently large, and estimate the threshold value for c.

5.2.4 (Convergence Threshold and Convergence Rate of the Method of Multipliers) www

Consider the quadratic problem

$$\begin{aligned} &\text{minimize} \;\; \tfrac{1}{2}x'Qx \\ &\text{subject to} \;\; Ax = b, \end{aligned}$$

where Q is symmetric and A is an $m \times n$ matrix of rank m. Let f^* be the optimal value of the problem and assume that the problem has a unique minimum x^* with associated Lagrange multiplier λ^*. Verify that for sufficiently large c, the penalized dual function is

$$q_c(\lambda) = -\tfrac{1}{2}(\lambda - \lambda^*)'A(Q + cA'A)^{-1}A'(\lambda - \lambda^*) + f^*.$$

(Use the quadratic programming duality theory of Section 4.4.2 to show that q_c is a quadratic function and to derive the Hessian matrix of q_c. Then use the fact that q_c is maximized at λ^* and that its maximum value is f^*.) Consider the first order method of multipliers.

(a) Use the theory of Section 1.3 to show that for all k

$$\|\lambda^{k+1} - \lambda^*\| \leq r^k \|\lambda^k - \lambda^*\|,$$

where

$$r^k = \max\bigl\{|1 - c^k E_{c^k}|,\ |1 - c^k e_{c^k}|\bigr\}$$

and E_c and e_c denote the maximum and the minimum eigenvalues of the matrix $A(Q + cA'A)^{-1}A'$.

(b) Assume that Q is invertible. Using the matrix identity

$$\bigl(I + c^k AQ^{-1}A'\bigr)^{-1} = I - c^k A(Q + c^k A'A)^{-1}A'$$

(cf. Section A.3 in Appendix A), relate the eigenvalues of the matrix $A(Q + c^k A'A)^{-1}A'$ with those of the matrix $AQ^{-1}A'$. Show that if $\gamma_1, \ldots, \gamma_m$ are the eigenvalues of $(AQ^{-1}A')^{-1}$, we have

$$r^k = \max_{i=1,\ldots,m} \left| \frac{\gamma_i}{\gamma_i + c^k} \right|.$$

(c) Show that the method converges to λ^* if $c > \bar{c}$, where $\bar{c} = 0$ if $\gamma_i \geq 0$ for all i, and $\bar{c} = -2\min\{\gamma_1, \ldots, \gamma_m\}$ otherwise.

5.2.5 (Stepsize Analysis of the Method of Multipliers) $\boxed{\text{www}}$

Consider the problem of Exercise 5.2.4. Use the results of that exercise to analyze the convergence and rate of convergence of the generalized method of multipliers

$$\lambda^{k+1} = \lambda^k + \alpha^k (Ax^k - b),$$

where α^k is a positive stepsize. Show in particular that if Q is positive definite and $c^k = c$ for all k, convergence is guaranteed if $\delta \leq \alpha^k \leq 2c$ for all k, where δ is some positive scalar. (For a solution and related analysis, see [Ber75d] and [Ber82a], p. 126.)

5.2.6

A weakness of the quadratic penalty method is that the augmented Lagrangian may not have a global minimum. As an example, show that the scalar problem

$$\begin{aligned} &\text{minimize} \quad -x^4 \\ &\text{subject to} \quad x = 0 \end{aligned}$$

has the unique global minimum $x^* = 0$ but its augmented Lagrangian

$$L_{c^k}(x, \lambda^k) = -x^4 + \lambda^k x + \frac{c^k}{2} x^2$$

has no global minimum for every c^k and λ^k. To overcome this difficulty, consider a penalty function of the form

$$\frac{c}{2} \|h(x)\|^2 + \|h(x)\|^\rho,$$

where $\rho > 4$, instead of $(c/2)\|h(x)\|^2$. Show that $L_{c^k}(x, \lambda^k)$ has a global minimum for every λ^k and $c^k > 0$.

5.2.7 (Two-Sided Inequality Constraints [Ber77], [Ber82a])

The purpose of this exercise is to show how to treat two-sided inequality constraints by using a *single* multiplier per constraint. Consider the problem

$$\begin{aligned} &\text{minimize} \quad f(x) \\ &\text{subject to} \quad \alpha_j \le g_j(x) \le \beta_j, \qquad j = 1, \ldots, r, \end{aligned}$$

where $f : \Re^n \mapsto \Re$ and $g_j : \Re^n \mapsto \Re$ are given functions, and α_j and β_j, $j = 1, \ldots, r$, are given scalars with $\alpha_j < \beta_j$. The method consists of sequential minimizations of the form

$$\begin{aligned} &\text{minimize} \quad f(x) + \sum_{j=1}^r P_j\big(g_j(x), \mu_j^k, c^k\big) \\ &\text{subject to} \quad x \in \Re^n, \end{aligned}$$

where

$$P_j\big(g_j(x), \mu_j^k, c^k\big) = \min_{u_j \in [g_j(x) - \beta_j,\, g_j(x) - \alpha_j]} \left\{ \mu_j^k u_j + \frac{c^k}{2} |u_j|^2 \right\}.$$

Each of these minimizations is followed by the multiplier iteration

$$\mu_j^{k+1} = \begin{cases} \mu_j^k + c^k\big(g_j(x^k) - \beta_j\big) & \text{if } \mu_j^k + c^k\big(g_j(x^k) - \beta_j\big) > 0, \\ \mu_j^k + c^k\big(g_j(x^k) - \alpha_j\big) & \text{if } \mu_j^k + c^k\big(g_j(x^k) - \alpha_j\big) < 0, \\ 0 & \text{otherwise}, \end{cases}$$

where x^k is a minimizing vector. Justify the method by introducing artificial variables u_j, by converting the problem to the equivalent form

$$\begin{aligned} &\text{minimize} \quad f(x) \\ &\text{subject to} \quad \alpha_j \le g_j(x) - u_j \le \beta_j, \qquad u_j = 0, \quad j = 1, \ldots, r, \end{aligned}$$

and by applying a multiplier method for this problem, where only the constraints $u_j = 0$ are eliminated by means of a quadratic penalty function (partial elimination of constraints).

5.2.8 (Proof of Ill-Conditioning as $c^k \to \infty$)

Consider the quadratic penalty method ($c^k \to \infty$) for the equality constrained problem of minimizing $f(x)$ subject to $h(x) = 0$, and assume that the generated sequence converges to a local minimum x^* that is also a regular point. Show that the condition number of the Hessian $\nabla^2_{xx} L_{c^k}(x^k, \lambda^k)$ tends to ∞. *Hint*: We have

$$\nabla^2_{xx} L_{c^k}(x^k, \lambda^k) = \nabla^2_{xx} L_0(x^k, \tilde{\lambda}^k) + c^k \nabla h(x^k) \nabla h(x^k)',$$

where $\tilde{\lambda}^k = \lambda^k + c^k h(x^k)$. The minimum eigenvalue $m(x^k, \lambda^k, c^k)$ of this matrix satisfies

$$\begin{aligned} m(x^k, \lambda^k, c^k) &= \min_{\|z\|=1} z' \nabla^2_{xx} L_{c^k}(x^k, \lambda^k) z \\ &\leq \min_{\|z\|=1, \nabla h(x^k)' z = 0} z' \nabla^2_{xx} L_{c^k}(x^k, \lambda^k) z \\ &= \min_{\|z\|=1, \nabla h(x^k)' z = 0} z' \nabla^2_{xx} L_0(x, \tilde{\lambda}^k) z. \end{aligned}$$

The maximum eigenvalue $M(x^k, \lambda^k, c^k)$ satisfies

$$\begin{aligned} M(x^k, \lambda^k, c^k) &= \max_{\|z\|=1} z' \nabla^2_{xx} L_{c^k}(x^k, \lambda^k) z \\ &\geq \min_{\|z\|=1} z' \nabla^2_{xx} L_0(x^k, \tilde{\lambda}^k) z + c^k \max_{\|z\|=1} z' \nabla h(x^k) \nabla h(x^k)' z. \end{aligned}$$

Use Prop. 5.2.2 to argue that $\nabla h(x^k)$ has rank m for sufficiently large k, and hence

$$\lim_{c^k \to \infty} \frac{M(x^k, \lambda^k, c^k)}{m(x^k, \lambda^k, c^k)} = \infty.$$

5.2.9 (www)

Let $\{x^k\}$ be a sequence generated by the logarithmic barrier method. Formulate conditions under which the sequences $\{-\epsilon^k / g_j(x^k)\}$ converge to corresponding Lagrange multipliers. *Hint*: Compare with the corresponding result of Prop. 5.2.2 for the quadratic penalty function.

5.2.10

State and prove analogs of Props. 5.2.1 and 5.2.2 for the case where the penalty function

$$\frac{c}{2} \|h(x)\|^2 + \|h(x)\|^\rho$$

with $\rho > 1$ is used in place of the quadratic $\frac{c}{2} \|h(x)\|^2$.

5.2.11 (Primal-Dual Methods not Using a Penalty) (www)

This exercise shows that an important dual ascent method, to be discussed in Section 7.2 (see also the end of Section 5.4.1), turns out to be equivalent to the first order method of multipliers for an artificial problem. Consider minimizing $f(x)$ subject to $h(x) = 0$, where f and h are twice continuously differentiable, and let x^* be a local minimum that is a regular point. Let λ^* be the associated Lagrange multiplier vector, and assume that the Hessian $\nabla^2_{xx}L(x^*, \lambda^*)$ of the (ordinary) Lagrangian

$$L(x, \lambda) = f(x) + \lambda' h(x)$$

is positive definite. (Note that this is stronger than what is required by the second order sufficiency conditions.) Consider the iteration

$$\lambda^{k+1} = \lambda^k + \alpha h(x^k),$$

where α is a positive scalar stepsize and x^k minimizes $L(x, \lambda^k)$ (the minimization is local in a suitable neighborhood of x^*). Show that there exists a threshold $\bar{\alpha} > 0$ and a sphere centered at λ^* such that if λ^0 belongs to this sphere and $\alpha < \bar{\alpha}$, then λ^k converges to λ^*. Consider first the case where f is quadratic and h is linear, and sketch an analysis for the more general case. *Hint*: Even though there is no penalty parameter here, the method can be viewed as a method of multipliers for the artificial problem

$$\begin{aligned} &\text{minimize} \quad f(x) - \frac{\alpha}{2}\big\|h(x)\big\|^2 \\ &\text{subject to} \quad h(x) = 0. \end{aligned}$$

Use the threshold of Exercise 5.2.4(c) to verify that $\bar{\alpha}$ can be taken to be equal to twice the minimum eigenvalue of $\nabla h(x^*)'\big(\nabla^2_{xx}L(x^*, \lambda^*)\big)^{-1}\nabla h(x^*)$. For analysis along this line, see [Ber82a], Section 2.6.

5.3 EXACT PENALTIES – SEQUENTIAL QUADRATIC PROGRAMMING

In this section we consider penalty methods that are *exact* in the sense that they require only one unconstrained minimization to obtain an optimal solution of the original constrained problem. We will use exact penalties as the basis for a broad class of algorithms, called *sequential quadratic programming*. These algorithms may also be viewed within the context of the Lagrangian methods of Section 5.4, and will be reencountered there.

To get a sense of how this is possible, consider the equality constrained problem

$$\begin{aligned} &\text{minimize} \quad f(x) \\ &\text{subject to} \quad h_i(x) = 0, \qquad i = 1, \ldots, m, \end{aligned} \tag{5.75}$$

where f and h_i are continuously differentiable, and let

$$L(x, \lambda) = f(x) + \lambda' h(x)$$

be the corresponding Lagrangian function. Then by minimizing the function

$$P(x, \lambda) = \|\nabla_x L(x, \lambda)\|^2 + \|h(x)\|^2 \tag{5.76}$$

over $(x, \lambda) \in \Re^{n+m}$ we can obtain local minima-Lagrange multiplier pairs (x^*, λ^*) satisfying the first order necessary conditions

$$\nabla_x L(x^*, \lambda^*) = 0, \qquad h(x^*) = 0.$$

We may view $P(x, \lambda)$ as an *exact penalty function*, i.e., a function whose unconstrained minima are (or strongly relate to) optimal solutions and/or Lagrange multipliers of a constrained problem.

The exact penalty function $P(x, \lambda)$ of Eq. (5.76) has (in effect) been used extensively in the special case where $m = n$ and the problem is to solve the system of constraint equations $h(x) = 0$ (in this case, any cost function f may be used). However, in the case where $m < n$, $P(x, \lambda)$ has significant drawbacks because it does not discriminate between local minima and local maxima, and it may also have local minima $(\bar{x}, \bar{\lambda})$ that are not global and do not satisfy the necessary optimality conditions, i.e., $P(\bar{x}, \bar{\lambda}) > 0$. There are, however, more sophisticated exact penalty functions that do not have these drawbacks, as we will see later.

We may distinguish between *differentiable* and *nondifferentiable* exact penalty functions. The former have the advantage that they can be minimized by the unconstrained methods we have already studied in Chapters 1 and 2. The latter involve nondifferentiabilities, so the methods of Chapters 1 and 2 are not directly applicable. We will develop special algorithms, called *sequential quadratic programming methods*, for their minimization.

Nondifferentiable exact penalty methods have been more popular in practice than their differentiable counterparts, and they will receive most of our attention. On the other hand, differentiable exact penalty methods have some interesting advantages; see the monograph [Ber82a] (Section 4.3), which treats extensively both types of methods.

5.3.1 Nondifferentiable Exact Penalty Functions

Our first objective in this section is to show that solutions of the equality constrained problem (5.75) are related to solutions of the (nondifferentiable) unconstrained problem

$$\begin{aligned} &\text{minimize} \quad f(x) + cP(x) \\ &\text{subject to} \quad x \in \Re^n, \end{aligned}$$

where $c > 0$ and P is the nondifferentiable penalty function defined by

$$P(x) = \max_{i=1,\ldots,m} |h_i(x)|.$$

We develop the main argument for equality constraints, then generalize to include inequality constraints, and finally state the result in Prop. 5.3.1.

Indeed let x^* be a local minimum, which is a regular point and satisfies together with a corresponding Lagrange multiplier vector λ^*, the second order sufficiency conditions of Prop. 4.2.1. Consider also the *primal function* p defined in a neighborhood of the origin by

$$p(u) = \min\big\{f(x) \mid h(x) = u, \|x - x^*\| < \epsilon\big\},$$

where $\epsilon > 0$ is some scalar; see the sensitivity theorem (Prop. 4.2.2). Then if we locally minimize $f + cP$ around x^*, we can split the minimization in two: first minimize over all x satisfying $h(x) = u$ and then minimize over all possible u. We have

$$\begin{aligned}
&\inf_{\|x-x^*\|<\epsilon} \left\{ f(x) + c \max_{i=1,\ldots,m} \big|h_i(x)\big| \right\} \\
&\qquad = \inf_{u \in U_\epsilon} \ \inf_{\{x \mid h(x)=u,\ \|x-x^*\|<\epsilon\}} \left\{ f(x) + c \max_{i=1,\ldots,m} \big|h_i(x)\big| \right\} \\
&\qquad = \inf_{u \in U_\epsilon} p_c(u),
\end{aligned}$$

where

$$p_c(u) = p(u) + c \max_{i=1,\ldots,m} |u_i|,$$

$$U_\epsilon = \big\{u \mid h(x) = u \text{ for some } x \text{ with } \|x - x^*\| < \epsilon\big\}.$$

We may view p_c as a *penalized primal function*. We will now show that for large enough c, p_c has a local minimum at $u = 0$; cf. Fig. 5.3.1.

Since, according to the sensitivity theorem, we have $\nabla p(0) = -\lambda^*$, we can use the mean value theorem to write for each u in a neighborhood of the origin

$$p(u) = p(0) - u'\lambda^* + \tfrac{1}{2} u' \nabla^2 p(\bar{\alpha} u) u,$$

where $\bar{\alpha}$ is some scalar in $[0, 1]$. Thus

$$p_c(u) = p(0) - \sum_{i=1}^m u_i \lambda_i^* + c \max_{i=1,\ldots,m} |u_i| + \tfrac{1}{2} u' \nabla^2 p(\bar{\alpha} u) u. \tag{5.77}$$

Assume that c is sufficiently large so that for some $\gamma > 0$,

$$c \geq \sum_{i=1}^m |\lambda_i^*| + \gamma.$$

Then it follows that

$$\begin{aligned}
c \max_{i=1,\ldots,m} |u_i| &\geq \left(\sum_{i=1}^m |\lambda_i^*| + \gamma \right) \max_{i=1,\ldots,m} |u_i| \\
&\geq \sum_{i=1}^m u_i \lambda_i^* + \gamma \max_{i=1,\ldots,m} |u_i|.
\end{aligned}$$

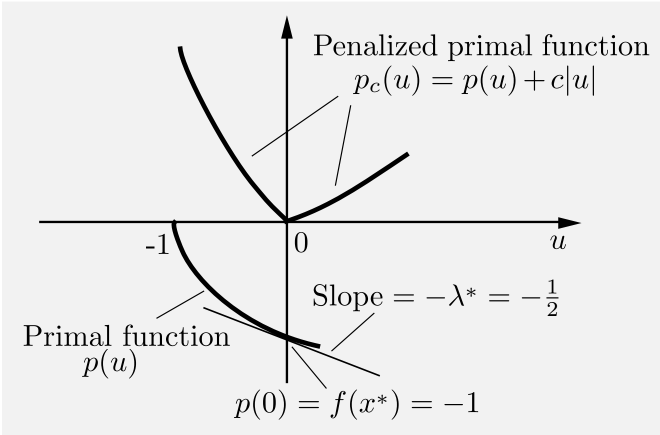

Figure 5.3.2. Illustration of how for c large enough, $u = 0$ is a strict local minimum of $p_c(u) = p(u) + c\max_{i=1,\ldots,m}|u_i|$. The figure corresponds to the two-dimensional problem where $f(x) = x_1$ and $h(x) = x_1^2 + x_2^2 - 1$. The optimal solution and Lagrange multiplier are $x^* = (-1, 0)$ and $\lambda^* = 1/2$, respectively. The primal function is defined for $u \geq -1$ and is given by

$$p(u) = \min_{x_1^2 + x_2^2 - 1 = u} x_1 = -\sqrt{1+u}.$$

Note that $\nabla p(0) = \lambda^*$ and that we must have $c > \lambda^*$ in order for the nondifferentiable penalty function to be exact.

Using this relation in Eq. (5.77), we obtain

$$p_c(u) \geq p(0) + \gamma \max_{i=1,\ldots,m} |u_i| + \tfrac{1}{2} u' \nabla^2 p(\bar{\alpha} u) u.$$

For u sufficiently close to zero, the last term is dominated by the next to last term, so

$$p_c(u) > p(0) = p_c(0)$$

for all $u \neq 0$ in a neighborhood N of the origin. Hence $u = 0$ is a strict local minimum of p_c as shown in Fig. 5.3.2. Since we have

$$f(x) + cP(x) \geq p(u) + c \max_{i=1,\ldots,m} |u_i| = p_c(u)$$

for all $u \neq 0$ in the neighborhood N and x such that $h(x) = u$ with $\|x - x^*\| < \epsilon$, and we also have $p_c(0) = f(x^*)$, we obtain that for all x in a neighborhood of x^* with $P(x) > 0$,

$$f(x) + cP(x) > f(x^*).$$

Since x^* is a strict local minimum of f over all x with $P(x) = 0$, it follows that x^* is a strict local minimum of $f + cP$, provided that $c > \sum_{i=1}^m |\lambda_i^*|$.

The above argument can be extended to the case where there are additional inequality constraints of the form $g_j(x) \le 0$, $j = 1, \ldots, r$. These constraints can be converted to the equality constraints

$$g_j(x) + z_j^2 = 0, \qquad j = 1, \ldots, r,$$

by introducing the squared slack variables z_j as in Section 4.3. The slack variables can be eliminated from the corresponding exact penalty function

$$f(x) + c \max\bigl\{|g_1(x) + z_1^2|, \ldots, |g_r(x) + z_r^2|, |h_1(x)|, \ldots, |h_m(x)|\bigr\} \tag{5.78}$$

by explicit minimization. In particular, we have

$$\min_{z_j} \bigl|g_j(x) + z_j^2\bigr| = \max\bigl\{0, g_j(x)\bigr\},$$

so minimization of the exact penalty function (5.78) is equivalent to minimization of the function

$$f(x) + c \max\bigl\{0, g_1(x), \ldots, g_r(x), |h_1(x)|, \ldots, |h_m(x)|\bigr\}.$$

Thus, by repeating the earlier argument given for equality constraints, we have the following proposition.

Proposition 5.3.1: Let x^* be a local minimum of the problem

$$\begin{aligned} &\text{minimize} \quad f(x) \\ &\text{subject to} \quad h_i(x) = 0, \quad i = 1, \ldots, m, \qquad g_j(x) \le 0, \quad j = 1, \ldots, r, \end{aligned}$$

which is regular and satisfies together with corresponding Lagrange multiplier vectors λ^* and μ^*, the second order sufficiency conditions of Prop. 4.3.2. Then, if

$$c > \sum_{i=1}^{m} |\lambda_i^*| + \sum_{j=1}^{r} \mu_j^*,$$

the vector x^* is a strict unconstrained local minimum of $f + cP$, where

$$P(x) = \max\bigl\{0, g_1(x), \ldots, g_r(x), |h_1(x)|, \ldots, |h_m(x)|\bigr\}.$$

An example illustrating the above proposition is given in Fig. 5.3.3. The proof of the proposition was relatively simple but made assumptions that are stronger than necessary. There are related results that do not

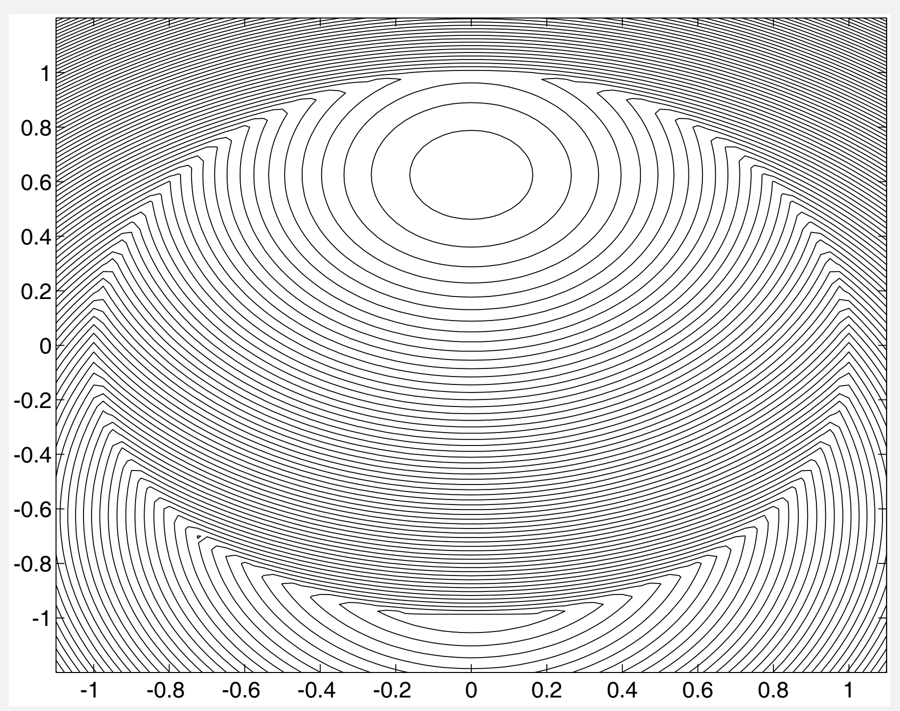

Figure 5.3.3. Equal cost surfaces of the function $f + cP$ for the two-dimensional problem where

$$f(x) = x_1, \qquad h(x) = x_1^2 + x_2^2 - 1$$

(cf. Fig. 5.3.2). For c greater than the Lagrange multiplier $\lambda^* = 1/2$, the optimal solution $x^* = (-1, 0)$ is a local minimum of $f + cP$. This is not so for $c < \lambda^*$. The figure corresponds to $c = 0.8$.

require second order differentiability assumptions. In particular, it is shown in Exercise 5.3.4 that if c is sufficiently large, a regular local minimum x^* is a "stationary" point of $f + cP$ (in a sense to be made precise shortly). The reverse is not necessarily true. In particular, there may exist local minima of $f + cP$ that do not correspond to constrained local minima of f for any c; see Exercise 5.3.1. There is also a more refined analysis that requires just first order differentiability; see [BNO03], Section 5.5.

Finally, let us note that under convexity assumptions, where f is convex, h_i are linear, g_j are convex, and there is an additional convex abstract set constraint, the results that connect global minima of the original problem and global minima of corresponding nondifferentiable exact penalty functions are more powerful. In particular, the analysis is not tied to a specific local minimum x^*, and there no need for the type of differentiability, sufficiency, and regularity assumptions that we are using in Prop. 5.3.1; see [Ber75a], and the textbook accounts [BNO03], Section 7.3, and [Ber15a], Section 1.5.

Descent Directions of Exact Penalties

We will now take the first steps towards algorithms for minimizing exact penalty functions, by characterizing their descent directions. In order to simplify notation, we will assume that all the constraints are inequalities. The analysis and algorithms to be given admit simple extensions to the equality constrained case simply by converting each equality constraint into two inequalities. We assume throughout that the cost and constraint functions are at least once continuously differentiable.

We will discuss properties of unconstrained minima of $f + cP$, with $c > 0$ and

$$P(x) = \max\{g_0(x), g_1(x), \ldots, g_r(x)\}, \qquad \forall\, x \in \Re^n, \tag{5.79}$$

where for notational convenience, we denote by g_0 the function that is identically zero:

$$g_0(x) \equiv 0, \qquad x \in \Re^n. \tag{5.80}$$

We first introduce some notation and definitions, and we develop some preliminary results.

For $x \in \Re^n$, $d \in \Re^n$, and $c > 0$, we consider the index set

$$J(x) = \{j \mid g_j(x) = P(x),\ j = 0, 1, \ldots, r\},$$

and we denote

$$\theta_c(x; d) = \max\{\nabla f(x)'d + c\nabla g_j(x)'d \mid j \in J(x)\}.$$

The function θ_c plays the role that the gradient would play if $f + cP$ were differentiable. In particular, the function

$$f(x) + cP(x) + \theta_c(x; d)$$

may be viewed as a linear approximation of $f + cP$ for variations d around x; see Fig. 5.3.4.

Since at an unconstrained local minimum x^*, $f + cP$ cannot decrease along any direction, the preceding interpretation of θ_c motivates us to call a vector x^* a *stationary point of* $f + cP$ if for all $d \in \Re^n$ there holds

$$\theta_c(x^*; d) \geq 0.$$

The following proposition shows that local minima of $f + cP$ must be stationary points of $f + cP$. Furthermore, *descent directions of* $f + cP$ *at a nonstationary point* x *can be obtained from the following convex quadratic program, in* $(d, \xi) \in \Re^{n+1}$:

$$\begin{aligned} &\text{minimize} \ \ \nabla f(x)'d + \tfrac{1}{2}d'Hd + c\xi \\ &\text{subject to} \ \ (d, \xi) \in \Re^{n+1}, \qquad g_j(x) + \nabla g_j(x)'d \leq \xi, \quad j = 0, 1, \ldots, r, \end{aligned} \tag{5.81}$$

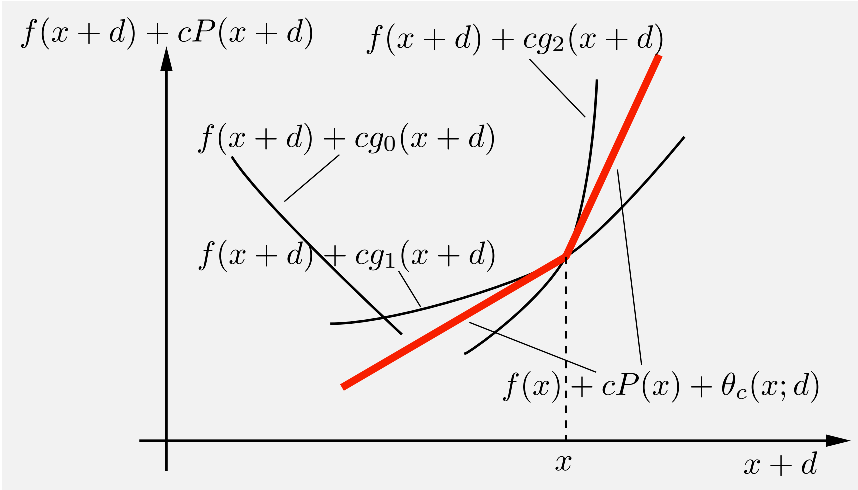

Figure 5.3.4. Illustration of $\theta_c(x;d)$ at a point x. It is the first order estimate of the variation

$$f(x+d)+cP(x+d)-f(x)-cP(x)$$

of $f+cP$ around x. Here the index set $J(x)$ is $\{1,2\}$.

where $c>0$ and H is a positive definite symmetric matrix.

To understand the role of this quadratic program, note that for a fixed d, the minimum with respect to ξ is attained at

$$\xi = \max_{j=0,1,\ldots,r} \big\{ g_j(x) + \nabla g_j(x)'d \big\}.$$

Thus, by eliminating the variable ξ, and by adding to the cost the constant term $f(x)$, we can write the quadratic program (5.81) in the alternative form

$$\begin{aligned} &\text{minimize} \quad \max_{j=0,1,\ldots,r} \big\{ f(x) + cg_j(x) + \nabla f(x)'d + c\nabla g_j(x)'d \big\} + \tfrac{1}{2} d'Hd \\ &\text{subject to} \quad d \in \Re^n. \end{aligned} \tag{5.82}$$

For small $\|d\|$, the maximum over j is attained for $j \in J(x)$, so we can substitute $P(x)$ in place of $g_j(x)$. The cost function then takes the form

$$f(x) + cP(x) + \theta_c(x;d) + \tfrac{1}{2} d'Hd,$$

so locally, for d near zero, the problem (5.82) can be viewed as minimization of a quadratic approximation of $f+cP$ around x; see Fig. 5.3.5. Also since the cost function of problem (5.82) is strictly convex in d, its optimal solution is unique, implying also that the quadratic program (5.81) has a unique optimal solution (d,ξ).

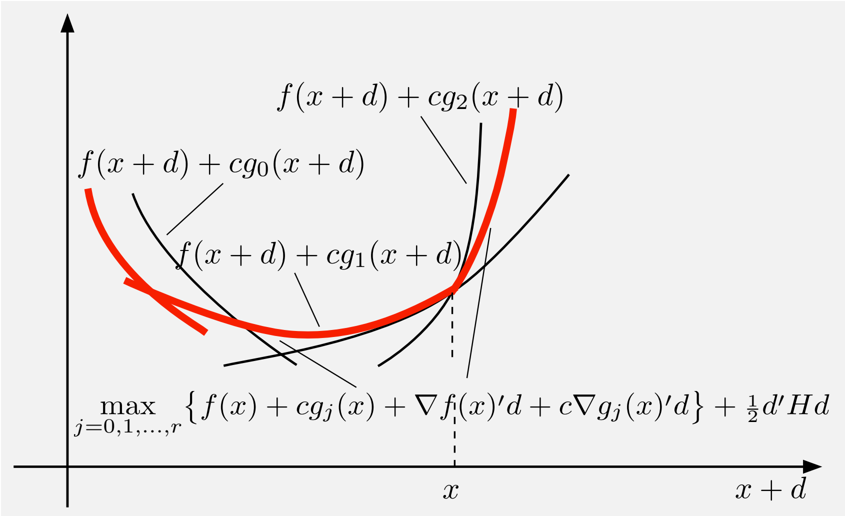

Figure 5.3.5. Illustration of the cost function

$$\max_{j=0,1,\ldots,r}\left\{f(x)+cg_j(x)+\nabla f(x)'d+c\nabla g_j(x)'d\right\}+\tfrac{1}{2}d'Hd$$

of the quadratic program (5.82). For small $\|d\|$ this function takes the form

$$f(x)+cP(x)+\theta_c(x;d)+\tfrac{1}{2}d'Hd,$$

and is a quadratic approximation of $f+cP$ around x. It can be seen that by minimizing this function over d we obtain a direction of descent of $f+cP$ at x.

Proposition 5.3.2: (Descent Directions of $f+cP$)

(a) For all $x\in\Re^n$, $d\in\Re^n$, and $\alpha>0$, we have

$$f(x+\alpha d)+cP(x+\alpha d)-f(x)-cP(x)=\alpha\theta_c(x;d)+o(\alpha), \quad (5.83)$$

where $\lim_{\alpha\to 0^+}o(\alpha)/\alpha=0$. As a result, if $\theta_c(x;d)<0$, then d is a descent direction, i.e., there exists $\bar{\alpha}>0$ such that

$$f(x+\alpha d)+cP(x+\alpha d)<f(x)+cP(x), \qquad \forall\ \alpha\in(0,\bar{\alpha}].$$

Moreover, a local minimum of $f+cP$ is a stationary point.

(b) If f and g_j are convex functions, then a stationary point of $f+cP$ is also a global minimum of $f+cP$.

(c) For any $x\in\Re^n$ and positive definite symmetric H, if (d,ξ) is the optimal solution of the quadratic program (5.81), then

$$\theta_c(x; d) \le -d'Hd. \tag{5.84}$$

(d) A vector x is a stationary point of $f + cP$ if and only if the quadratic program (5.81) has $\big\{d = 0, \xi = P(x)\big\}$ as its optimal solution.

Proof: (a) We have for all $\alpha > 0$ and $j \in J(x)$,

$$\begin{aligned} f(x + \alpha d) + cg_j(x + \alpha d) = {} & f(x) + \alpha \nabla f(x)'d + c\big(g_j(x) + \alpha \nabla g_i(x)'d\big) \\ & + o_j(\alpha), \end{aligned}$$

where $\lim_{\alpha \to 0^+} o_j(\alpha)/\alpha = 0$. Hence, by using the fact $g_j(x) = P(x)$ for all $j \in J(x)$,

$$\begin{aligned} & f(x + \alpha d) + c \max\big\{g_j(x + \alpha d) \mid j \in J(x)\big\} \\ & \quad = f(x) + \alpha \nabla f(x)'d + c \max\big\{g_j(x) + \alpha \nabla g_j(x)'d \mid j \in J(x)\big\} + o(\alpha) \\ & \quad = f(x) + cP(x) + \alpha \theta_c(x; d) + o(\alpha), \end{aligned}$$

where $\lim_{\alpha \to 0^+} o(\alpha)/\alpha = 0$. We have, for all α that are sufficiently small,

$$\begin{aligned} \max\big\{g_j(x + \alpha d) \mid j \in J(x)\big\} & = \max\big\{g_j(x + \alpha d) \mid j = 0, 1, \ldots, r\big\} \\ & = P(x + \alpha d). \end{aligned}$$

Combining the two above relations, we obtain

$$f(x + \alpha d) + cP(x + \alpha d) = f(x) + cP(x) + \alpha \theta_c(x; d) + o(\alpha),$$

which is Eq. (5.83).

If x^* is a local minimum of $f + cP$, then by Eq. (5.83), we have, for all d and $\alpha > 0$ such that $\|d\|$ and α are sufficiently small,

$$\alpha \theta_c(x^*; d) + o(\alpha) \ge 0.$$

Dividing by α and taking the limit as $\alpha \to 0$, we obtain $\theta_c(x^*; d) \ge 0$, so x^* is a stationary point of $f + cP$.

(b) By convexity, we have [cf. Prop. B.3(a) in Appendix B] for all j and $x \in \Re^n$,

$$f(x) + cg_j(x) \ge f(x^*) + cg_j(x^*) + \big(\nabla f(x^*) + c\nabla g_j(x^*)\big)'(x - x^*).$$

Taking the maximum over j, we obtain

$$f(x) + cP(x) \ge \max_{j = 0, 1, \ldots, r} \big\{f(x^*) + cg_j(x^*) + \big(\nabla f(x^*) + c\nabla g_j(x^*)\big)'(x - x^*)\big\}.$$

For a sufficiently small scalar ϵ and for all x with $\|x-x^*\| < \epsilon$, the maximum above is attained for some $j \in J(x^*)$. Since $g_j(x^*) = P(x^*)$ for all $j \in J(x^*)$, we obtain for all x with $\|x - x^*\| < \epsilon$,

$$f(x) + cP(x) \geq f(x^*) + cP(x^*) + \theta_c(x^*; x - x^*) \geq f(x^*) + cP(x^*),$$

where the last inequality holds because x^* is a stationary point of $f + cP$. Hence x^* is a local minimum of $f + cP$, and in view of the convexity of $f + cP$, x^* is a global minimum.

(c) We have $g_j(x) + \nabla g_j(x)'d \leq \xi$ for all j. Since $g_j(x) = P(x)$ for all $j \in J(x)$, it follows that $\nabla g_j(x)'d \leq \xi - P(x)$ for all $j \in J(x)$ and therefore using the definition of θ_c we have

$$\theta_c(x; d) \leq \nabla f(x)'d + c\big(\xi - P(x)\big). \tag{5.85}$$

Let $\{\mu_j\}$ be a set of Lagrange multipliers for the quadratic program (5.81). The optimality conditions yield

$$\nabla f(x) + Hd + \sum_{j=0}^{r} \mu_j \nabla g_j(x) = 0, \tag{5.86}$$

$$c - \sum_{j=0}^{r} \mu_j = 0, \tag{5.87}$$

$$g_j(x) + \nabla g_j(x)'d \leq \xi, \qquad \mu_j \geq 0, \qquad j = 0, 1, \ldots, r,$$
$$\mu_j\big(g_j(x) + \nabla g_j(x)'d - \xi\big) = 0, \qquad j = 0, 1, \ldots, r.$$

By adding the last equation over all j and using Eq. (5.87), we have

$$\begin{aligned}
\sum_{j=0}^{r} \mu_j \nabla g_j(x)'d &= \sum_{j=0}^{r} \mu_j \xi - \sum_{j=0}^{r} \mu_j g_j(x) \\
&\geq \sum_{j=0}^{r} \mu_j \big(\xi - \max_{m=0,1,\ldots,r} g_m(x)\big) \\
&= \sum_{j=0}^{r} \mu_j \Big(\xi - P(x)\Big) \\
&= c\big(\xi - P(x)\big).
\end{aligned}$$

Combining this equation with Eq. (5.86) we obtain

$$\nabla f(x)'d + d'Hd + c\big(\xi - P(x)\big) \leq 0, \tag{5.88}$$

which in conjunction with Eq. (5.85) yields

$$\theta_c(x; d) + d'Hd \leq 0,$$

thus proving Eq. (5.84).

(d) We have that x is a stationary point of $f + cP$ if and only if $\theta_c(x; d) \geq 0$ for all d, which by Eq. (5.84) is true if and only if $\big\{d = 0,\ \xi = P(x)\big\}$ is the optimal solution of the quadratic program (5.81). **Q.E.D.**

5.3.2 Sequential Quadratic Programming

We now introduce an iterative descent algorithm for minimizing the exact penalty function $f+cP$. It is called the *linearization algorithm* or *sequential quadratic programming*. Like the gradient projection method, it calculates the descent direction by solving a quadratic programming subproblem of the form (5.81). The algorithm is given by

$$x^{k+1} = x^k + \alpha^k d^k,$$

where α^k is a nonnegative scalar stepsize, and d^k is a direction obtained by solving the quadratic program in (d, ξ)

$$\begin{aligned} &\text{minimize} \;\; \nabla f(x^k)'d + \tfrac{1}{2} d'H^k d + c\xi \\ &\text{subject to} \;\; g_j(x^k) + \nabla g_j(x^k)'d \le \xi, \qquad j = 0, 1, \ldots, r, \end{aligned} \tag{5.89}$$

where H^k is a positive definite symmetric matrix. Proposition 5.3.2(c) implies that the solution d is a descent direction of $f + cP$ at x^k.

The initial vector x^0 is arbitrary and the stepsize α^k is chosen by any one of the stepsize rules listed below:

(a) *Minimization rule*: Here α^k is chosen so that

$$f(x^k + \alpha^k d^k) + cP(x^k + \alpha^k d^k) = \min_{\alpha \ge 0} \big\{ f(x^k + \alpha d^k) + cP(x^k + \alpha d^k) \big\}.$$

(b) *Limited minimization rule*: Here a fixed scalar $s > 0$ is selected and α^k is chosen so that

$$f(x^k + \alpha^k d^k) + cP(x^k + \alpha^k d^k) = \min_{\alpha \in [0,s]} \big\{ f(x^k + \alpha d^k) + cP(x^k + \alpha d^k) \big\}.$$

(c) *Armijo rule*: Here fixed scalars s, β, and σ with $s > 0$, $\beta \in (0,1)$, and $\sigma \in (0, \frac{1}{2})$, are selected, and we set $\alpha^k = \beta^{m_k} s$, where m_k is the first nonnegative integer m for which

$$f(x^k) + cP(x^k) - f(x^k + \beta^m s d^k) - cP(x^k + \beta^m s d^k) \ge \sigma \beta^m s d^{k\prime} H^k d^k. \tag{5.90}$$

It can be shown that if $d^k \neq 0$, the Armijo rule will yield a stepsize after a finite number of arithmetic operations. To see this, note that by Prop. 5.3.2(a) and Eq. (5.84), we have for all $\alpha > 0$,

$$\begin{aligned} f(x^k) + cP(x^k) - f(x^k + \alpha d^k) - cP(x^k + \alpha d^k) &= -\alpha \theta_c(x^k; d^k) + o(\alpha) \\ &\ge \alpha d^{k\prime} H^k d^k + o(\alpha). \end{aligned}$$

We then obtain

$$f(x^k) + cP(x^k) - f(x^k + \alpha d^k) - cP(x^k + \alpha d^k) \ge \sigma \alpha d^{k\prime} H^k d^k, \qquad \forall\, \alpha \in (0, \bar{\alpha}],$$

where $\bar{\alpha} > 0$ is such that we have $(1-\sigma)\alpha d^{k\prime} H^k d^k + o(\alpha) \geq 0$ for all $\alpha \in (0, \bar{\alpha}]$. Therefore, if $d^k \neq 0$, there is an integer m such that the Armijo test (5.90) is passed, while if $d^k = 0$, by Prop. 5.3.2(d), x^k is a stationary point of $f + cP$.

We have the following convergence result. Its proof is patterned after the corresponding proof for gradient methods for unconstrained minimization (cf. Prop. 1.2.1 in Section 1.2), but is considerably more complicated due to the constraints.

Proposition 5.3.3: Let $\{x^k\}$ be a sequence generated by the linearization algorithm, where the stepsize α^k is chosen by the minimization rule, the limited minimization rule, or the Armijo rule. Assume that there exist positive scalars γ and Γ such that

$$\gamma \|z\|^2 \leq z' H^k z \leq \Gamma \|z\|^2, \qquad \forall\, z \in \Re^n, \qquad k = 0, 1, \ldots,$$

(this condition corresponds to the assumption of a gradient-related direction sequence in unconstrained optimization). Then every limit point of $\{x^k\}$ is a stationary point of $f + cP$.

Proof: We argue by contradiction. Assume that a subsequence $\{x^k\}_K$ generated by the algorithm using the Armijo rule converges to a vector $\bar{x}$ that is not a stationary point of $f + cP$. Since $f(x^k) + cP(x^k)$ is monotonically decreasing, we have

$$f(x^k) + cP(x^k) \to f(\bar{x}) + cP(\bar{x})$$

and hence also

$$f(x^k) + cP(x^k) - f(x^{k+1}) - cP(x^{k+1}) \to 0.$$

By the definition of the Armijo rule, we have

$$f(x^k) + cP(x^k) - f(x^{k+1}) - cP(x^{k+1}) \geq \sigma \alpha^k d^{k\prime} H^k d^k.$$

Hence

$$\alpha^k d^{k\prime} H^k d^k \to 0. \tag{5.91}$$

Since for $k \in K$, d^k is the optimal solution of the quadratic program (5.89), we must have for some set of Lagrange multipliers $\{\mu_j^k\}$ and all $k \in K$,

$$\nabla f(x^k) + \sum_{j=0}^{r} \mu_j^k \nabla g_j(x^k) + H^k d^k = 0, \qquad c = \sum_{j=0}^{r} \mu_j^k, \tag{5.92}$$

$$\mu_j^k \geq 0, \qquad \mu_j^k \big(g_j(x^k) + \nabla g_j(x^k)' d^k - \xi^k \big) = 0, \qquad j = 0, 1, \ldots, r, \tag{5.93}$$

where

$$\xi^k = \max_{j=0,1,\ldots,r} \big\{ g_j(x^k) + \nabla g_j(x^k)' d^k \big\}.$$

The relations $c = \sum_{j=0}^r \mu_j^k$ and $\mu_j^k \geq 0$ imply that the subsequences $\{\mu_j^k\}$ are bounded. Hence, without loss of generality, we may assume that for some μ_j, $j = 0, 1, \ldots, r$, we have

$$\{\mu_j^k\}_K \to \bar{\mu}_j, \qquad j = 0, 1, \ldots, r. \tag{5.94}$$

Using the assumption $\gamma \|z\|^2 \leq z' H^k z \leq \Gamma \|z\|^2$, we may also assume without loss of generality that

$$\{H^k\}_K \to \bar{H} \tag{5.95}$$

for some positive definite matrix $\bar{H}$.

Now from the fact $\alpha^k d^{k\prime} H^k d^k \to 0$ [cf. Eq. (5.91)], we see that there are two possibilities. Either

$$\liminf_{k \to \infty,\, k \in K} \|d^k\| = 0, \tag{5.96}$$

or else

$$\liminf_{k \to \infty,\, k \in K} \alpha^k = 0, \qquad \liminf_{k \to \infty,\, k \in K} \|d^k\| > 0. \tag{5.97}$$

If Eq. (5.96) holds, then we may assume without loss of generality that $\{d^k\}_K \to 0$, and by taking the limit in Eqs. (5.92) and (5.93), and using Eq. (5.94), we have

$$\nabla f(\bar{x}) + \sum_{j=0}^r \bar{\mu}_j \nabla g_j(\bar{x}) = 0, \qquad c = \sum_{j=0}^r \bar{\mu}_j,$$

$$\bar{\mu}_j \geq 0, \qquad \bar{\mu}_j \big(g_j(\bar{x}) - \xi \big) = 0, \qquad j = 0, 1, \ldots, r,$$

where $\xi = \max_{j=0,1,\ldots,r} g_j(\bar{x})$. Hence the quadratic program (5.81) corresponding to $\bar{x}$ has $\big\{ d = 0,\, \xi = P(\bar{x}) \big\}$ as its optimal solution. From Prop. 5.3.2(d), it follows that $\bar{x}$ is a stationary point of $f + cP$, thus contradicting the hypothesis made earlier.

It will thus suffice to assume that Eq. (5.97) holds and to arrive at a contradiction. We may assume without loss of generality that

$$\{\alpha^k\}_K \to 0.$$

Since Eqs. (5.92), (5.94), and (5.95) show that $\{d^k\}_K$ is a bounded sequence, we may also assume without loss of generality that

$$\{d^k\}_K \to \bar{d},$$

where $\bar{d}$ is some vector that cannot be zero in view of Eq. (5.97).

Since $\{\alpha^k\}_K \to 0$, it follows, in view of the definition of the Armijo rule, that the initial stepsize s will be reduced at least once for all $k \in K$ after some index $\bar{k}$. This means that for all $k \in K$, $k \geq \bar{k}$,

$$f(x^k) + cP(x^k) - f(x^k + \bar{\alpha}^k d^k) - cP(x^k + \bar{\alpha}^k d^k) < \sigma \bar{\alpha}^k d^{k\prime} H^k d^k, \quad (5.98)$$

where $\bar{\alpha}^k = \alpha^k/\beta$.

Define for all k and d,

$$\zeta^k(d) = \nabla f(x^k)'d + c \max_{j \in J(x^k)} \big\{ g_j(x^k) + \nabla g_j(x^k)'d \big\} - cP(x^k),$$

and restrict attention to $k \in K$, $k \geq \bar{k}$, that are sufficiently large so that $\bar{\alpha}^k \leq 1$, $J(x^k) \subset J(\bar{x})$, and $J(x^k + \bar{\alpha}^k d^k) \subset J(\bar{x})$. We will show that

$$f(x^k) + cP(x^k) - f(x^k + \bar{\alpha}^k d^k) - cP(x^k + \bar{\alpha}^k d^k) = -\zeta^k(\bar{\alpha}^k d^k) + o(\bar{\alpha}^k), \quad (5.99)$$

where

$$\lim_{k \to \infty,\, k \in K} \frac{o(\bar{\alpha}^k)}{\bar{\alpha}^k} = 0. \quad (5.100)$$

Indeed, we have

$$f(x^k + \bar{\alpha}^k d^k) = f(x^k) + \bar{\alpha}^k \nabla f(x^k)'d^k + o_0(\bar{\alpha}^k \|d^k\|)$$

$$g_j(x^k + \bar{\alpha}^k d^k) = g_j(x^k) + \bar{\alpha}^k \nabla g_j(x^k)'d^k + o_j(\bar{\alpha}^k \|d^k\|), \qquad j \in J(x^k),$$

where $o_j(\cdot)$ are functions satisfying $\lim_{k \to \infty} o_j\big(\bar{\alpha}^k \|d^k\|\big)/\bar{\alpha}^k = 0$. Adding and taking the maximum over $j \in J(x)$, and using the fact $J(x^k + \bar{\alpha}^k d^k) \subset J(\bar{x})$ [implying that $P(x^k + \bar{\alpha}^k d^k) = \max_j g_j(x^k + \bar{\alpha}^k d^k)$], we obtain for sufficiently large k,

$$\begin{aligned} f(x^k + \bar{\alpha}^k d^k) + cP(x^k + \bar{\alpha}^k d^k) &= f(x^k) + \bar{\alpha}^k \nabla f(x^k)'d^k \\ &\quad + c \max_{j \in J(x^k)} \big\{ g_j(x^k) + \bar{\alpha}^k \nabla g_j(x^k)'d^k \big\} + o(\bar{\alpha}^k \|d^k\|) \\ &= f(x^k) + cP(x^k) + \zeta^k(\bar{\alpha}^k d^k) + o(\bar{\alpha}^k), \end{aligned}$$

thus proving Eq. (5.99).

We also claim that

$$-\frac{\zeta^k(\bar{\alpha}^k d^k)}{\bar{\alpha}^k} \geq -\zeta^k(d^k) \geq d^{k\prime} H^k d^k. \quad (5.101)$$

Indeed, let (d^k, ξ^k) be the optimal solution of the quadratic program

$$\begin{aligned} &\text{minimize} \ \ \nabla f(x^k)'d + \tfrac{1}{2} d' H^k d + c\xi \\ &\text{subject to} \ \ g_j(x^k) + \nabla g_j(x^k)'d \leq \xi, \quad j = 0, 1, \ldots, r. \end{aligned}$$

We have

$$
\begin{aligned}
\xi^k &= \max_{j=0,1,\ldots,r} \bigl\{ g_j(x^k) + \nabla g_j(x^k)' d^k \bigr\} \\
&\geq \max_{j \in J(x)} \bigl\{ g_j(x^k) + \nabla g_j(x^k)' d^k \bigr\} \\
&= \frac{\zeta^k(d^k) - \nabla f(x^k)' d^k}{c} + P(x^k).
\end{aligned}
$$

On the other hand, in the proof of Prop. 5.3.2(c) we showed [cf. Eq. (5.88)] that

$$
c\bigl(\xi^k - P(x^k)\bigr) - \nabla f(x^k)' d^k \geq d^{k\prime} H^k d^k.
$$

The last two equations, together with the relation $\zeta^k(\bar{\alpha}^k d) \leq \bar{\alpha}^k \zeta^k(d)$, which follows from the convexity of $\zeta^k(\cdot)$, prove Eq. (5.101).

By dividing Eq. (5.98) with $\bar{\alpha}^k$ and by combining it with Eq. (5.99), we obtain

$$
\sigma d^{k\prime} H^k d^k > -\frac{\zeta^k(\bar{\alpha}^k d)}{\bar{\alpha}^k} + \frac{o(\bar{\alpha}^k)}{\bar{\alpha}^k},
$$

which in view of Eq. (5.101), yields

$$
(1-\sigma) d^{k\prime} H^k d^k + \frac{o(\bar{\alpha}^k)}{\bar{\alpha}^k} < 0.
$$

Since $\{H^k\}_K \to \bar{H}$, $\{d^k\}_K \to \bar{d}$, $\bar{H}$ is positive definite, $\bar{d} \neq 0$, and $o(\bar{\alpha}^k)/\bar{\alpha}^k \to 0$ [cf. Eq. (5.100)], we obtain a contradiction. This completes the proof of the proposition for the case of the Armijo rule.

Consider now the minimization rule and let $\{x^k\}_K$ converge to a vector $\bar{x}$, which is not a stationary point of $f + cP$. Let $\tilde{x}^{k+1}$ be the point that would be generated from x^k via the Armijo rule and let $\tilde{\alpha}^k$ be the corresponding stepsize. We have

$$
f(x^k) - f(x^{k+1}) \geq f(x^k) - f(\tilde{x}^{k+1}) \geq \sigma \tilde{\alpha}^k d^{k\prime} H^k d^k.
$$

By replacing α^k by $\tilde{\alpha}^k$ in the arguments of the earlier proof, we obtain a contradiction. This line of argument establishes that any stepsize rule that gives a larger reduction in the value of $f + cP$ at each iteration than the Armijo rule inherits its convergence properties, so it also proves the proposition for the limited minimization rule. **Q.E.D.**

Application to Constrained Optimization Problems

Given the inequality constrained problem

$$
\begin{aligned}
&\text{minimize } \ f(x) \\
&\text{subject to } \ g_j(x) \leq 0, \qquad j = 1, \ldots, r,
\end{aligned}
\tag{5.102}
$$

we can attempt its solution by using the linearization algorithm to minimize the corresponding exact penalty function $f+cP$ for a value of c that exceeds the threshold $\sum_{j=1}^r \mu_j^*$ (cf. Prop. 5.3.1).

There are a number of complex implementation issues here. One difficulty is that we may not know a suitable threshold value for c. Under these circumstances, a possible approach is to choose an initial value c^0 for c and increase it as necessary at each iteration k if the algorithm indicates that the current value c^k is inadequate. An important question is to decide on the conditions that would prompt an increase of c^k. The most common approach is based on trying to solve the quadratic program

$$\begin{aligned} &\text{minimize} \ \ \nabla f(x^k)'d + \tfrac{1}{2}d'H^k d \\ &\text{subject to} \ \ g_j(x^k) + \nabla g_j(x^k)'d \le 0, \qquad j = 1,\dots,r, \end{aligned} \tag{5.103}$$

which differs from the direction finding quadratic program

$$\begin{aligned} &\text{minimize} \ \ \nabla f(x^k)'d + \tfrac{1}{2}d'H^k d + c\xi \\ &\text{subject to} \ \ g_j(x^k) + \nabla g_j(x^k)'d \le \xi, \qquad j = 0,1,\dots,r, \end{aligned} \tag{5.104}$$

of the linearization method in that ξ has been set to zero. If program (5.103) has a feasible solution, then it must have a unique optimal solution d^k and at least one set of Lagrange multipliers $\mu_1^k,\dots,\mu_r^k$ (since its cost function is a strictly convex quadratic and its constraints are linear). It can then be verified by checking the corresponding optimality conditions (Exercise 5.3.3) that for all $c > \sum_{j=1}^r \mu_j^k$, the pair $\{d^k, \xi = 0\}$ is the optimal solution of the quadratic program (5.104). Thus the direction d^k can be used as a direction of descent for minimizing $f + c^k P$, where $\sum_{j=1}^r \mu_j^k$ provides an underestimate for the appropriate value for c^k. The penalty parameter may be updated by

$$c^k = \max\left\{c^{k-1}, \sum_{j=1}^r \mu_j^k + \gamma\right\},$$

where γ is some positive scalar. Note that by the optimality conditions of the quadratic program (5.103) we have approximately, for small $\|d^k\|$,

$$\nabla f(x^k) + \sum_{j=1}^r \mu_j^k \nabla g_j(x^k) \approx 0, \qquad \mu_j^k g_j(x^k) \approx 0, \quad j = 1,\dots,r,$$

so near convergence, the scalars μ_j^k are approximately equal to Lagrange multipliers of the constrained problem (5.102). This is consistent with the strategy of setting c^k at a somewhat higher value than $\sum_{j=1}^r \mu_j^k$. If on the other hand, the quadratic program (5.103) has no feasible solution, one can set $c^k = c^{k-1}$ and obtain a direction of descent d^k for the quadratic program (5.104).†

One of the drawbacks of this approach is that the value of the penalty parameter c^k may increase rapidly during the early stages of the algorithm, while during the final stage of the algorithm a much smaller value of c^k may be adequate. A large value of c^k results in very sharp corners of the surfaces of equal cost of the penalized cost $f + c^k P$ along the boundary of the constraint set, and can have a substantial adverse effect on the effectiveness of the stepsize procedure and thus on algorithmic progress (see Fig. 5.3.3). In this connection, it is interesting to note that if the system

$$g_j(x^k) + \nabla g_j(x^k)'d \le 0, \qquad j = 1, \ldots, r,$$

is feasible, then the direction d^k obtained from the quadratic program (5.103) is independent of c^k, while the stepsize α^k depends strongly on c^k. For this reason, it may be important to provide schemes that allow for the reduction of c^k if circumstances appear to be favorable. The details of this can become quite complicated and we refer to the book [Ber82a] for a discussion of some possibilities.

An important question relates to the choice of the matrices H^k. In unconstrained minimization, one tries to employ a stepsize $\alpha^k = 1$ together with matrices H^k that approximate the Hessian of the cost function at a solution. A natural analog for the constrained case would be to choose H^k close to the Hessian of the Lagrangian function

$$L(x, \mu) = f(x) + \mu' g(x),$$

evaluated at (x^*, μ^*); a justification for this is provided in the next section, where it is shown that the direction d^k calculated by the linearization algorithm, with this choice of H^k, can be viewed as a Newton step.

There are two difficulties relating to such an approach. The first is that $\nabla^2_{xx} L(x^*, \mu^*)$ may not be positive definite. Actually this is not as serious as might appear. As we discuss more fully in the next section, what is important is that H^k approximate closely $\nabla^2_{xx} L(x^*, \mu^*)$ only on the

---

† It is possible that because of the constraint nonlinearities the quadratic program (5.103) has no feasible solution. This will not happen if the constraint functions g_j are convex and the original inequality constrained problem (5.102) has at least one feasible solution, say $\bar{x}$; it can be seen that the vector $\bar{d}^k = \bar{x} - x^k$ is a feasible solution of the quadratic program (5.103) since

$$g_j(x^k) + \nabla g_j(x^k)'(\bar{x} - x^k) \le g_j(\bar{x}) \le 0$$

by Prop. B.3(a) of Appendix B. Usually, even if the constraints are nonconvex, the quadratic problem (5.103) is feasible, provided the constrained problem (5.102) is feasible.

subspace tangent to the active constraints. Under second order sufficiency assumptions on (x^*, μ^*), this can be done with positive definite H^k, since then $\nabla^2_{xx} L(x^*, \mu^*)$ is positive definite on this subspace.

The second difficulty relates to the fact that even if we were to choose H^k equal to the (generally unknown) matrix $\nabla^2_{xx} L(x^*, \mu^*)$ and even if this matrix is positive definite, it may happen that arbitrarily close to x^* a stepsize $\alpha^k = 1$ is not acceptable by the algorithm because it does not decrease the value of $f + cP$; this can happen even for very simple problems (see Exercise 5.3.9). The book [Ber82a] (p. 290) discusses this point in detail, and introduces modifications to the basic linearization algorithm that allow a superlinear convergence rate.

Extension to Equality Constraints

The development given earlier for inequality constraints can be extended to the case of additional equality constraints simply by converting each equality constraint $h_i(x) = 0$ to the two inequalities

$$h_i(x) \le 0, \qquad -h_i(x) \le 0.$$

For example, the direction finding quadratic program of the linearization method is

$$\begin{aligned} \text{minimize} \quad & \nabla f(x^k)'d + \tfrac{1}{2} d'H^k d + c\xi \\ \text{subject to} \quad & g_j(x^k) + \nabla g_j(x^k)'d \le \xi, \quad j = 0, 1, \ldots, r, \\ & \big| h_i(x^k) + \nabla h_i(x^k)'d \big| \le \xi, \quad i = 1, \ldots, m. \end{aligned}$$

This program yields a descent direction for the exact penalty function

$$f(x) + c \max\big\{0, g_1(x), \ldots, g_r(x), |h_1(x)|, \ldots, |h_m(x)|\big\},$$

and can be used as a basis for an algorithm similar to the one developed for inequality constraints.

5.3.3 Differentiable Exact Penalty Functions

We now discuss briefly differentiable exact penalty functions for the equality constrained problem

$$\begin{aligned} \text{minimize} \quad & f(x) \\ \text{subject to} \quad & h_i(x) = 0, \qquad i = 1, \ldots, m. \end{aligned} \tag{5.105}$$

We assume that f and h_i are twice continuously differentiable. Furthermore, we assume that the matrix $\nabla h(x)$ has rank m for all x, although much of the following analysis can be conducted assuming $\nabla h(x)$ has rank m in a suitable open subset of $\Re^n$. Motivated by the exact penalty function

$$\big\|\nabla_x L(x,\lambda)\big\|^2 + \big\|h(x)\big\|^2$$

discussed earlier, we consider the function

$$P_c(x,\lambda) = L(x,\lambda) + \tfrac{1}{2}\big\|W(x)\nabla_x L(x,\lambda)\big\|^2 + \tfrac{c}{2}\big\|h(x)\big\|^2, \tag{5.106}$$

where

$$L(x,\lambda) = f(x) + \lambda' h(x)$$

is the Lagrangian function, c is a positive parameter, and $W(x)$ is any continuously differentiable $m \times n$ matrix such that the $m \times m$ matrix $W(x)\nabla h(x)$ is nonsingular for all x.

The idea here is that by introducing the Lagrangian $L(x,\lambda)$ in the penalized cost P_c, we build a preference towards local minima rather than local maxima. The use of the matrix function $W(x)$ cannot be motivated easily, but will be justified by subsequent developments. Two examples of choices of $W(x)$ that turn out to be useful are

$$W(x) = \nabla h(x)', \tag{5.107}$$

$$W(x) = \big(\nabla h(x)'\nabla h(x)\big)^{-1}\nabla h(x)'. \tag{5.108}$$

The monograph [Ber82a] discusses in greater detail the role of the matrix $W(x)$ and also considers a different type of method whereby $W(x)$ is taken equal to the identity matrix.

Let us write $W(x)$ in the form

$$W(x) = \begin{pmatrix} w_1(x)' \\ \vdots \\ w_m(x)' \end{pmatrix},$$

where $w_i : \Re^n \mapsto \Re^n$ are some functions, and let $e_1, \ldots, e_m$ be the columns of the $m \times m$ identity matrix. It is then straightforward to verify that

$$\nabla_x P_c = \nabla_x L + \left(\nabla^2_{xx} L W' + \sum_{i=1}^m \nabla w_i \nabla_x L e_i'\right) W \nabla_x L + c\nabla h h, \tag{5.109}$$

$$\nabla_\lambda P_c = h + \nabla h' W' W \nabla_x L, \tag{5.110}$$

where all functions and gradients in the above expressions are evaluated at the typical pair (x,λ).

It can be seen that if (x^*, λ^*) is a local minimum-Lagrange multiplier pair of the original problem (5.105), then (x^*, λ^*) is also a stationary point of $P_c(x, \lambda)$, i.e.,

$$\nabla_x P_c(x^*, \lambda^*) = 0, \qquad \nabla_\lambda P_c(x^*, \lambda^*) = 0.$$

Under appropriate conditions, the reverse assertions are possible, namely that stationary points (x^*, λ^*) of $P_c(x, \lambda)$ satisfy the first order necessary conditions for the original constrained optimization problem. There are several results of this type, of which the following is typical. We outline the proof in Exercise 5.3.7 and we also refer to [Ber82a] for an extensive analysis.

Proposition 5.3.4: For every compact subset $X \times \Lambda$ of $\Re^{n+m}$ there exists a $\bar{c} > 0$ such that for all $c \geq \bar{c}$, every stationary point (x^*, λ^*) of P_c that belongs to $X \times \Lambda$ satisfies the first order necessary conditions

$$\nabla_x L(x^*, \lambda^*) = 0, \qquad \nabla_\lambda L(x^*, \lambda^*) = 0.$$

Differentiable Exact Penalty Functions Depending Only on x

One approach to minimizing $P_c(x, \lambda)$ is to first minimize it with respect to λ and then minimize it with respect to x. To simplify the subsequent formulas, let us focus on the function

$$W(x) = \big(\nabla h(x)'\nabla h(x)\big)^{-1}\nabla h(x)'$$

of Eq. (5.108). For this function, $W(x)\nabla h(x)$ is equal to the identity matrix and from Eq. (5.106) we have

$$P_c(x, \lambda) = f(x) + \lambda' h(x) + \tfrac{1}{2}\big\|W(x)\nabla f(x) + \lambda\big\|^2 + \tfrac{c}{2}\big\|h(x)\big\|^2. \tag{5.111}$$

We can minimize explicitly this function with respect to λ by setting

$$\nabla_\lambda P_c(x, \lambda) = h(x) + W(x)\nabla f(x) + \lambda = 0.$$

Substituting λ from this equation into Eq. (5.111), we obtain

$$\min_\lambda P_c(x, \lambda) = f(x) + \hat{\lambda}(x)' h(x) + \frac{c-1}{2}\big\|h(x)\big\|^2,$$

where

$$\hat{\lambda}(x) = -W(x)\nabla f(x).$$

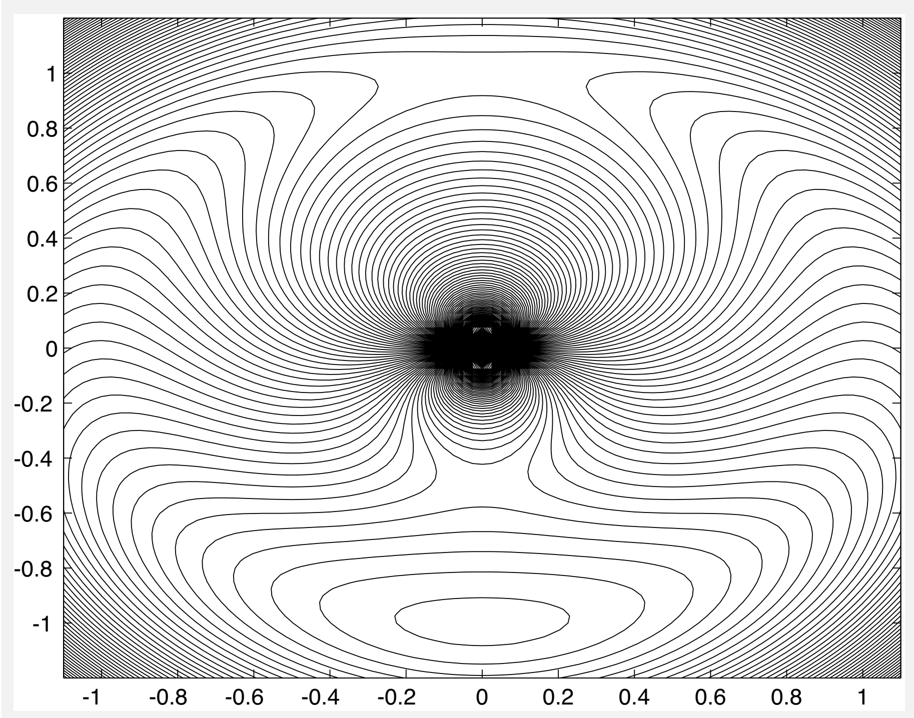

Figure 5.3.6. Equal cost surfaces of the differentiable exact penalty function $\hat{P}_c(x)$ for the two-dimensional problem where

$$f(x) = x_1, \qquad h(x) = x_1^2 + x_2^2 - 1$$

(cf. Figs. 5.3.2 and 5.3.3). The figure corresponds to $c = 2$. Note that there is a singularity at $(0, 0)$, which is a nonregular point at which $\hat{\lambda}(x)$ is undefined. The function $\hat{P}_c(x)$ takes arbitrarily large and arbitrarily small values sufficiently close to $(0, 0)$. This type of singularity can be avoided by using a modification of the exact penalty function (see Exercise 5.3.8).

Replacing $c - 1$ by c, it is seen that the function

$$\hat{P}_c(x) = f(x) + \hat{\lambda}(x)' h(x) + \frac{c}{2} \|h(x)\|^2 \tag{5.112}$$

is an exact penalty function, inheriting its properties from the exact penalty function $P_c(x, \lambda)$. Figure 5.3.6 illustrates the function $\hat{P}_c(x)$ for the same example problem that we used to illustrate the nondifferentiable exact penalty function $f + cP$ in Fig. 5.3.3. It can be seen that the two exact penalty functions $\hat{P}_c$ and $f + cP$ have quite different structures, including that $\hat{P}_c$ is not defined at nonregular points. For a detailed analysis of algorithms for minimization of $\hat{P}_c$ (including superlinearly converging Newton-like methods), the associated convergence and implementation issues, and extensions to inequality constraints we refer to the monograph [Ber82a] (Section 4.3.3).

EXERCISES

5.3.1

Consider a one-dimensional problem with two inequality constraints where $f(x) = 0$, $g_1(x) = -x$, $g_2(x) = 1 - x^2$. Show that for all c, $x = (1/2)(1 - \sqrt{5})$ and $x = 0$ are stationary points of $f + cP$, where P is the nondifferentiable exact penalty function (5.79)-(5.80), but are infeasible for the constrained problem. Plot $P(x)$ and discuss the behavior of the linearization method for this problem.

5.3.2

Let H be a positive definite symmetric matrix. Show that the pair (x^*, μ^*) satisfies the first order necessary conditions of Prop. 4.3.1 for the problem

$$\begin{aligned} &\text{minimize} \quad f(x) \\ &\text{subject to} \quad g_j(x) \le 0, \qquad j = 1, \ldots, r, \end{aligned}$$

if and only if $(0, \mu^*)$ is a global minimum-Lagrange multiplier pair of the quadratic program

$$\begin{aligned} &\text{minimize} \quad \nabla f(x^*)'d + \tfrac{1}{2}d'Hd \\ &\text{subject to} \quad g_j(x^*) + \nabla g_j(x^*)'d \le 0, \qquad j = 1, \ldots, r. \end{aligned}$$

(See [Ber82a], Section 4.1 for a solution.)

5.3.3

Show that if (d, μ) is a global minimum-Lagrange multiplier pair of the quadratic program

$$\begin{aligned} &\text{minimize} \quad \nabla f(x)'d + \tfrac{1}{2}d'Hd \\ &\text{subject to} \quad g_j(x) + \nabla g_j(x)'d \le 0, \qquad j = 1, \ldots, r, \end{aligned}$$

where H is positive definite symmetric, and

$$c \ge \sum_{j=1}^{r} \mu_j,$$

then $(d, \xi = 0, \bar{\mu})$ is a global minimum-Lagrange multiplier pair of the quadratic program

$$\begin{aligned} &\text{minimize} \quad \nabla f(x)'d + \tfrac{1}{2}d'Hd + c\xi \\ &\text{subject to} \quad (x, \xi) \in \Re^{n+1}, \qquad g_j(x) + \nabla g_j(x)'d \le \xi, \quad j = 0, 1, \ldots, r, \end{aligned}$$

where $\bar{\mu}_j = \mu_j$ for $j = 1, \ldots, r$, $\bar{\mu}_0 = c - \sum_{j=1}^{r} \mu_j$, and $g_0(x) \equiv 0$. (See [Ber82a], Section 4.1 for a solution.)

5.3.4

Show that if the pair (x^*, μ^*) satisfies the first order necessary conditions of Prop. 4.3.1 for the problem

$$\begin{aligned} &\text{minimize} \quad f(x) \\ &\text{subject to} \quad g_j(x) \le 0, \quad j = 1, \ldots, r, \end{aligned}$$

then x^* is a stationary point of $f + cP$ for all $c \ge \sum_{j=1}^{r} \mu_j^*$. *Hint*: Combine the results of Exercises 5.3.2 and 5.3.3.

5.3.5

Show that when the constraints are linear, the linearization method based on the quadratic program (5.103) is equivalent to one of the gradient projection methods of Section 3.3.

5.3.6

For the one-dimensional problem of minimizing $(1/6)x^3$ subject to $x = 0$, consider the differentiable exact penalty function $P_c(x, \lambda)$ of Eq. (5.106) with $W(x)$ given by Eq. (5.107) or Eq. (5.108). Show that it has two stationary points: the pairs $(0, 0)$ and $\big(c-1, (1-c^2)/2\big)$. Are both of these local minima of $P_c(x, \lambda)$? Discuss how your analysis is consistent with Prop. 5.3.4.

5.3.7

Prove Prop. 5.3.4. *Hint*: By Eq. (5.110), the condition $\nabla_\lambda P_c = 0$ at some point of $X \times \Lambda$ implies $W \nabla_x L = -(\nabla h' W')^{-1} h$. If at this point $\nabla_x P_c = 0$ also holds, we obtain after some calculation

$$\begin{aligned} 0 &= W \nabla_x P_c \\ &= \left\{ cW\nabla h - \left(I + W \left(\nabla_{xx}^2 L W' + \sum_{i=1}^{m} \nabla w_i \nabla_x L e_i' \right) \right) (\nabla h' W')^{-1} \right\} h. \end{aligned}$$

Show that there exists $\bar{c} > 0$ such that for all $c \ge \bar{c}$ and stationary points within $X \times \Lambda$, the matrix within braces is nonsingular, implying that at such points $h = 0$. Conclude that we also have $W \nabla_x L = 0$ so that from Eq. (5.109), $\nabla_x L = 0$.

5.3.8 (Dealing with Singularities [Ber82a], p. 215)

A difficulty with the penalty function $P_c(x, \lambda)$ of Eq. (5.106) is the assumption that the matrix $\nabla h(x)$ has rank m for all x. When this assumption is violated, the λ-dependent terms of P_c may be unbounded below. Furthermore, the function $\hat{P}_c$ of Eq. (5.112) is undefined at some points and singularities of the type shown in Fig. 5.3.3 at $x = 0$ may arise. To deal with this difficulty, introduce the following modified version of P_c:

$$P_{c,\tau}(x, \lambda) = L(x, \lambda) + \tfrac{1}{2}\big\|\nabla h(x)\nabla_x L(x, \lambda)\big\|^2 + \tfrac{c+\tau\|\lambda\|^2}{2}\big\|h(x)\big\|^2,$$

where τ is an additional positive parameter.

(a) Show that $P_{c,\tau}(x, \lambda)$ is bounded from below if the function $f(x)+(c/2)\big\|h(x)\big\|^2$ is bounded from below.

(b) Obtain a corresponding differentiable penalty function depending only on x, by minimizing $P_{c,\tau}(x, \lambda)$ with respect to λ.

(c) Plot the contours of this function for the problem of Fig. 5.3.6 and verify that the singularity exhibited in that figure does not occur.

5.3.9 (Maratos' Effect [Mar78])

This example illustrates a fundamental difficulty in attaining superlinear convergence when using the nondifferentiable exact penalty function for monitoring descent. (This difficulty does not arise for differentiable exact penalty functions; see [Ber82a], pp. 271-277.) Consider the problem

$$\begin{aligned} &\text{minimize} \ \ f(x) = x_1 \\ &\text{subject to} \ \ h(x) = x_1^2 + x_2^2 - 1 = 0, \end{aligned}$$

with optimal solution $x^* = (-1, 0)$ and Lagrange multiplier $\lambda^* = 1/2$ (see Figs. 5.3.2, 5.3.3, and 5.3.6). For any x, let (d, λ) be an optimal solution-Lagrange multiplier pair of the problem

$$\begin{aligned} &\text{minimize} \ \ \nabla f(x) + \tfrac{1}{2}d'\nabla_{xx}^2 L(x^*, \lambda^*)d \\ &\text{subject to} \ \ h(x) + \nabla h(x)'d = 0. \end{aligned}$$

(Note that d is the Newton direction; see also the next section.) Show that for all c,

$$f(x+d) + c\big|h(x+d)\big| - f(x) - c\big|h(x)\big| = \lambda h(x) - c\big|h(x)\big| + (c - \lambda^*)\|d\|^2.$$

Conclude that for $c > 2\lambda^*$, there are points x arbitrarily close to x^* for which the exact penalty function $f(x) + c\big|h(x)\big|$ is not reduced by a pure Newton step. (For a solution of the exercise and for a broader discussion of this phenomenon, see [Ber82a], p. 290.)

5.4 LAGRANGIAN METHODS

In this section we consider the direct solution of the system of necessary optimality conditions of the equality constrained problem

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

Thus we view the optimality conditions

$$\nabla f(x) + \nabla h(x)\lambda = 0, \qquad h(x) = 0, \tag{5.113}$$

as a system of $(n+m)$ nonlinear equations with $(n+m)$ unknowns, the vectors x and λ. We refer to this as the *Lagrangian system*, and we will aim to solve it with a variety of first and second order methods, called *Lagrangian methods*. We will also discuss supplementary schemes, based on merit function descent, which offer improved convergence guarantees. Moreover, we will provide alternative or modified forms of the optimality conditions to accommodate inequality constraints and approximations.

An important example of the Lagrangian approach is the primal-dual methodology for linear programming that we have discussed in Section 5.1.2. In that case, we considered solution of the system of primal and dual optimality conditions (5.16), Newton-like solution methods, and the use of the merit function (5.15) to enforce global convergence. Here our approach is more closely related to traditional lines of analysis of algorithms for solution of differentiable systems of nonlinear equations.

We will first consider Lagrangian algorithms for the system (5.113) that have the generic form

$$x^{k+1} = G(x^k, \lambda^k), \qquad \lambda^{k+1} = H(x^k, \lambda^k), \tag{5.114}$$

where $G : \Re^{n+m} \mapsto \Re^n$ and $H : \Re^{n+m} \mapsto \Re^n$ are continuously differentiable functions. Since this iteration can only converge to a pair (x^*, λ^*) such that

$$x^* = G(x^*, \lambda^*), \qquad \lambda^* = H(x^*, \lambda^*),$$

the functions G and H must be chosen so that local minima-Lagrange multiplier pairs satisfy the above equations.

We start with a first order method, which does not require second derivatives. We then consider Newton-like methods, and various ways to implement them. The difficulty with all these methods, as well as with most other methods for solving nonlinear systems of equations, is that they guarantee only local convergence, i.e., convergence from a starting point that is sufficiently close to a solution. To enlarge the region of convergence, it is necessary to use some type of line search based on the improvement of some merit function. While the existence of such a function for the Lagrangian system (5.113) is not obvious, we will see that a number of functions, such as the augmented Lagrangian function, the exact penalty functions of Section 5.3, and other functions can serve as the basis for globally convergent versions.

5.4.1 First Order Lagrangian Methods

The simplest Lagrangian method for solving the system of optimality conditions (5.113) is given by

$$x^{k+1} = x^k - \alpha \nabla_x L(x^k, \lambda^k), \tag{5.115}$$

$$\lambda^{k+1} = \lambda^k + \alpha h(x^k), \tag{5.116}$$

where L is the Lagrangian function

$$L(x, \lambda) = f(x) + \lambda' h(x)$$

and $\alpha > 0$ is a scalar stepsize. To motivate this method, consider the function

$$P(x, \lambda) = \tfrac{1}{2}\big\|\nabla_x L(x, \lambda)\big\|^2 + \tfrac{1}{2}\big\|h(x)\big\|^2.$$

This function is minimized at a local minimum-Lagrange multiplier pair, so it can be viewed as an exact penalty function (cf. the discussion of Section 5.3).

Let us consider the direction

$$d(x^k, \lambda^k) = \big(-\nabla_x L(x^k, \lambda^k), h(x^k)\big)$$

used in the first order iteration (5.115)-(5.116) and derive conditions under which it is a descent direction of the exact penalty function $P(x, \lambda)$. We have

$$\nabla P(x, \lambda) = \begin{pmatrix} \nabla^2_{xx} L(x, \lambda) \nabla_x L(x, \lambda) + \nabla h(x) h(x) \\ \nabla h(x)' \nabla_x L(x, \lambda) \end{pmatrix},$$

so that

$$\begin{aligned} d(x^k, \lambda^k)' \nabla P(x^k, \lambda^k) &= -\nabla_x L(x^k, \lambda^k)' \big(\nabla^2_{xx} L(x^k, \lambda^k) \nabla_x L(x^k, \lambda^k) \\ &\qquad + \nabla h(x^k) h(x^k)\big) + h(x^k)' \nabla h(x^k)' \nabla_x L(x^k, \lambda^k) \\ &= -\nabla_x L(x^k, \lambda^k)' \nabla^2_{xx} L(x^k, \lambda^k) \nabla_x L(x^k, \lambda^k). \end{aligned}$$

If the Hessian of the Lagrangian $\nabla^2_{xx} L(x^k, \lambda^k)$ is positive definite, we see that $d(x^k, \lambda^k)$ is a descent direction of the exact penalty function P [assuming that $\nabla_x L(x^k, \lambda^k) \neq 0$]. Note, however, that *positive definiteness of $\nabla^2_{xx} L(x^k, \lambda^k)$ is essential and is a stronger requirement than the second order sufficiency conditions of Section 4.2.*

To analyze the convergence of the first order iteration (5.115)-(5.116), we cannot quite use the global convergence methodology for descent methods of Chapters 1 and 2 because $d(x^k, \lambda^k)$ need not be a descent direction of the exact penalty function P when far from (x^*, λ^*). Thus we need some new tools, and to this end, we develop a general result on the local convergence of methods for solving systems of nonlinear equations. A pair (x^*, λ^*) is said to be a *point of attraction* of the iteration (5.114) if there exists an open set $S \subset \Re^{n+m}$ such that if $(x^0, \lambda^0) \in S$, then the sequence $\{(x^k, \lambda^k)\}$ generated by the iteration belongs to S and converges to (x^*, λ^*). The following proposition is very useful for our purposes.

Proposition 5.4.1: Let $G : \Re^{n+m} \mapsto \Re^n$ and $H : \Re^{n+m} \mapsto \Re^n$ be continuously differentiable functions. Assume that (x^*, λ^*) satisfies

$$x^* = G(x^*, \lambda^*), \qquad \lambda^* = H(x^*, \lambda^*),$$

and that all eigenvalues of the $(n+m) \times (n+m)$ matrix

$$R^* = \begin{pmatrix} \nabla_x G(x^*, \lambda^*) & \nabla_x H(x^*, \lambda^*) \\ \nabla_\lambda G(x^*, \lambda^*) & \nabla_\lambda H(x^*, \lambda^*) \end{pmatrix} \tag{5.117}$$

lie strictly within the unit circle of the complex plane. Then (x^*, λ^*) is a point of attraction of the iteration

$$x^{k+1} = G(x^k, \lambda^k), \qquad \lambda^{k+1} = H(x^k, \lambda^k), \tag{5.118}$$

and when the generated sequence $\{(x^k, \lambda^k)\}$ converges to (x^*, λ^*), the rate of convergence of $\|x^k - x^*\|$ and $\|\lambda^k - \lambda^*\|$ is linear.

Proof: Denote $y = (x, \lambda)$, $y^k = (x^k, \lambda^k)$, $y^* = (x^*, \lambda^*)$, and consider the function $M : \Re^{n+m} \mapsto \Re^{n+m}$ given by $M(y) = \big(G(x, \lambda), H(x, \lambda)\big)$. By the mean value theorem, we have for any two vectors y and $\tilde{y}$,

$$M(\tilde{y}) - M(y) = R'(\tilde{y} - y),$$

where R is the matrix having as ith column the gradient $\nabla M_i(\hat{y}^i)$ of the ith component of M evaluated at some vector $\hat{y}^i$ on the line segment connecting y and $\tilde{y}$. By taking $\tilde{y}$ and y sufficiently close to y^*, we can make R as close to the matrix R^* of Eq. (5.117) as desired, and therefore we can make the eigenvalues of the transpose R' lie within the unit circle [the eigenvalues of R and R' coincide by Prop. A.13(f) of Appendix A]. It follows from Prop. A.15 of Appendix A that there exists a norm $\|\cdot\|$ and an open sphere S with respect to that norm centered at (x^*, λ^*) such that, within S, the induced matrix norm of R' is less than $1 - \epsilon$ where ϵ is some positive scalar. Since

$$\big\|M(\tilde{y}) - M(y)\big\| \le \|R'\| \, \|\tilde{y} - y\|,$$

it follows that within the sphere S, the mapping M is a contraction as defined in Appendix A. The result then follows from the contraction mapping theorem (Prop. A.26 in Appendix A). **Q.E.D.**

We now prove the local convergence of the first order Lagrangian iteration (5.115)-(5.116).

Proposition 5.4.2: Assume that f and h are twice continuously differentiable, and let (x^*, λ^*) be a local minimum-Lagrange multiplier pair. Assume also that x^* is regular and that the matrix $\nabla^2_{xx} L(x^*, \lambda^*)$ is positive definite. Then there exists $\bar{\alpha} > 0$, such that for all $\alpha \in (0, \bar{\alpha}]$, (x^*, λ^*) is a point of attraction of iteration (5.115)-(5.116), and if the generated sequence $\{(x^k, \lambda^k)\}$ converges to (x^*, λ^*), then the rate of convergence of $\|x^k - x^*\|$ and $\|\lambda^k - \lambda^*\|$ is linear.

Proof: The proof consists of showing that, for α sufficiently small, the hypothesis of Prop. 5.4.1 is satisfied. Indeed for $\alpha > 0$, consider the mapping $M_\alpha : \Re^{n+m} \mapsto \Re^{n+m}$ defined by

$$M_\alpha(x, \lambda) = \begin{pmatrix} x - \alpha \nabla_x L(x, \lambda) \\ \lambda + \alpha \nabla_\lambda L(x, \lambda) \end{pmatrix}.$$

Clearly $(x^*, \lambda^*) = M_\alpha(x^*, \lambda^*)$, and we have

$$\nabla M_\alpha(x^*, \lambda^*)' = I - \alpha B, \tag{5.119}$$

where

$$B = \begin{pmatrix} \nabla^2_{xx} L(x^*, \lambda^*) & \nabla h(x^*) \\ -\nabla h(x^*)' & 0 \end{pmatrix}. \tag{5.120}$$

We will show that the real part of each eigenvalue of B is strictly positive, and then the result will follow from Eq. (5.119) by using Prop. 5.4.1. For any complex vector y, denote by $\hat{y}$ its complex conjugate, and for any complex number γ, denote by $Re(\gamma)$ its real part. Let β be an eigenvalue of B, and let $(z, w) \neq 0$ be a corresponding eigenvector where z and w are complex vectors of dimension n and m, respectively. We have

$$Re\left\{(\hat{z}' \quad \hat{w}') B \begin{pmatrix} z \\ w \end{pmatrix}\right\} = Re\left\{\beta (\hat{z}' \quad \hat{w}') \begin{pmatrix} z \\ w \end{pmatrix}\right\} = Re(\beta)\big(\|z\|^2 + \|w\|^2\big), \tag{5.121}$$

while at the same time, by using Eq. (5.120),

$$Re\left\{(\hat{z}' \quad \hat{w}') B \begin{pmatrix} z \\ w \end{pmatrix}\right\} = Re\big\{\hat{z}' \nabla^2_{xx} L(x^*, \lambda^*) z + \hat{z}' \nabla h(x^*) w - \hat{w}' \nabla h(x^*)' z\big\}. \tag{5.122}$$

Since for any real $n \times m$ matrix Q, we have

$$Re\{\hat{z}' Q' w\} = Re\{\hat{w}' Q z\},$$

it follows from Eqs. (5.121) and (5.122) that

$$Re\{\hat{z}'\nabla^2_{xx}L(x^*,\lambda^*)z\} = Re\left\{(\hat{z}' \quad \hat{w}')B\begin{pmatrix} z \\ w \end{pmatrix}\right\} = Re(\beta)\big(\|z\|^2 + \|w\|^2\big). \tag{5.123}$$

Since for any positive definite matrix A, we have

$$Re\{\hat{z}'Az\} > 0, \qquad \forall\ z \neq 0,$$

it follows from Eq. (5.123) and the positive definiteness assumption on $\nabla^2_{xx}L(x^*,\lambda^*)$ that either $Re(\beta) > 0$ or else $z = 0$. But if $z = 0$, the equation

$$B\begin{pmatrix} z \\ w \end{pmatrix} = \beta\begin{pmatrix} z \\ w \end{pmatrix}$$

yields

$$\nabla h(x^*)w = 0.$$

Since $\nabla h(x^*)$ has rank m, it follows that $w = 0$. This contradicts our earlier assumption that $(z, w) \neq 0$. Consequently, we must have $Re(\beta) > 0$. **Q.E.D.**

We note that by appropriately scaling the vectors x and λ, we can show that the result of Prop. 5.4.2 holds also for the more general iteration

$$x^{k+1} = x^k - \alpha D\nabla_x L(x^k, \lambda^k), \tag{5.124}$$

$$\lambda^{k+1} = \lambda^k + \alpha E h(x^k), \tag{5.125}$$

where D and E are any positive definite symmetric matrices of appropriate dimension [reduce the preceding iteration to an iteration of the form (5.115)-(5.116) using a change of variables $x = D^{1/2}y$ and $\lambda = E^{1/2}\mu$; cf. the discussion of Section 1.3.2].

Augmented Lagrangian Convexification

We noted that positive definiteness of $\nabla^2_{xx}L(x^*,\lambda^*)$ is necessary for the validity of the Lagrangian method (5.115)-(5.116) and its scaled version (5.124)-(5.125). This essentially requires that the problem has a "locally convex" structure in the neighborhood of x^*. On the other hand when this structure is not present, we can remedy the situation by convexification, through the use of an augmented Lagrangian of the form

$$L_c(x,\lambda) = f(x) + \lambda' h(x) + \frac{c}{2}\big\|h(x)\big\|^2. \tag{5.126}$$

In particular, we may apply the method to the equivalent problem

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

where c is chosen sufficiently large to ensure that the corresponding matrix

$$\nabla^2_{xx} L_c(x^*, \lambda^*) = \nabla^2_{xx} L(x^*, \lambda^*) + c\nabla h(x^*)\nabla h(x^*)'$$

is positive definite [assuming of course that (x^*, λ^*) satisfy the second order sufficiency conditions so that the augmented Lagrangian theory applies; cf. Lemma 4.2.1 and the subsequent discussion].

The method (5.124)-(5.125) applied to the preceding equivalent problem takes the form

$$x^{k+1} = x^k - \alpha D \nabla_x L_c(x^k, \lambda^k), \tag{5.127}$$

$$\lambda^{k+1} = \lambda^k + \alpha E h(x^k), \tag{5.128}$$

where D and E are positive definite symmetric matrices of appropriate dimension. By using Prop. 5.4.2, we see that this method converges linearly to (x^*, λ^*), assuming that the second order sufficiency conditions are satisfied, c is sufficiently large to ensure that $\nabla^2_{xx} L_c(x^*, \lambda^*)$ is positive definite, the starting pair (x^0, λ^0) is sufficiently close to (x^*, λ^*), and α is sufficiently small.

Lagrangian Method in the Space of Primal Variables

An important observation for our purposes is that given a good approximation of a local minimum x^* that is a regular point [$\nabla h(x^*)$ has rank m], we can obtain analytically a good approximation of the associated Lagrange multiplier λ^*. One way to do this is to use the function $\hat{\lambda}$ defined by

$$\hat{\lambda}(x) = \big(\nabla h(x)'\nabla h(x)\big)^{-1}\big(h(x) - \nabla h(x)'\nabla f(x)\big), \tag{5.129}$$

for all x such that $\nabla h(x)$ has rank m. Indeed, by multiplying the necessary condition $\nabla f(x^*)+\nabla h(x^*)\lambda^* = 0$ with $\nabla h(x^*)'$, we obtain $\nabla h(x^*)'\nabla f(x^*)+\nabla h(x^*)'\nabla h(x^*)\lambda^* = 0$, so that

$$\lambda^* = -\big(\nabla h(x^*)'\nabla h(x^*)\big)^{-1}\nabla h(x^*)'\nabla f(x^*).$$

Thus, using the fact $h(x^*) = 0$ in Eq. (5.129), we obtain $\hat{\lambda}(x^*) = \lambda^*$. Since $\hat{\lambda}(\cdot)$ is a continuous function, it follows that $\hat{\lambda}(x)$ is near λ^* if x is near x^*.

One benefit of this observation is to alleviate the requirement for a good initial choice of (x^0, λ^0) in the preceding Lagrangian methods (cf. Prop. 5.4.2). If a good initial choice x^0 is available, we can obtain a good initial choice λ^0 from $\lambda^0 = \hat{\lambda}(x^0)$.

Carrying this idea further, we may consider a Lagrangian method where λ^k is taken to be equal to $\hat{\lambda}(x^k)$ rather than updated according to Eq. (5.116) or (5.125), leading to the algorithm

$$x^{k+1} = x^k - \alpha \nabla_x L\big(x^k, \hat{\lambda}(x^k)\big), \tag{5.130}$$

which iterates exclusively within the space of the vector x. To show local convergence to x^* of this iteration for sufficiently small stepsize α, it is necessary that the matrix

$$I - \alpha G(x^*)$$

where

$$G(x) = \nabla\big(\nabla_x L\big(x, \hat{\lambda}(x)\big)\big),$$

has eigenvalues strictly within the unit circle (cf. Prop. 5.4.1).

Indeed, remarkably, it can be shown that if (x^*, λ^*) satisfy the second order sufficiency conditions of Prop. 4.2.1 (in addition to x^* being a regular point), the eigenvalues of $G(x^*)$ are all real-valued and positive, so the eigenvalues of $I - \alpha G(x^*)$ lie within the unit circle for sufficiently small α. The proof is fairly complicated, and is given in Prop. 4.26 of [Ber82a]. A superlinearly converging Newton-like Lagrangian method that operates exclusively within the space of x is also described in [Ber82a] (Prop. 4.27).

An interesting observation is that [assuming that $\nabla h(x^k)$ has rank m] we can obtain both $\hat{\lambda}(x^k)$ and $\nabla_x L\big(x^k, \hat{\lambda}(x^k)\big)$ by solving the quadratic program

$$\begin{aligned} &\text{minimize} \quad \nabla f(x^k)'d + \tfrac{1}{2}\|d\|^2 \\ &\text{subject to} \quad h(x^k) + \nabla h(x^k)'d = 0. \end{aligned} \tag{5.131}$$

Indeed the optimality conditions for this program are

$$\nabla f(x^k) + \nabla h(x^k)\lambda + d = 0, \qquad h(x^k) + \nabla h(x^k)'d = 0,$$

and it can be seen that the unique Lagrange multiplier vector is $\hat{\lambda}(x^k)$, while the unique optimal solution is

$$d(x^k) = -\nabla_x L\big(x^k, \hat{\lambda}(x^k)\big).$$

As a special case we note that if h is linear and x^k is a feasible point, $-\nabla_x L\big(x^k, \hat{\lambda}(x^k)\big)$ is equal to $-\nabla f(x^k)$ projected onto the feasible set, so the iteration reduces to a gradient projection iteration. In the more general case, where h is linear but x^k is infeasible, $-\nabla_x L\big(x^k, \hat{\lambda}(x^k)\big)$ has two components; one component is $-\nabla f(x^k)$ projected onto the feasible set and aims at cost function reduction, while the other component is orthogonal to the feasible set, and aims at infeasibility reduction (for sufficiently small α). Note also that based on the quadratic programming implementation (5.131), the method is related to the linearization method (5.103) of Section 5.3.2.

A noteworthy fact here is that *positive definiteness of the matrix $\nabla^2_{xx} L(x^*, \lambda^*)$ is not required*. Convergence is guaranteed assuming just the second order sufficiency conditions of Prop. 4.2.1 and regularity of x^*. To get a sense of this, we note that *the iteration (5.130) when applied to the equivalent problem*

$$\begin{aligned} &\text{minimize} \quad f_c(x) \overset{\text{def}}{=} f(x) + \frac{c}{2}\big\|h(x)\big\|^2 \\ &\text{subject to} \quad h(x) = 0, \end{aligned}$$

is independent of the value of c, and embodies any desired amount of augmented Lagrangian convexification. Indeed it can be seen that the solution of the quadratic program (5.131) is not affected if ∇f is replaces by ∇f_c, since subject to the constraint $h(x^k) + \nabla h(x^k)'d = 0$, the inner products $\nabla f(x^k)'d$ and $\nabla f_c(x^k)'d$ differ by the constant $c\|h(x^k)\|^2$.

Let us finally note that the iteration (5.130) involves the calculation of $\hat{\lambda}(x^k)$, which may be significant extra overhead, particularly if h is nonlinear and/or the number of constraints m is large. Under favorable conditions, however, the method is viable and applies to problems that do not have the "locally convex" structure [positive definiteness of $\nabla^2_{xx}L(x^*, \lambda^*)$], which is required for the Lagrangian iteration (5.115)-(5.116).

Decomposition and Parallelization in Separable Problems

The Lagrangian methods and their variations in this section are well-suited for separable problems of the form

$$\begin{aligned} &\text{minimize} \quad \sum_{i=1}^{n} f_i(x_i) \\ &\text{subject to} \quad \sum_{i=1}^{n} h_{ij}(x_i) = 0, \quad j = 1, \ldots, m, \end{aligned} \tag{5.132}$$

where $f_i : \Re \mapsto \Re$ and $h_{ij} : \Re \mapsto \Re$ are twice continuously differentiable scalar functions. Problems of this type arise naturally in many contexts and they will be discussed in greater detail in the context of convex programming in Section 6.1.5, and in a variety of algorithmic contexts in Chapter 7.

Here we wish to point out the mechanism by which Lagrangian methods can exploit the structure of separable problems. Indeed the Lagrangian function of the problem takes the form

$$L(x, \lambda) = \sum_{i=1}^{n} \left\{ f_i(x_i) + \sum_{j=1}^{m} \lambda_j h_{ij}(x_i) \right\},$$

and is separable with respect to x_i. As a result, the Lagrangian method (5.115)-(5.116) takes the form

$$x_i^{k+1} = x_i^k - \alpha \left(\frac{\partial f_i(x_i^k)}{\partial x_i} + \sum_{j=1}^{m} \lambda_j^k \frac{\partial h_{ij}(x_i^k)}{\partial x_i} \right), \qquad i = 1, \ldots, n, \tag{5.133}$$

$$\lambda_j^{k+1} = \lambda_j^k + \alpha \sum_{i=1}^{n} h_{ij}(x_i^k), \qquad j = 1, \ldots, m. \tag{5.134}$$

Note that the iteration (5.133) decomposes with respect to the coordinates x_i, and is well-suited for parallel computation. For example, one may consider a parallel computing system with n processors, each updating a single scalar coordinate x_i according to iteration (5.133), and another (central) processor updating the multiplier vector λ according to Eq. (5.134). The ith processor communicates with the central processor, sending its current values x_i^k or $h_{ij}(x_i^k)$, $j = 1, \ldots, m$, while receiving from the central processor the current value of λ^k.

For a small enough stepsize α, this parallel algorithmic process is convergent under the conditions of Prop. 5.4.2 [including the requirement that the Hessian of the Lagrangian $\nabla^2_{xx} L(x^*, \lambda^*)$ is positive definite]. Also, a certain amount of asynchronism can be allowed into the algorithm, based on the totally and partially asynchronous guidelines of Section 2.5. Moreover, the iteration (5.130) admits similar parallelization.

Let us finally note an alternative method for separable problems, which uses a Lagrangian minimization in place of the gradient iteration (5.133). The method also requires that the Hessian of the Lagrangian $\nabla^2_{xx} L(x^*, \lambda^*)$ is positive definite, and has the form

$$x_i^{k+1} = \arg\min_{x_i} \left\{ f_i(x_i) + \sum_{j=1}^{m} \lambda_j^k h_{ij}(x_i) \right\}, \qquad i = 1, \ldots, n,$$

$$\lambda_j^{k+1} = \lambda_j^k + \alpha \sum_{i=1}^{n} h_{ij}(x_i^k), \qquad j = 1, \ldots, m,$$

where the minimization is assumed to be local in a neighborhood of x^*. This method can be understood by viewing it as a method of multipliers for the problem

$$\begin{aligned} &\text{minimize} \ \ f(x) - \frac{\alpha}{2} \big\| h(x) \big\|^2 \\ &\text{subject to} \ \ h(x) = 0. \end{aligned}$$

In particular, for convergence, α must not exceed twice the minimum eigenvalue of $\nabla h(x^*)' \big(\nabla^2_{xx} L(x^*, \lambda^*) \big)^{-1} \nabla h(x^*)$; see Exercise 5.2.11.

5.4.2 Newton-Like Methods for Equality Constraints

We will now turn to second order methods for solving the Lagrangian system

$$\nabla f(x) + \nabla h(x) \lambda = 0, \qquad h(x) = 0,$$

by viewing it as the vector equation

$$\nabla L(x, \lambda) = 0.$$

Newton's method for this equation is

$$x^{k+1} = x^k + \Delta x^k, \qquad \lambda^{k+1} = \lambda^k + \Delta \lambda^k, \tag{5.135}$$

where $(\Delta x^k, \Delta\lambda^k) \in \Re^{n+m}$ is obtained by solving the system

$$\nabla^2 L(x^k, \lambda^k) \begin{pmatrix} \Delta x^k \\ \Delta\lambda^k \end{pmatrix} = -\nabla L(x^k, \lambda^k), \tag{5.136}$$

the linearized version of the optimality condition $\nabla L(x, \lambda) = 0$.

We say that (x^{k+1}, λ^{k+1}) is *well-defined* by the Newton iteration (5.135)-(5.136) if the matrix $\nabla^2 L(x^k, \lambda^k)$ is invertible. Note that if x^* is a local minimum that is regular and together with a Lagrange multiplier λ^* satisfies the second order sufficiency condition of Prop. 4.2.1, then $\nabla^2 L(x^*, \lambda^*)$ is invertible; this was shown as part of the proof of the sensitivity theorem (Prop. 4.2.2). As a result, $\nabla^2 L(x, \lambda)$ is invertible in a neighborhood of (x^*, λ^*), and within this neighborhood, points generated by the Newton iteration are well-defined. In the subsequent discussion, when stating various local convergence properties of the Newton iteration in connection with such a pair, we implicitly restrict the iteration within a neighborhood where it is well-defined.

The local convergence properties of the method can be inferred from the results of Section 1.4. For purposes of convenient reference, we provide the corresponding result in the following proposition.

Proposition 5.4.3: Let x^* be a strict local minimum that is regular and satisfies together with a corresponding Lagrange multiplier vector λ^* the second order sufficiency conditions of Prop. 4.2.1. Then (x^*, λ^*) is a point of attraction of the Newton iteration (5.135)-(5.136). Furthermore, if the generated sequence converges to (x^*, λ^*), the rate of convergence of $\big\{\|(x^k, \lambda^k) - (x^*, \lambda^*)\|\big\}$ is superlinear (at least order two if $\nabla^2 f$ and $\nabla^2 h_i$, $i = 1, \ldots, m$, are Lipschitz continuous in a neighborhood of x^*).

Proof: Use Prop. 1.4.1 of Section 1.4. **Q.E.D.**

The Newton iteration (5.135)-(5.136) has a rich structure that can be used to provide interesting implementations, three of which we discuss below.

A First Implementation of Newton's Method

Let us write the gradient and Hessian of the Lagrangian function as

$$\nabla L(x^k, \lambda^k) = \begin{pmatrix} \nabla_x L(x^k, \lambda^k) \\ h(x^k) \end{pmatrix}, \qquad \nabla^2 L(x^k, \lambda^k) = \begin{pmatrix} H^k & N^k \\ N^{k\prime} & 0 \end{pmatrix},$$

where

$$H^k = \nabla^2_{xx} L(x^k, \lambda^k), \qquad N^k = \nabla h(x^k).$$

Thus, the Newton system (5.136) takes the form

$$\begin{pmatrix} H^k & N^k \\ N^{k\prime} & 0 \end{pmatrix} \begin{pmatrix} \Delta x^k \\ \Delta \lambda^k \end{pmatrix} = - \begin{pmatrix} \nabla_x L(x^k, \lambda^k) \\ h(x^k) \end{pmatrix}. \tag{5.137}$$

Let us assume that *H^k is invertible and N^k has rank m.* Then we can provide a more explicit expression for the Newton iteration. Indeed the Newton system (5.137) can be written as

$$H^k \Delta x^k + N^k \Delta \lambda^k = -\nabla_x L(x^k, \lambda^k), \tag{5.138}$$

$$N^{k\prime} \Delta x^k = -h(x^k). \tag{5.139}$$

By multiplying the first equation with $N^{k\prime}(H^k)^{-1}$ and by using the second equation, it follows that

$$-h(x^k) + N^{k\prime}(H^k)^{-1} N^k \Delta \lambda^k = -N^{k\prime}(H^k)^{-1} \nabla_x L(x^k, \lambda^k).$$

Since N^k has rank m, the matrix $N^{k\prime}(H^k)^{-1}N^k$ is nonsingular, and we obtain

$$\lambda^{k+1} - \lambda^k = \Delta \lambda^k = \big(N^{k\prime}(H^k)^{-1}N^k\big)^{-1}\big(h(x^k) - N^{k\prime}(H^k)^{-1}\nabla_x L(x^k, \lambda^k)\big). \tag{5.140}$$

We have

$$\begin{aligned} \nabla_x L(x^k, \lambda^k) &= \nabla f(x^k) + N^k \lambda^k \\ &= \nabla f(x^k) + N^k \lambda^{k+1} - N^k \Delta \lambda^k \\ &= \nabla_x L(x^k, \lambda^{k+1}) - N^k \Delta \lambda^k, \end{aligned}$$

and by using this equation to substitute $\nabla_x L(x^k, \lambda^k)$ in Eqs. (5.138) and (5.140), we finally obtain the two-step iteration

$$\lambda^{k+1} = \big(N^{k\prime}(H^k)^{-1}N^k\big)^{-1}\big(h(x^k) - N^{k\prime}(H^k)^{-1}\nabla f(x^k)\big), \tag{5.141}$$

$$x^{k+1} = x^k - (H^k)^{-1} \nabla_x L(x^k, \lambda^{k+1}). \tag{5.142}$$

This is a first implementation of the Newton iteration (under the assumption that H^k is invertible and N^k has rank m). It has the advantage that it requires the solution of systems of dimension at most n [as opposed to $n + m$, which is the dimension of $\nabla^2 L(x^k, \lambda^k)$].

An Implementation of Newton's Method Based on Augmented Lagrangian Functions

We will now derive another way to write the system of equations (5.138)-(5.139). It is based on the observation that, for every scalar c, we have from Eq. (5.139),

$$cN^k N^{k\prime} \Delta x = -cN^k h(x^k),$$

which added to Eq. (5.138), yields

$$(H^k + cN^kN^{k\prime})\Delta x^k + N^k\big(\Delta\lambda^k + ch(x^k)\big) = -\nabla_x L(x^k,\lambda^k). \tag{5.143}$$

Equivalently, by writing $\Delta\lambda^k = \lambda^{k+1} - \lambda^k$,

$$(H^k + cN^kN^{k\prime})\Delta x^k = -\nabla_x L\big(x^k, \lambda^{k+1} + ch(x^k)\big) = -\nabla_x L_c(x^k,\lambda^{k+1}), \tag{5.144}$$

where L_c is the augmented Lagrangian function

$$L_c(x,\lambda) = L(x,\lambda) + \frac{c}{2}\big\|h(x)\big\|^2.$$

Also if $(H^k + cN^kN^{k\prime})^{-1}$ exists, by multiplying Eq. (5.143) with $N^{k\prime}(H^k + cN^kN^{k\prime})^{-1}$, we obtain

$$\begin{aligned} N^{k\prime}&(H^k + cN^kN^{k\prime})^{-1}N^k\big(\Delta\lambda^k + ch(x^k)\big) \\ &= -N^{k\prime}\Delta x^k - N^{k\prime}(H^k + cN^kN^{k\prime})^{-1}\nabla_x L(x^k,\lambda^k), \end{aligned}$$

which by writing $\Delta\lambda^k = \lambda^{k+1} - \lambda^k$ and $N^{k\prime}\Delta x^k = -h(x^k)$ [cf. Eq. (5.139)], yields

$$\begin{aligned} N^{k\prime}&(H^k + cN^kN^{k\prime})^{-1}N^k\big(\lambda^{k+1} - \lambda^k + ch(x^k)\big) \\ &= h(x^k) - N^{k\prime}(H^k + cN^kN^{k\prime})^{-1}\nabla_x L(x^k,\lambda^k), \end{aligned}$$

or by using the fact $\nabla_x L(x^k,\lambda^k) = \nabla f(x^k) + N^k\lambda^k$,

$$\begin{aligned} N^{k\prime}&(H^k + cN^kN^{k\prime})^{-1}N^k\big(\lambda^{k+1} + ch(x^k)\big) \\ &= h(x^k) - N^{k\prime}(H^k + cN^kN^{k\prime})^{-1}\nabla f(x^k). \end{aligned} \tag{5.145}$$

Thus, from Eqs. (5.144) and (5.145), we obtain the following equivalent form of Newton's method

$$\begin{aligned} \lambda^{k+1} = -ch(x^k) + &\big(N^{k\prime}(H^k + cN^kN^{k\prime})^{-1}N^k\big)^{-1} \\ &\big(h(x^k) - N^{k\prime}(H^k + cN^kN^{k\prime})^{-1}\nabla f(x^k)\big), \end{aligned} \tag{5.146}$$

$$x^{k+1} = x^k - (H^k + cN^kN^{k\prime})^{-1}\nabla_x L_c(x^k,\lambda^{k+1}). \tag{5.147}$$

An advantage that this implementation may offer over the one of Eqs. (5.141)-(5.142) (which corresponds to $c = 0$) is that the matrix H^k may not be invertible while $H^k + cN^kN^{k\prime}$ may be invertible for some values of c. For example, for c sufficiently large, we have that $H^k + cN^kN^{k\prime}$ is not only invertible but also positive definite near (x^*,λ^*) (cf. Lemma 4.2.1 in Section 4.2). An additional benefit of this property is that it allows us to differentiate between local minima and local maxima (near a local maximum-Lagrange multiplier pair, $H^k + cN^kN^{k\prime}$ is not likely to be positive definite for any positive value of c). Note that positive definiteness of $H^k + cN^k N^{k'}$ can be easily detected if the Cholesky factorization method is used for solving the various linear systems of equations in Eqs. (5.146) and (5.147) (cf. the discussion of Section 1.4).

Another property, which is particularly useful for enlarging the region of convergence of Newton's method (see Section 5.4.3), can be inferred from Eq. (5.147): if c is large enough so that $H^k + cN^k N^{k'}$ is positive definite, $x^{k+1} - x^k$ is a descent direction of the augmented Lagrangian function $L_c(\cdot, \lambda^{k+1})$ at x^k.

An Implementation of Newton's Method Based on Quadratic Programming

We will now derive a third implementation of Newton's method. It is based on the observation that the Newton system (5.138)-(5.139) is written as

$$\nabla f(x^k) + H^k \Delta x^k + N^k \lambda^{k+1} = 0, \qquad h(x^k) + N^{k'} \Delta x^k = 0,$$

which are the necessary optimality conditions for $(\Delta x^k, \lambda^{k+1})$ to be a global minimum-Lagrange multiplier pair of the quadratic program

$$\begin{aligned} &\text{minimize} \ \ \nabla f(x^k)' \Delta x + \tfrac{1}{2} \Delta x' H^k \Delta x \\ &\text{subject to} \ \ h(x^k) + N^{k'} \Delta x = 0. \end{aligned} \tag{5.148}$$

Thus we can obtain $(\Delta x^k, \lambda^{k+1})$ by solving this problem.

This implementation is not particularly useful for practical purposes but provides an interesting connection with the linearization method of Section 5.3. This connection can be made more explicit by noting that the solution Δx^k of the quadratic program (5.148) is unaffected if H^k is replaced by any matrix of the form $H^k + cN^k N^{k'}$, where c is a scalar, thereby obtaining the program

$$\begin{aligned} &\text{minimize} \ \ \nabla f(x^k)' \Delta x + \tfrac{1}{2} \Delta x' (H^k + cN^k N^{k'}) \Delta x \\ &\text{subject to} \ \ h(x^k) + N^{k'} \Delta x = 0. \end{aligned} \tag{5.149}$$

To see that problems (5.148) and (5.149) have the same solution Δx^k, simply note that they have the same constraints while their cost functions differ by the term $c\Delta x' N^k N^{k'} \Delta x$, which is equal to $c\big\|h(x^k)\big\|^2$ and is therefore constant. Near a local minimum-Lagrange multiplier pair (x^*, λ^*) satisfying the sufficiency conditions, we have that $H^k + cN^k N^{k'}$ is positive definite if c is sufficiently large (Lemma 4.2.1 in Section 4.2), and the quadratic program (5.149) is positive definite.

We see therefore that, under these circumstances, the Newton iteration can be viewed in effect as a special case of the linearization method of Section 5.3 with a constant unity stepsize, and scaling matrix

$$\bar{H}^k = H^k + cN^k N^{k'},$$

where c is any scalar for which $\bar{H}^k$ is positive definite.

Merit Functions and Descent Properties of Newton's Method

Since we would like to improve the global convergence properties of Newton's method, it is interesting to look for appropriate merit functions, i.e., functions for which $(x^{k+1} - x^k)$ is a descent direction at x^k or $(x^{k+1} - x^k, \lambda^{k+1} - \lambda^k)$ is a descent direction at (x^k, λ^k). By this, we mean functions F such that for a sufficiently small positive scalar $\bar{\alpha}$, we have

$$F\big(x^k + \alpha(x^{k+1} - x^k)\big) < F(x^k), \qquad \forall\, \alpha \in (0, \bar{\alpha}],$$

or

$$F\big(x^k + \alpha(x^{k+1} - x^k), \lambda^k + \alpha(\lambda^{k+1} - \lambda^k)\big) < F(x^k, \lambda^k), \qquad \forall\, \alpha \in (0, \bar{\alpha}].$$

The following proposition shows that there are several possible merit functions.

Proposition 5.4.4: (Merit Functions for Lagrangian Methods) Let x^* be a local minimum that is a regular point and satisfies together with a corresponding Lagrange multiplier vector λ^* the second order sufficiency conditions of Prop. 4.2.1. There exists a neighborhood S of (x^*, λ^*) such that if $(x^k, \lambda^k) \in S$ and $x^k \neq x^*$, then (x^{k+1}, λ^{k+1}) is well-defined by the Newton iteration and the following hold:

(a) There exists a scalar $\bar{c}$ such that for all $c \geq \bar{c}$, the vector $(x^{k+1} - x^k)$ is a descent direction at x^k for the exact penalty function

$$f(x) + c \max_{i=1,\ldots,m} \big|h_i(x)\big|. \tag{5.150}$$

(b) The vector $(x^{k+1} - x^k, \lambda^{k+1} - \lambda^k)$ is a descent direction at (x^k, λ^k) for the exact penalty function

$$P(x, \lambda) = \tfrac{1}{2}\big\|\nabla L(x, \lambda)\big\|^2.$$

Furthermore, given any scalar $r > 0$, there exists a $\delta > 0$ such that if

$$\big\|(x^k - x^*, \lambda^k - \lambda^*)\big\| < \delta,$$

we have

$$P(x^{k+1}, \lambda^{k+1}) \leq r P(x^k, \lambda^k). \tag{5.151}$$

(c) For every scalar c such that $H^k + cN^kN^{k\prime}$ is positive definite, the vector $(x^{k+1} - x^k)$ is a descent direction at x^k of the augmented Lagrangian function $L_c(\cdot, \lambda^{k+1})$.

Proof: (a) Take $\bar{c} > 0$ sufficiently large and a neighborhood S of (x^*, λ^*), which is sufficiently small, so that for $(x^k, \lambda^k) \in S$, the matrix $H^k + \bar{c} N^k N^{k'}$ is positive definite. Since Δx^k is the solution of the quadratic program (5.149), it follows from Prop. 5.3.2 that if $x^k \neq x^*$, then Δx^k is a descent direction of the exact penalty function (5.150) for all $c \geq \bar{c}$.

(b) We have

$$\begin{pmatrix} x^{k+1} - x^k \\ \lambda^{k+1} - \lambda^k \end{pmatrix} = -\nabla^2 L(x^k, \lambda^k)^{-1} \nabla L(x^k, \lambda^k)$$

and

$$\nabla P(x^k, \lambda^k) = \nabla^2 L(x^k, \lambda^k) \nabla L(x^k, \lambda^k),$$

so

$$\big((x^{k+1} - x^k)', (\lambda^{k+1} - \lambda^k)' \big) \nabla P(x^k, \lambda^k) = -\big\| \nabla L(x^k, \lambda^k) \big\|^2 < 0,$$

and the descent property follows.

From Prop. 1.4.1, we have that, given any $\bar{r} > 0$, there exists a $\bar{\delta} > 0$ such that for $\big\|(x^k - x^*, \lambda^k - \lambda^*)\big\| < \bar{\delta}$, we have

$$\big\|(x^{k+1} - x^*, \lambda^{k+1} - \lambda^*)\big\| \leq \bar{r} \big\|(x^k - x^*, \lambda^k - \lambda^*)\big\|. \tag{5.152}$$

For every (x, λ), we have, by the mean value theorem,

$$\nabla L(x, \lambda) = B \begin{pmatrix} x - x^* \\ \lambda - \lambda^* \end{pmatrix},$$

where each row of B is the corresponding row of $\nabla^2 L$ evaluated at a point between (x, λ) and (x^*, λ^*). Since $\nabla^2 L(x^*, \lambda^*)$ is invertible, it follows that there is an $\epsilon > 0$ and scalars $\mu > 0$ and $M > 0$ such that for $\big\|(x - x^*, \lambda - \lambda^*)\big\| < \epsilon$, we have

$$\mu \big\|(x - x^*, \lambda - \lambda^*)\big\| \leq \big\|\nabla L(x, \lambda)\big\| \leq M \big\|(x - x^*, \lambda - \lambda^*)\big\|. \tag{5.153}$$

From Eqs. (5.152) and (5.153), it follows that for each $\bar{r} > 0$ there exists $\delta > 0$ such that, for $\big\|(x^k - x^*, \lambda^k - \lambda^*)\big\| < \delta$,

$$\big\|\nabla L(x^{k+1}, \lambda^{k+1})\big\| \leq (M\bar{r}/\mu) \big\|\nabla L(x^k, \lambda^k)\big\|,$$

or, equivalently,

$$P(x^{k+1}, \lambda^{k+1}) \leq (M^2 \bar{r}^2 / \mu^2) P(x^k, \lambda^k).$$

Given $r > 0$, we take $\bar{r} = (\mu/M)\sqrt{r}$ in the preceding relation, and Eq. (5.151) follows.

(c) We have shown that $x^{k+1} - x^k = -(H^k + cN^kN^{k\prime})^{-1}\nabla_x L_c\big(x^k, \lambda^{k+1}\big)$ [cf. Eq. (5.147)], which implies the conclusion. **Q.E.D.**

It is also possible to use the differentiable exact penalty functions of Section 5.3.3 as merit functions for Newton's method. The verification of this is somewhat tedious, so we refer to [Ber82a], p. 219, and [Ber82c] for an analysis. Moreover, it can be shown that differentiable exact penalty functions, while more complicated, have an interesting advantage over the nondifferentiable penalty function (5.150): they are not susceptible to the Maratos' effect discussed in Exercise 5.3.9 of Section 5.3, and they allow superlinear convergence of Newton-like methods without any complex modifications. For this analysis, we refer to the paper [Ber80b] (see also [Ber82a], pp. 271-279).

Variations of Newton's Method

There are a number of variations of Newton's method, which aim at some beneficial effect, and are obtained by introducing some extra terms in the left-hand side of the Newton system. These variations have the general form

$$x^{k+1} = x^k + \Delta x^k, \qquad \lambda^{k+1} = \lambda^k + \Delta\lambda^k,$$

where

$$\big(\nabla^2 L(x^k, \lambda^k) + V^k(x^k, \lambda^k)\big)\begin{pmatrix}\Delta x^k \\ \Delta\lambda^k\end{pmatrix} = -\nabla L(x^k, \lambda^k),$$

with the extra term $V^k(x^k, \lambda^k)$ being "small" enough relative to $\nabla^2 L(x^k, \lambda^k)$, so that the eigenvalues of the matrix

$$I - \big(\nabla^2 L(x^k, \lambda^k) + V^k(x^k, \lambda^k)\big)^{-1}\nabla^2 L(x^k, \lambda^k)$$

are within the unit circle and the convergence result of Prop. 5.4.1 applies. In the case where the extra term $V^k(x^k, \lambda^k)$ converges to zero, superlinear convergence is attained; otherwise, the rate of convergence is linear (cf. Prop. 5.4.1).

An interesting approximation of Newton's method is obtained by adding a term $-(1/c^k)\Delta\lambda^k$ in the left-hand side of the equation $N^{k\prime}\Delta x^k = -h(x^k)$, where c^k is a positive parameter, so that Δx^k and $\Delta\lambda^k$ are obtained by solving the system

$$H^k\Delta x^k + N^k\Delta\lambda^k = -\nabla_x L(x^k, \lambda^k), \tag{5.154}$$

$$N^{k\prime}\Delta x^k - (1/c^k)\Delta\lambda^k = -h(x^k). \tag{5.155}$$

As $c^k \to \infty$, the system asymptotically becomes identical to the one corresponding to Newton's method.

We can show that the system (5.154)-(5.155) has a unique solution if either $(H^k)^{-1}$ or $(H^k+c^kN^kN^{k'})^{-1}$ exists. Indeed when $(H^k)^{-1}$ exists, we can write explicitly the solution. By multiplying Eq. (5.154) by $N^{k'}(H^k)^{-1}$ and by using Eq. (5.155), we obtain

$$(1/c^k)\Delta\lambda^k - h(x^k) + N^{k'}(H^k)^{-1}N^k\Delta\lambda^k = -N^{k'}(H^k)^{-1}\nabla_x L(x^k,\lambda^k),$$

from which

$$\Delta\lambda^k = \big((1/c^k)I + N^{k'}(H^k)^{-1}N^k\big)^{-1}\big(h(x^k) - N^{k'}(H^k)^{-1}\nabla_x L(x^k,\lambda^k)\big)$$

and

$$\lambda^{k+1} = \lambda^k + \big((1/c^k)I + N^{k'}(H^k)^{-1}N^k\big)^{-1}\big(h(x^k) - N^{k'}(H^k)^{-1}\nabla_x L(x^k,\lambda^k)\big).$$

From Eq. (5.154), we then obtain

$$x^{k+1} = x^k - (H^k)^{-1}\nabla_x L(x^k,\lambda^{k+1}).$$

Also if $(H^k + c^kN^kN^{k'})^{-1}$ exists, by multiplying Eq. (5.155) with c^kN^k and adding the resulting equation to Eq. (5.154), we obtain

$$(H^k + c^kN^kN^{k'})\Delta x^k = -\nabla_x L(x^k,\lambda^k) - c^kN^kh(x^k),$$

and finally,

$$x^{k+1} = x^k - (H^k + c^kN^kN^{k'})^{-1}\nabla L_{c^k}(x^k,\lambda^k), \tag{5.156}$$

where L_{c^k} is the augmented Lagrangian function. Furthermore, from Eq. (5.155), we obtain

$$\lambda^{k+1} = \lambda^k + c^k\big(h(x^k) + N^{k'}(x^{k+1} - x^k)\big). \tag{5.157}$$

Note that the preceding development shows that N^k *need not have rank m in order for the system (5.154)-(5.155) to have a unique solution, while this is not true for the Newton iteration*. Thus by introducing the term $(1/c^k)\Delta\lambda^k$ in the second equation, we avoid potential difficulties due to linear dependence of the constraint gradients.

The preceding analysis suggests that if c^k is taken sufficiently large, then the approximate Newton iteration (5.156)-(5.157) should converge locally to a local minimum-Lagrange multiplier pair (x^*,λ^*) under the same conditions as the exact Newton iteration (cf. Prop. 5.4.3). Furthermore, the rate of convergence should be superlinear if $c^k\to\infty$. The proof of this, using Prop. 5.4.1, is straightforward, but is tedious and will not be given; see [Ber82a], pp. 240-243, where some variations of the method of Eqs. (5.156)-(5.157) are also discussed.

Another type of approximate Newton's method is obtained by introducing extra terms in the right-hand side (rather than the left-hand side) of the Newton system. For local convergence to (x^*,λ^*), it is essential that the extra terms tend to zero. The primal-dual methods for linear programming of Section 5.1.2 are of this type.

Connection with the First Order Method of Multipliers

From Eq. (5.156) it is seen that if $H^k + c^k N^k N^{k\prime}$ is positive definite, then $(x^{k+1} - x^k)$ is a descent direction for the augmented Lagrangian function $L_{c^k}(\cdot, \lambda^k)$. Furthermore, if the constraint functions h_i are linear, then Eq. (5.157) can be written as

$$\lambda^{k+1} = \lambda^k + c^k h(x^{k+1}), \tag{5.158}$$

while if in addition f is quadratic and $H^k + c^k N^k N^{k\prime}$ is positive definite, then from Eq. (5.156), x^{k+1} is the unique minimizing point of the augmented Lagrangian $L_{c^k}(\cdot, \lambda^k)$. Hence, it follows that *if the constraints are linear and the cost function is quadratic, then the iteration (5.156)-(5.157) is equivalent to the first order method of multipliers of Section 5.2.*

In the more general case where the constraints are nonlinear, it is natural to consider the iteration

$$x^{k+1} = x^k - \nabla^2_{xx} L_{c^k}(x^k, \lambda^k)^{-1} \nabla_x L_{c^k}(x^k, \lambda^k), \tag{5.159}$$

followed by the first order multiplier iteration

$$\lambda^{k+1} = \lambda^k + c^k h(x^{k+1}). \tag{5.160}$$

This is simply the first order multiplier iteration where the minimization of the augmented Lagrangian is replaced by a *single* pure Newton step, a method known as the *diagonalized method of multipliers*.

Note that for c^k large and x^k close to x^*, the Hessian $\nabla^2_{xx} L_{c^k}(x^k, \lambda^k)$ is nearly equal to $H^k + c^k N^k N^{k\prime}$, and $h(x^{k+1})$ is nearly equal to $h(x^k) + N^{k\prime}(x^{k+1} - x^k)$. Thus the diagonalized method of multipliers (5.159)-(5.160) can be viewed as an approximation to the variation of Newton's method (5.156)-(5.157) discussed earlier. This suggests that if c^k is taken larger than some threshold for all k, then the method should converge locally to a local minimum-Lagrange multiplier pair (x^*, λ^*) under the conditions of Prop. 5.4.3. This is indeed true, and the proof may be found in [Tap77] and [Ber82a], pp. 241-243, where it is also shown that the rate of convergence is superlinear if $c^k \to \infty$.

Extension to Inequality Constraints

Let us consider the inequality constrained problem

$$\begin{aligned} &\text{minimize} \quad f(x) \\ &\text{subject to} \quad g_j(x) \le 0, \qquad j = 1, \ldots, r, \end{aligned}$$

and focus on a local minimum x^* that is regular and together with a Lagrange multiplier μ^*, satisfies the second order sufficiency conditions of Prop. 4.3.2.

We can develop a Newton method for this problem, which is an extension of the quadratic programming implementation given earlier for equality constraints [cf. Eq. (5.148)]. This method is also similar to the constrained version of Newton's method for convex constraint sets given in Section 3.3 (in fact the two methods coincide when all the constraints are linear). In particular, given (x^k, μ^k), we obtain (x^{k+1}, μ^{k+1}) as an optimal solution-Lagrange multiplier pair of the quadratic program

$$\begin{aligned} &\text{minimize} \;\; \nabla f(x^k)'(x - x^k) + \tfrac{1}{2}(x - x^k)'\nabla^2_{xx}L(x^k, \mu^k)(x - x^k) \\ &\text{subject to} \;\; g_j(x^k) + \nabla g_j(x^k)'(x - x^k) \le 0, \qquad j = 1, \ldots, r. \end{aligned}$$

It is possible to show that there exists a neighborhood S of (x^*, μ^*) such that if (x^k, μ^k) is within S, then (x^{k+1}, μ^{k+1}) is uniquely defined as an optimal solution-Lagrange multiplier pair within S (an application of the implicit function theorem is needed to formalize this statement). Furthermore, (x^k, μ^k) converges to (x^*, μ^*) superlinearly. The details of this development are quite complex, and we refer to the book [Ber82a], Section 4.4.3, and the literature cited at the end of the chapter for further material.

5.4.3 Global Convergence

In order to enlarge the region of convergence of Lagrangian methods, it is necessary to combine them with some other method that has satisfactory global convergence properties. We refer to such a method as a *global method*. The main idea here is to construct a method that when sufficiently close to a local minimum switches automatically to a fast Lagrangian method, while when far away from such a point switches automatically to the global method, which is designed to make steady progress towards approaching the set of optimal solutions. The Lagrangian method can be any method that updates both vectors x and λ, based on the guidelines of this section. Prime candidates for use as global methods are multiplier methods and exact penalty methods.

There are many possibilities for combining global and Lagrangian methods, and the suitability of any one of these depends strongly on the problem at hand. For this reason, our main purpose in this section is not to develop and recommend specific algorithms, but rather to focus on the main guidelines for harmoniously interfacing global and Lagrangian methods while retaining the advantages of both. We note that combinations of global and Lagrangian methods, which involve augmented Lagrangian functions for convexification purposes, underlie several practical algorithms, including some popular nonlinear programming codes [CGT92], [MuS87].

Once a global and a Lagrangian method have been selected, the main issue to be settled is the choice of what we shall call the *switching rule* and the *acceptance rule*. The switching rule determines at each iteration, on the basis of certain tests, whether a switch should be made to the Lagrangian method. The tests depend on the information currently available, and their purpose is to decide whether an iteration of the Lagrangian method has a reasonable chance of success. As an example, such tests might include verification that ∇h has rank m and that $\nabla^2_{xx}L$ is positive definite on the subspace $\{y \mid \nabla h'y = 0\}$. We hasten to add here that these tests should not require excessive computational overhead. In some cases a switch might be made without any test at all, subject only to the condition that the Lagrangian iteration is well-defined.

The acceptance rule determines whether the results of the Lagrangian iteration will be accepted as they are, whether they will be modified, or whether they will be rejected completely and a switch will be made back to the global method. Typically, acceptance of the results of the Lagrangian iteration is based on reducing the value of a suitable merit function.

Combinations with Multiplier Methods

One possibility for enlarging the region of convergence of Lagrangian methods is to combine them with the first or second order methods of multipliers discussed in Section 5.2. The resulting methods tend to be very reliable since they inherit the robustness of the method of multipliers. At the same time they typically require fewer iterations to converge within the same accuracy than pure methods of multipliers.

To convey the general idea, let us discuss a few of the many possibilities. The simplest one is to switch to a Lagrangian method at the end of each (perhaps approximate) unconstrained minimization of a method of multipliers and continue using the Lagrangian method as long as the value of the exact penalty function $\|\nabla L\|^2$ is being decreased by a certain factor at each iteration. If satisfactory progress in decreasing $\|\nabla L\|^2$ is not observed, a switch back to the method of multipliers is made. Another possibility is to attempt a switch to a Lagrangian method at each iteration. As an example, consider a method for the equality constrained problem, which combines Newton's method for unconstrained minimization of the augmented Lagrangian together with the approximate Newton/Lagrangian iterations (5.156)-(5.160), which correspond to the method of multipliers.

At iteration k, we have x^k, λ^k, and a penalty parameter c^k. We also have a positive scalar w^k, which represents a target value of the exact penalty function $\|\nabla L\|^2$ that must be attained in order to accept the Lagrangian iteration, and a positive scalar ϵ^k that controls the accuracy of the unconstrained minimization of the method of multipliers. At the kth iteration, we determine x^{k+1}, λ^{k+1}, w^{k+1}, and ϵ^{k+1} as follows:

We form the Cholesky factorization $L^kL^{k\prime}$ of $\nabla^2_{xx}L_{c^k}(x^k,\lambda^k)$ as in Section 1.4. In the process, we modify $\nabla^2_{xx}L_{c^k}(x^k,\lambda^k)$ if it is not "sufficiently positive definite" (compare with Section 1.4). We then find the Newton direction

$$d^k = -(L^kL^{k\prime})^{-1}\nabla_x L_{c^k}(x^k,\lambda^k), \tag{5.161}$$

and if $\nabla^2_{xx} L_{c^k}(x^k, \lambda^k)$ was found "sufficiently positive definite" during the factorization process, we also carry out the Lagrangian iteration [compare with Eqs. (5.159) and (5.160)]:

$$\bar{x}^k = x^k + d^k, \tag{5.162}$$

$$\bar{\lambda}^k = \lambda^k + c^k h(\bar{x}^k). \tag{5.163}$$

[The analog of Eq. (5.157) could also be used in place of Eq. (5.163).]

If

$$\big\|\nabla L(\bar{x}^k, \bar{\lambda}^k)\big\|^2 \le w^k,$$

then we accept the Lagrangian iteration and we set

$$x^{k+1} = \bar{x}^k, \qquad \lambda^{k+1} = \bar{\lambda}^k, \qquad c^{k+1} = c^k, \qquad \epsilon^{k+1} = \epsilon^k,$$

$$w^{k+1} = \gamma \big\|\nabla L(\bar{x}^k, \bar{\lambda}^k)\big\|^2,$$

where γ is a fixed scalar with $0 < \gamma < 1$.

Otherwise, we do not accept the results of the Lagrangian iteration, that is we do not update λ^k. Instead we revert to minimization of the augmented Lagrangian $L_{c^k}(\cdot, \lambda^k)$ by performing an Armijo-type line search. In particular, we set

$$x^{k+1} = x^k + \alpha^k d^k,$$

where the stepsize is obtained as

$$\alpha^k = \beta^{m_k},$$

where m_k is the first nonnegative integer m such that

$$L_{c^k}(x^k, \lambda^k) - L_{c^k}(x^k + \beta^m d^k, \lambda^k) \ge -\sigma \beta^m d^{k\prime} \nabla_x L_{c^k}(x^k, \lambda^k),$$

and β and σ are fixed scalars with $\beta \in (0, 1)$ and $\sigma \in (0, \frac{1}{2})$. If

$$\big\|\nabla_x L_{c^k}(x^{k+1}, \lambda^k)\big\| \le \epsilon^k,$$

implying termination of the current unconstrained minimization, we do the ordinary first order multiplier iteration, setting

$$\lambda^{k+1} = \lambda^k + c^k h(x^k), \tag{5.164}$$

$$\epsilon^{k+1} = \gamma \epsilon^k, \qquad c^{k+1} = r c^k, \qquad w^{k+1} = \gamma \big\|\nabla L(x^{k+1}, \lambda^{k+1})\big\|^2,$$

where r is a fixed scalar with $r > 1$. If

$$\big\|\nabla_x L_{c^k}(x^{k+1}, \lambda^k)\big\| > \epsilon^k,$$

we set

$$\lambda^{k+1} = \lambda^k, \qquad \epsilon^{k+1} = \epsilon^k, \qquad c^{k+1} = c^k, \qquad w_{k=1} = w^k,$$

and proceed with the next iteration.

An alternative combined algorithm is obtained by using, in place of the first order iteration (5.163), the second order iteration

$$\bar{\lambda}^k = \left(N^{k\prime}(H^k + c^k N^k N^{k\prime})^{-1} N^k\right)^{-1} \left(h(\tilde{x}^k) - N^{k\prime}(H^k + c^k N^k N^{k\prime})^{-1} \nabla f(\tilde{x}^k)\right) - c^k h(\tilde{x}^k),$$

where $\tilde{x}^k$ is obtained by a pure Newton step

$$\tilde{x}^k = x^k + d^k = x^k - (L^k L^{k\prime})^{-1} \nabla_x L(x^k, \lambda^k),$$

[cf. Eq. (5.161)]. This corresponds to the second implementation of Newton's method of Eqs. (5.146)-(5.147)]. One could then obtain the vector $\bar{x}^k$ by a line search on the augmented Lagrangian $L_{c^k}(\cdot, \bar{\lambda}^k)$ along the direction d^k. The first order multiplier update (5.164) could also be replaced by a second order update. This combination of Newton's method and the second order multiplier method has outstanding rate of convergence properties, particularly if relatively good starting points are known. The combination given earlier based on the first order multiplier updates (5.163)-(5.164) is simpler, particularly if second derivatives are hard to compute and/or a quasi-Newton approximation is used in Eq. (5.161) in place of the inverse Hessian of the augmented Lagrangian $(L^k L^{k\prime})^{-1}$.

5.4.4 A Comparison of Various Methods

Quite a few barrier, penalty, and Lagrange multiplier methods were given in this chapter, so it is worth reflecting on their suitability for different types of problems. Even though it is hard to provide reliable guidelines, one may at least delineate the relative strengths and weaknesses of the various methods in specific practical contexts.

The barrier methods of Section 5.1, generally must solve a sequence of minimization problems that are increasingly ill-conditioned. This is a disadvantage relative to the multiplier methods of Section 5.2, whose sequence of minimization problems need not be ill-conditioned, and also relative to the exact penalty methods of Section 5.3, which require solution of only one minimization problem. However, for linear and for quadratic programs there is special structure that makes the logarithmic barrier and also the primal-dual interior point methods of Section 5.1.2 preferable to multiplier and exact penalty methods, from the theoretical and apparently the practical point of view. Whether, there are other important classes of problems for which this is also true, is an open question.

Multiplier methods are excellent general purpose constrained optimization methods. Their main advantages are simplicity and robustness. They rely on well-developed unconstrained optimization technology, and they require fewer assumptions for their validity relative to their competitors. In particular, they can deal with nonlinear equality constraints, and they do not require the existence of second derivatives and the regularity of the generated iterates (although they can be made more efficient when second derivatives can be used and when the iterates are regular). For these reasons some of the most popular software packages for solving nonlinear programming problems are based on multiplier methods.

The main disadvantage of multiplier methods relative to exact penalty methods is that they require a sequence of unconstrained minimizations as opposed to a single minimization. This disadvantage can be ameliorated by making the minimizations inexact or by combining the multiplier method with a Lagrangian method as described in Section 5.4.3. Still, practice has shown that minimization of an exact penalty function by a Newton-like method can require substantially fewer iterations relative to a multiplier method. Note, however, that each of these iterations may require a potentially costly subproblem (as in the linearization method) or may require complex calculations (as in differentiable exact penalty methods).

Generally, both multiplier methods and exact penalty methods require some trial and error to obtain appropriate values of the penalty parameter and also to ensure that there are no difficulties with ill-conditioning. However, multiplier methods typically are easier to "tune" than exact penalty methods, and deal more comfortably with the absence of a good starting point. Thus, if only a limited number of optimization runs are required in a given practical problem after development of the optimization code, one is typically better off with a method of multipliers than with an exact penalty method. If on the other hand, repetitive solution of the same problem with minor variations is envisioned, solution time is an issue, and the associated overhead per iteration is reasonable, one may prefer to use an exact penalty method.

EXERCISES

5.4.1

Consider the problem

$$\begin{aligned} &\text{minimize} \quad -x_1 x_2 \\ &\text{subject to} \quad x_1 + x_2 = 2. \end{aligned}$$

Write the implementation (5.141)-(5.142) of Newton's method, and show that it finds the optimal solution in a single iteration, regardless of the starting point. Write also the approximate implementation (5.156)-(5.157) for the starting point $x^0 = (0,0)$, $\lambda^0 = 0$, and for the two values $c = 10^{-2}$ and $c = 10^2$. How does the error after a single iteration depend on c?

5.4.2

Use Prop. 5.4.1 to derive a local convergence result for the approximate implementation (5.156)-(5.157) of Newton's method. Do the same for iteration (5.159)-(5.160).

5.5 NOTES AND SOURCES

Section 5.1: The logarithmic barrier method dates to the work of Frisch in the middle 50s [Fri56]. Other barrier methods have been proposed by Carroll [Car61]. An important early reference on penalty and barrier methods is the book by Fiacco and McCormick [FiM68]. Properties of the central path are investigated by McLinden [McL80], Sonnevend [Son86], Bayer and Lagarias [BaL89], and Guler [Gul94]. For surveys of interior point methods for linear programming, which give many additional references, see Gonzaga [Gon92], den Hertog [Her94], and Forsgren, Gill, and Wright [FGW02]. The line of analysis that we use is due to Tseng [Tse89], which also gives a computational complexity result along the lines of Exercise 5.1.5 (see also Gonzaga [Gon91]).

There has been a lot of effort to apply interior point methods to nonlinear problems, such as quadratic programming (Anstreicher, den Hertog, Roos, and Terlaky [AHR93], Wright [Wri96]), linear complementarity problems (Kojima, Meggido, Noma, and Yoshise [KMN91], Tseng [Tse92], Wright [Wri93c]), matrix inequalities (Alizadeh [Ali92], [Ali95], Nesterov and Nemirovskii [NeN94], Vandenberghe and Boyd [VaB95]), general convex programming problems (Wright [Wri92], Kortanek and Zhu [KoZ93], Anstreicher and Vial [AnV94], Jarre and Saunders [JaS95], Kortanek and Zhu [KoZ95]), and optimal control (Wright [Wri93b]). The research monographs by Nesterov and Nemirovskii [NeN94], Wright [Wri97a], Ye [Ye97], and Renegar [Ren01] are devoted to interior point methods for linear, quadratic, and convex programming.

Among interior point algorithms for linear programming, primal-dual methods are widely considered as generally the most effective; see e.g., McShane, Monma, and Shanno [MMS91]. They were introduced for linear programming through the study of the primal-dual central path by Megiddo [Meg88]. They were turned into path-following algorithms in the subsequent papers by Kojima, Mizuno, and Yoshise [KMY89], and Monteiro and Adler [MoA89a]. Their convergence, rate of convergence, and computational complexity are discussed in Zhang and Tapia [ZhT92], [ZTP93], [ZhT93], Potra [Pot94], and Tapia, Zhang, and Ye [TZY95]. The predictor-corrector variant was proposed by Mehrotra [Meh92]. There have been several extensions to broader classes of problems, including quadratic programming (Monteiro and Adler [MoA89b]), and linear complementarity (Wright [Wri94] and Tseng [Tse95b]). The research monograph by Wright [Wri97a] provides an extensive development of primal-dual interior point methods.

Section 5.2: The method of multipliers for equality constraints was independently proposed by Hestenes [Hes69], Powell [Pow69], and Haarhoff and Buys [HaB70]. These references contained very little analysis and empirical evidence, but much subsequent work established the convergence properties of the method and proposed various extensions. Surveys by Bertsekas [Ber76b] and Rockafellar [Roc76c] summarize the work up to 1976, and the author's research monograph [Ber82a] provides a detailed analysis and many references. The global convergence of the method is discussed by Poljak and Tretjakov [PoT73b], Bertsekas [Ber76a], Polak and Tits [PoT80a], Bertsekas [Ber82a], and Conn, Gould, and Toint [CGT91]. Global and superlinear convergence results for second order methods of multipliers, which are analogous to Prop. 5.2.3, are given in [Ber82a], Section 2.3.2.

The class of nonquadratic penalty methods for inequality constraints, given in Section 5.2.5, was introduced by Kort and Bertsekas [KoB72], with a special focus on the exponential method of multipliers. The convergence properties of the sequence $\{x^k\}$ generated by this method are quite intricate and are discussed by Bertsekas [Ber82a], Tseng and Bertsekas [TsB93], and Iusem [Ius99] for convex problems. For nonconvex problems under second order sufficiency conditions, the convergence analysis follows the pattern of the corresponding analysis for the quadratic method of multipliers (see Nguyen and Strodiot [NgS79]). The exponential method was applied to the solution of systems of nonlinear inequalities by Bertsekas ([Ber82a], Section 5.1.3), and Schnabel [Sch82]. The modified barrier method was proposed and developed by Polyak; see [Pol92] and the references given therein.

For subsequent research on the exponential penalty, the modified barrier, and other related methods that use nonquadratic penalty functions; see Freund [Fre91], Guler [Gul92], Teboulle [Teb92], Chen and Teboulle [ChT93], Tseng and Bertsekas [TsB93], Eckstein [Eck94a], Iusem, Svaiter, and Teboulle [IST94], Iusem and Teboulle [IuT95], Ben-Tal and Zibulevsky [BeZ97], Polyak and Teboulle [PoT97], Wei, Qi, and Birge [WQB98], and Iusem [Ius99].

Section 5.3: Nondifferentiable exact penalty methods were first proposed by Zangwill [Zan67b]; see also Han and Mangasarian [HaM79], who survey the subject and give many references. The linearization method is due to Pschenichny [Psc70] (see also Pschenichny and Danilin [PsD76]). It was independently derived later by Han [Han77]. Convergence rate issues and modifications to improve the convergence rate of sequential quadratic programming algorithms and to avoid the Maratos' effect (Exercise 5.3.9) are discussed in Boggs, Tolle, and Wang [BTW82], Coleman and Conn [CoC82a], [CoC82b], Gabay [Gab82], Mayne and Polak [MaP82], Panier and Tits [PaT91], Bonnans, Panier, Tits, and Zhou [BPT92], and Bonnans [Bon89b], [Bon94]. Combinations of the linearization method and the two-metric projection method of Section 3.4 have been proposed by Heinkenschloss [Hei96]. Note that since the linearization method can minimize the nondifferentiable exact penalty function $f(x)+cP(x)$, it can also be used to minimize

$$P(x) = \max\{g_0(x), g_1(x), \ldots, g_r(x)\},$$

which is a typical case of a minimax problem.

Exact differentiable penalty methods involving only x were introduced by Fletcher [Fle70b]. Exact differentiable penalty methods involving both x and λ were introduced by DiPillo and Grippo [DiG79]. The relation between these two types of methods, their utility for sequential quadratic programming, and a number of variations were derived by Bertsekas [Ber82c] (see also [Ber82a], Section 4.3, which contains a detailed convergence analysis). Extensions to inequality constraints are given by Glad and Polak [GlP79], and Bertsekas [Ber82a]; see also Mukai and Polak [MuP75], Boggs and Tolle [BoT80], and DiPillo and Grippo [DiG89]. Differentiable exact penalty functions are used by Nazareth [Naz96], and Nazareth and Qi [NaQ96] to extend the region of convergence of Newton-like methods for solving systems of nonlinear equations.

Section 5.4: First order Lagrangian methods were introduced by Arrow, Hurwicz, and Uzawa [AHU58]. They were also analyzed by Poljak [Pol70], and Psenichnyi and Danilin [PsD75], whom we follow in our presentation. Combinations of Lagrangian methods with the proximal algorithm were proposed in more general form for convex programming problems by Chen and Teboulle [ChT94], together with applications in decomposition of separable problems.

A Newton-like Lagrangian method for inequality constraints was proposed by Wilson [Wil63]. For this method, a superlinear convergence rate was established by Robinson [Rob74] under second order sufficiency conditions, including strict complementarity. Superlinear convergence results for a variant of the method were shown under weaker conditions by Wright [Wri98] and Hager [Hag99]. Combinations of Lagrangian methods with first order methods of multipliers were given by Glad [Gla79]. An extensive discussion of Newton-like Lagrangian methods is given in the author's research monograph [Ber82a].