D.G Luenberger,
Linear & Nonlinear
Programming,
2nd ed.

Chapter 8 CONJUGATE DIRECTION METHODS

Conjugate direction methods can be regarded as being somewhat intermediate between the method of steepest descent and Newton's method. They are motivated by the desire to accelerate the typically slow convergence associated with steepest descent while avoiding the information requirements associated with the evaluation, storage, and inversion of the Hessian (or at least solution of a corresponding system of equations) as required by Newton's method.

Conjugate direction methods invariably are invented and analyzed for the purely quadratic problem

$$\text{minimize} \quad \tfrac{1}{2}\mathbf{x}^T\mathbf{Q}\mathbf{x} - \mathbf{b}^T\mathbf{x},$$

where $\mathbf{Q}$ is an $n \times n$ symmetric positive definite matrix. The techniques once worked out for this problem are then extended, by approximation, to more general problems; it being argued that, since near the solution point every problem is approximately quadratic, convergence behavior is similar to that for the pure quadratic situation.

The area of conjugate direction algorithms has been one of great creativity in the nonlinear programming field, illustrating that detailed analysis of the pure quadratic problem can lead to significant practical advances. Indeed, conjugate direction methods, especially the method of conjugate gradients, have proved to be extremely effective in dealing with general objective functions and are considered among the best general purpose methods presently available.

8.1 CONJUGATE DIRECTIONS

Definition. Given a symmetric matrix $\mathbf{Q}$, two vectors $\mathbf{d}_1$ and $\mathbf{d}_2$ are said to be $\mathbf{Q}$-*orthogonal,* or *conjugate with respect to* $\mathbf{Q}$, if $\mathbf{d}_1^T\mathbf{Q}\mathbf{d}_2 = 0$.

In the applications that we consider, the matrix $\mathbf{Q}$ will be positive definite but this is not inherent in the basic definition. Thus if $\mathbf{Q} = \mathbf{0}$, any two vectors are conjugate, while if $\mathbf{Q} = \mathbf{I}$, conjugacy is equivalent to the usual notion of orthogonality. A finite set of vectors $\mathbf{d}_0, \mathbf{d}_1, \ldots, \mathbf{d}_k$ is said to be a **Q**-orthogonal set if $\mathbf{d}_i^T\mathbf{Q}\mathbf{d}_j = 0$ for all $i \neq j$.

Proposition. *If* **Q** *is positive definite and the set of nonzero vectors* $\mathbf{d}_0, \mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_k$ *are* **Q**-*orthogonal, then these vectors are linearly independent.*

Proof. Suppose there are constants α_i, $i = 0, 1, 2, \ldots, k$ such that

$$\alpha_0\mathbf{d}_0 + \cdots + \alpha_k\mathbf{d}_k = \mathbf{0}.$$

Multiplying by **Q** and taking the scalar product with $\mathbf{d}_i$ yields

$$\alpha_i\mathbf{d}_i^T\mathbf{Q}\mathbf{d}_i = 0.$$

Or, since $\mathbf{d}_i^T\mathbf{Q}\mathbf{d}_i > 0$ in view of the positive definiteness of **Q**, we have $\alpha_i = 0$. ∎

Before discussing the general conjugate direction algorithm, let us investigate just why the notion of **Q**-orthogonality is useful in the solution of the quadratic problem

$$\text{minimize} \quad \tfrac{1}{2}\mathbf{x}^T\mathbf{Q}\mathbf{x} - \mathbf{b}^T\mathbf{x}, \tag{1}$$

when **Q** is positive definite. Recall that the unique solution to this problem is also the unique solution to the linear equation

$$\mathbf{Q}\mathbf{x} = \mathbf{b}, \tag{2}$$

and hence that the quadratic minimization problem is equivalent to a linear equation problem.

Corresponding to the $n \times n$ positive definite matrix **Q** let $\mathbf{d}_0, \mathbf{d}_1, \ldots, \mathbf{d}_{n-1}$ be n nonzero **Q**-orthogonal vectors. By the above proposition they are linearly independent, which implies that the solution $\mathbf{x}^*$ of (1) or (2) can be expanded in terms of them as

$$\mathbf{x}^* = \alpha_0\mathbf{d}_0 + \cdots + \alpha_{n-1}\mathbf{d}_{n-1} \tag{3}$$

for some set of α_i's. In fact, multiplying by **Q** and then taking the scalar product with $\mathbf{d}_i$ yields directly

$$\alpha_i = \frac{\mathbf{d}_i^T\mathbf{Q}\mathbf{x}^*}{\mathbf{d}_i^T\mathbf{Q}\mathbf{d}_i} = \frac{\mathbf{d}_i^T\mathbf{b}}{\mathbf{d}_i^T\mathbf{Q}\mathbf{d}_i}. \tag{4}$$

This shows that the α_i's and consequently the solution $\mathbf{x}^*$ can be found by evaluation of simple scalar products. The end result is

$$\mathbf{x}^* = \sum_{i=0}^{n-1} \frac{\mathbf{d}_i^T\mathbf{b}}{\mathbf{d}_i^T\mathbf{Q}\mathbf{d}_i}\mathbf{d}_i. \tag{5}$$

There are two basic ideas imbedded in (5). The first is the idea of selecting an orthogonal set of $\mathbf{d}_i$'s so that by taking an appropriate scalar product, all terms on the right side of (3), except the ith, vanish. This could, of course, have been accomplished by making the $\mathbf{d}_i$'s orthogonal in the ordinary sense instead of making them $\mathbf{Q}$-orthogonal. The second basic observation, however, is that by using $\mathbf{Q}$-orthogonality the resulting equation for α_i can be expressed in terms of the known vector $\mathbf{b}$ rather than the unknown vector $\mathbf{x}^*$; hence the coefficients can be evaluated without knowing $\mathbf{x}^*$.

The expansion for $\mathbf{x}^*$ can be considered to be the result of an iterative process of n steps where at the ith step $\alpha_i\mathbf{d}_i$ is added. Viewing the procedure this way, and allowing for an arbitrary initial point for the iteration, the basic conjugate direction method is obtained.

Conjugate Direction Theorem. *Let $\{\mathbf{d}_i\}_{i=0}^{n-1}$ be a set of nonzero $\mathbf{Q}$-orthogonal vectors. For any $\mathbf{x}_0 \in E^n$ the sequence $\{\mathbf{x}_k\}$ generated according to*

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k\mathbf{d}_k, \qquad k \geqslant 0 \tag{6}$$

with

$$\alpha_k = -\frac{\mathbf{g}_k^T\mathbf{d}_k}{\mathbf{d}_k^T\mathbf{Q}\mathbf{d}_k} \tag{7}$$

and

$$\mathbf{g}_k = \mathbf{Q}\mathbf{x}_k - \mathbf{b},$$

converges to the unique solution, $\mathbf{x}^*$, of $\mathbf{Q}\mathbf{x} = \mathbf{b}$ after n steps, that is, $\mathbf{x}_n = \mathbf{x}^*$.

Proof. Since the $\mathbf{d}_k$'s are linearly independent, we can write

$$\mathbf{x}^* - \mathbf{x}_0 = \alpha_0\mathbf{d}_0 + \alpha_1\mathbf{d}_1 + \cdots + \alpha_{n-1}\mathbf{d}_{n-1}$$

for some set of α_k's. As we did to get (4), we multiply by $\mathbf{Q}$ and take the scalar product with $\mathbf{d}_k$ to find

$$\alpha_k = \frac{\mathbf{d}_k^T\mathbf{Q}(\mathbf{x}^* - \mathbf{x}_0)}{\mathbf{d}_k^T\mathbf{Q}\mathbf{d}_k}. \tag{8}$$

Now following the iterative process (6) from $\mathbf{x}_0$ up to $\mathbf{x}_k$ gives

$$\mathbf{x}_k - \mathbf{x}_0 = \alpha_0\mathbf{d}_0 + \alpha_1\mathbf{d}_1 + \cdots + \alpha_{k-1}\mathbf{d}_{k-1}, \tag{9}$$

and hence by the $\mathbf{Q}$-orthogonality of the $\mathbf{d}_k$'s it follows that

$$\mathbf{d}_k^T\mathbf{Q}(\mathbf{x}_k - \mathbf{x}_0) = 0. \tag{10}$$

Substituting (10) into (8) produces

$$\alpha_k = \frac{\mathbf{d}_k^T\mathbf{Q}(\mathbf{x}^* - \mathbf{x}_k)}{\mathbf{d}_k^T\mathbf{Q}\mathbf{d}_k} = -\frac{\mathbf{g}_k^T\mathbf{d}_k}{\mathbf{d}_k^T\mathbf{Q}\mathbf{d}_k},$$

which is identical with (7). ▮

To this point the conjugate direction method has been derived essentially through the observation that solving (1) is equivalent to solving (2). The conjugate direction method has been viewed simply as a somewhat special, but nevertheless straightforward, orthogonal expansion for the solution to (2). This viewpoint, although important because of its underlying simplicity, ignores some of the most important aspects of the algorithm; especially those aspects that are important when extending the method to nonquadratic problems. These additional properties are discussed in the next section.

Also, methods for selecting or generating sequences of conjugate directions have not yet been presented. Some methods for doing this are discussed in the exercises; while the most important method, that of conjugate gradients, is discussed in Section 8.3.

8.2 DESCENT PROPERTIES OF THE CONJUGATE DIRECTION METHOD

We define $\mathcal{B}_k$ as the subspace of E^n spanned by $\{\mathbf{d}_0, \mathbf{d}_1, \ldots, \mathbf{d}_{k-1}\}$. We shall show that as the method of conjugate directions progresses each $\mathbf{x}_k$ minimizes the objective over the k-dimensional linear variety $\mathbf{x}_0 + \mathcal{B}_k$.

Expanding Subspace Theorem. *Let $\{\mathbf{d}_i\}_{i=0}^{n-1}$ be a sequence of nonzero* **Q**-*orthogonal vectors in E^n. Then for any $\mathbf{x}_0 \in E^n$ the sequence $\{\mathbf{x}_k\}$ generated according to*

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k\mathbf{d}_k \tag{11}$$

$$\alpha_k = -\frac{\mathbf{g}_k^T\mathbf{d}_k}{\mathbf{d}_k^T\mathbf{Q}\mathbf{d}_k} \tag{12}$$

has the property that $\mathbf{x}_k$ minimizes $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{Q}\mathbf{x} - \mathbf{b}^T\mathbf{x}$ on the line $\mathbf{x} = \mathbf{x}_{k-1} + \alpha\mathbf{d}_{k-1}$, $-\infty < \alpha < \infty$, as well as on the linear variety $\mathbf{x}_0 + \mathcal{B}_k$.

Proof. It need only be shown that $\mathbf{x}_k$ minimizes f on the linear variety $\mathbf{x}_0 + \mathcal{B}_k$, since it contains the line $\mathbf{x} = \mathbf{x}_{k-1} + \alpha\mathbf{d}_{k-1}$. Since f is a strictly convex function, the conclusion will hold if it can be shown that $\mathbf{g}_k$ is orthogonal to $\mathcal{B}_k$ (that is, the gradient of f at $\mathbf{x}_k$ is orthogonal to the subspace $\mathcal{B}_k$). The situation is illustrated in Fig. 8.1. (Compare Theorem 2, p. 181.)

We prove $\mathbf{g}_k \perp \mathcal{B}_k$ by induction. Since $\mathcal{B}_0$ is empty that hypothesis is true for $k = 0$. Assuming that it is true for k, that is, assuming $\mathbf{g}_k \perp \mathcal{B}_k$, we

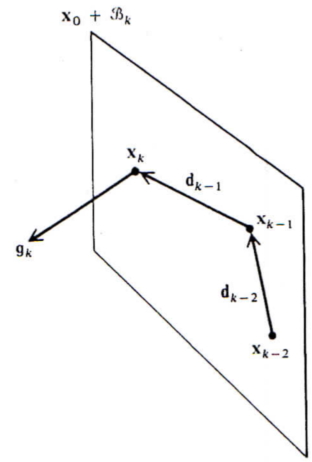

Fig. 8.1 Conjugate direction method

show that $\mathbf{g}_{k+1} \perp \mathcal{B}_{k+1}$. We have

$$\mathbf{g}_{k+1} = \mathbf{g}_k + \alpha_k \mathbf{Q}\mathbf{d}_k, \tag{13}$$

and hence

$$\mathbf{d}_k^T \mathbf{g}_{k+1} = \mathbf{d}_k^T \mathbf{g}_k + \alpha_k \mathbf{d}_k^T \mathbf{Q}\mathbf{d}_k = 0 \tag{14}$$

by definition of α_k. Also for $i < k$

$$\mathbf{d}_i^T \mathbf{g}_{k+1} = \mathbf{d}_i^T \mathbf{g}_k + \alpha_k \mathbf{d}_i^T \mathbf{Q}\mathbf{d}_k. \tag{15}$$

The first term on the right-hand side of (15) vanishes because of the induction hypothesis, while the second vanishes by the $\mathbf{Q}$-orthogonality of the $\mathbf{d}_i$'s. Thus $\mathbf{g}_{k+1} \perp \mathcal{B}_{k+1}$. ∎

Corollary. *In the method of conjugate directions the gradients* $\mathbf{g}_k$, $k = 0, 1, \ldots, n$ *satisfy*

$$\mathbf{g}_k^T \mathbf{d}_i = 0 \quad \text{for} \quad i < k.$$

The above theorem is referred to as the Expanding Subspace Theorem, since the $\mathcal{B}_k$'s form a sequence of subspaces with $\mathcal{B}_{k+1} \supset \mathcal{B}_k$. Since $\mathbf{x}_k$ minimizes f over $\mathbf{x}_0 + \mathcal{B}_k$, it is clear that $\mathbf{x}_n$ must be the overall minimum of f.

To obtain another interpretation of this result we again introduce the function

$$E(\mathbf{x}) = \tfrac{1}{2}(\mathbf{x} - \mathbf{x}^*)^T \mathbf{Q}(\mathbf{x} - \mathbf{x}^*) \tag{16}$$

as a measure of how close the vector $\mathbf{x}$ is to the solution $\mathbf{x}^*$. Since $E(\mathbf{x}) =$

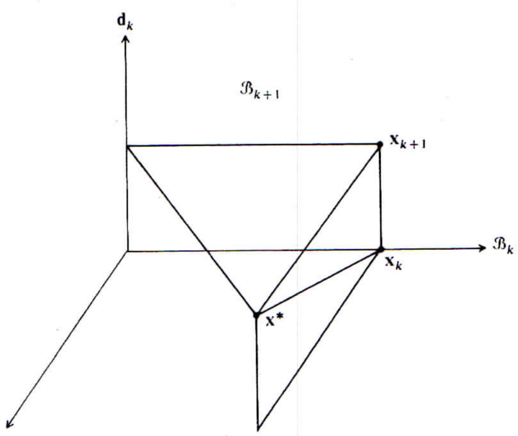

Fig. 8.2 Interpretation of expanding subspace theorem

$f(\mathbf{x}) + \frac{1}{2}\mathbf{x}^{*T}\mathbf{Q}\mathbf{x}^*$ the function E can be regarded as the objective that we seek to minimize.

By considering the minimization of E we can regard the original problem as one of minimizing a generalized distance from the point $\mathbf{x}^*$. Indeed, if we had $\mathbf{Q} = \mathbf{I}$, the generalized notion of distance would correspond (within a factor of two) to the usual Euclidean distance. For an arbitrary positive-definite $\mathbf{Q}$ we say E is a generalized Euclidean metric or distance function. Vectors $\mathbf{d}_i$, $i = 0, 1, \ldots, n-1$ that are $\mathbf{Q}$-orthogonal may be regarded as orthogonal in this generalized Euclidean space and this leads to the simple interpretation of the Expanding Subspace Theorem illustrated in Fig. 8.2. For simplicity we assume $\mathbf{x}_0 = \mathbf{0}$. In the figure $\mathbf{d}_k$ is shown as being orthogonal to $\mathcal{B}_k$ with respect to the generalized metric. The point $\mathbf{x}_k$ minimizes E over $\mathcal{B}_k$ while $\mathbf{x}_{k+1}$ minimizes E over $\mathcal{B}_{k+1}$. The basic property is that, since $\mathbf{d}_k$ is orthogonal to $\mathcal{B}_k$, the point $\mathbf{x}_{k+1}$ can be found by minimizing E along $\mathbf{d}_k$ and adding the result to $\mathbf{x}_k$.

8.3 THE CONJUGATE GRADIENT METHOD

The conjugate gradient method is the conjugate direction method that is obtained by selecting the successive direction vectors as a conjugate version of the successive gradients obtained as the method progresses. Thus, the directions are not specified beforehand, but rather are determined sequentially at each step of the iteration. At step k one evaluates the current negative gradient vector and adds to it a linear combination of the previous direction vectors to obtain a new conjugate direction vector along which to move.

There are three primary advantages to this method of direction selection. First, unless the solution is attained in less than n steps, the gradient is always nonzero and linearly independent of all previous direction vectors. Indeed, the gradient $\mathbf{g}_k$ is orthogonal to the subspace $\mathcal{B}_k$ generated by $\mathbf{d}_0, \mathbf{d}_1, \ldots, \mathbf{d}_{k-1}$. If the solution is reached before n steps are taken, the gradient vanishes and the process terminates—it being unnecessary, in this case, to find additional directions.

Second, a more important advantage of the conjugate gradient method is the especially simple formula that is used to determine the new direction vector. This simplicity makes the method only slightly more complicated than steepest descent.

Third, because the directions are based on the gradients, the process makes good uniform progress toward the solution at every step. This is in contrast to the situation for arbitrary sequences of conjugate directions in which progress may be slight until the final few steps. Although for the pure quadratic problem uniform progress is of no great importance, it is important for generalizations to nonquadratic problems.

Conjugate Gradient Algorithm

Starting at any $\mathbf{x}_0 \in E^n$ define $\mathbf{d}_0 = -\mathbf{g}_0 = \mathbf{b} - \mathbf{Q}\mathbf{x}_0$ and

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{d}_k \tag{17}$$

$$\alpha_k = -\frac{\mathbf{g}_k^T \mathbf{d}_k}{\mathbf{d}_k^T \mathbf{Q} \mathbf{d}_k} \tag{18}$$

(~ 3.27)

$$\mathbf{d}_{k+1} = -\mathbf{g}_{k+1} + \beta_k \mathbf{d}_k \tag{19}$$

$$\beta_k = \frac{\mathbf{g}_{k+1}^T \mathbf{Q} \mathbf{d}_k}{\mathbf{d}_k^T \mathbf{Q} \mathbf{d}_k}, \tag{20}$$

where $\mathbf{g}_k = \mathbf{Q}\mathbf{x}_k - \mathbf{b}$.

In the algorithm the first step is identical to a steepest descent step; each succeeding step moves in a direction that is a linear combination of the current gradient and the preceding direction vector. The attractive feature of the algorithm is the simple formulae, (19) and (20), for updating the direction vector. The method is only slightly more complicated to implement than the method of steepest descent but converges in a finite number of steps.

Verification of the Algorithm

To verify that the algorithm is a conjugate direction algorithm, it is necessary to verify that the vectors $\{\mathbf{d}_k\}$ are $\mathbf{Q}$-orthogonal. It is easiest to prove this by simultaneously proving a number of other properties of the algorithm. This is done in the theorem below where the notation $[\mathbf{d}_0, \mathbf{d}_1, \ldots, \mathbf{d}_k]$ is used to denote the subspace spanned by the vectors $\mathbf{d}_0, \mathbf{d}_1, \ldots, \mathbf{d}_k$.

Conjugate Gradient Theorem. *The conjugate gradient algorithm* (17)–(20) *is a conjugate direction method. If it does not terminate at* $\mathbf{x}_k$, *then*

a) $[\mathbf{g}_0, \mathbf{g}_1, \ldots, \mathbf{g}_k] = [\mathbf{g}_0, \mathbf{Q}\mathbf{g}_0, \ldots, \mathbf{Q}^k\mathbf{g}_0]$

b) $[\mathbf{d}_0, \mathbf{d}_1, \ldots, \mathbf{d}_k] = [\mathbf{g}_0, \mathbf{Q}\mathbf{g}_0, \ldots, \mathbf{Q}^k\mathbf{g}_0]$

c) $\mathbf{d}_k^T\mathbf{Q}\mathbf{d}_i = 0$ *for* $i \leqslant k - 1$

d) $\alpha_k = \mathbf{g}_k^T\mathbf{g}_k/\mathbf{d}_k^T\mathbf{Q}\mathbf{d}_k$

e) $\beta_k = \mathbf{g}_{k+1}^T\mathbf{g}_{k+1}/\mathbf{g}_k^T\mathbf{g}_k$. (~3.33)

Proof. We first prove (a), (b) and (c) simultaneously by induction. Clearly, they are true for $k = 0$. Now suppose they are true for k, we show that they are true for $k + 1$. We have

$$\mathbf{g}_{k+1} = \mathbf{g}_k + \alpha_k\mathbf{Q}\mathbf{d}_k.$$

By the induction hypothesis both $\mathbf{g}_k$ and $\mathbf{Q}\mathbf{d}_k$ belong to $[\mathbf{g}_0, \mathbf{Q}\mathbf{g}_0, \ldots, \mathbf{Q}^{k+1}\mathbf{g}_0]$, the first by (a) and the second by (b). Thus $\mathbf{g}_{k+1} \in [\mathbf{g}_0, \mathbf{Q}\mathbf{g}_0, \ldots, \mathbf{Q}^{k+1}\mathbf{g}_0]$. Furthermore $\mathbf{g}_{k+1} \notin [\mathbf{g}_0, \mathbf{Q}\mathbf{g}_0, \ldots, \mathbf{Q}^k\mathbf{g}_0] = [\mathbf{d}_0, \mathbf{d}_1, \ldots, \mathbf{d}_k]$ since otherwise $\mathbf{g}_{k+1} = \mathbf{0}$, because for any conjugate direction method $\mathbf{g}_{k+1}$ is orthogonal to $[\mathbf{d}_0, \mathbf{d}_1, \ldots, \mathbf{d}_k]$. (The induction hypothesis on (c) guarantees that the method is a conjugate direction method up to $\mathbf{x}_{k+1}$.) Thus, finally we conclude that (by expanding subspaces)

$$[\mathbf{g}_0, \mathbf{g}_1, \ldots, \mathbf{g}_{k+1}] = [\mathbf{g}_0, \mathbf{Q}\mathbf{g}_0, \ldots, \mathbf{Q}^{k+1}\mathbf{g}_0],$$

which proves (a).

To prove (b) we write

$$\mathbf{d}_{k+1} = -\mathbf{g}_{k+1} + \beta_k\mathbf{d}_k, \quad (19)$$

and (b) immediately follows from (a) and the induction hypothesis on (b).

Next, to prove (c) we have

$$\mathbf{d}_{k+1}^T\mathbf{Q}\mathbf{d}_i = -\mathbf{g}_{k+1}^T\mathbf{Q}\mathbf{d}_i + \beta_k\mathbf{d}_k^T\mathbf{Q}\mathbf{d}_i.$$

For $i = k$ the right side is zero by definition of β_k. For $i < k$ both terms vanish. The first term vanishes since $\mathbf{Q}\mathbf{d}_i \in [\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_{i+1}]$, the induction hypothesis which guarantees the method is a conjugate direction method up to $\mathbf{x}_{k+1}$, and by the Expanding Subspace Theorem that guarantees that $\mathbf{g}_{k+1}$ is orthogonal to $[\mathbf{d}_0, \mathbf{d}_1, \ldots, \mathbf{d}_{i+1}]$. The second term vanishes by the induction hypothesis on (c). This proves (c), which also proves that the method is a conjugate direction method.

To prove (d) we have

$$-\mathbf{g}_k^T\mathbf{d}_k = \mathbf{g}_k^T\mathbf{g}_k - \beta_{k-1}\mathbf{g}_k^T\mathbf{d}_{k-1},$$

and the second term is zero by the Expanding Subspace Theorem.

Finally, to prove (e) we note that $\mathbf{g}_{k+1}^T\mathbf{g}_k = 0$, because $\mathbf{g}_k \in [\mathbf{d}_0, \ldots, \mathbf{d}_k]$

and g_{k+1} is orthogonal to $[d_0, \ldots, d_k]$. Thus since

$$Qd_k = \frac{1}{\alpha_k}(g_{k+1} - g_k),$$

we have

$$g_{k+1}^T Q d_k = \frac{1}{\alpha_k} g_{k+1}^T g_{k+1}. \blacksquare$$

Parts (a) and (b) of this theorem are a formal statement of the interrelation between the direction vectors and the gradient vectors. Part (c) is the equation that verifies that the method is a conjugate direction method. Parts (d) and (e) are identities yielding alternative formulae for α_k and β_k that are often more convenient than the original ones.

8.4 THE C–G METHOD AS AN OPTIMAL PROCESS

We turn now to the description of a special viewpoint that leads quickly to some very profound convergence results for the method of conjugate gradients. The basis of the viewpoint is part (b) of the Conjugate Gradient Theorem. This result tells us the spaces $\mathcal{B}_k$ over which we successively minimize are determined by the original gradient g_0 and multiplications of it by Q. Each step of the method brings into consideration an additional power of Q times g_0. It is this observation we exploit.

Let us consider a new general approach for solving the quadratic minimization problem. Given an arbitrary starting point x_0, let

$$x_{k+1} = x_0 + P_k(Q)g_0, \tag{21}$$

where P_k is a polynomial of degree k. Selection of a set of coefficients for each of the polynomials P_k determines a sequence of x_k's. We have

$$\begin{aligned} x_{k+1} - x^* &= x_0 - x^* + P_k(Q)Q(x_0 - x^*) \\ &= [I + QP_k(Q)](x_0 - x^*), \end{aligned} \tag{22}$$

and hence

$$\begin{aligned} E(x_{k+1}) &= \tfrac{1}{2}(x_{k+1} - x^*)^T Q(x_{k+1} - x^*) \\ &= \tfrac{1}{2}(x_0 - x^*)^T Q[I + QP_k(Q)]^2(x_0 - x^*). \end{aligned} \tag{23}$$

We may now pose the problem of selecting the polynomial P_k in such a way as to minimize $E(x_{k+1})$ with respect to all possible polynomials of degree k. Expanding (21), however, we obtain

$$x_{k+1} = x_0 + \gamma_0 g_0 + \gamma_1 Q g_0 + \cdots + \gamma_k Q^k g_0, \tag{24}$$

where the γ_i's are the coefficients of P_k. In view of

$$\mathcal{B}_{k+1} = [d_0, d_1, \ldots, d_k] = [g_0, Qg_0, \ldots, Q^k g_0],$$