IMA Journal of Numerical Analysis (1990) **10**, 155–170

Conjugate Gradient Type Methods for Indefinite, Asymmetric, and Complex Systems

G. MARKHAM

National Power-Technology and Environmental Centre, Kelvin Avenue, Leatherhead, Surrey KT22 7SE

[Received 25 February 1986 and in revised form 25 May 1989]

Methods based upon preconditioned formulations of the conjugate gradient method, and its extensions to asymmetric and complex systems, are currently proving highly successful in solving linear algebraic systems of equations across a wide range of applications. As the range of applications increases it is timely to review the derivation and properties of conjugate gradient type methods.

The present discussion pursues two particular but interrelated themes, the extension of the CG method for real symmetric systems to methods of greater generality, and the so-called 'minimization' properties by which the methods have been characterized. The opportunity is taken to illuminate the behaviour of CG type methods and their performance under various circumstances, and in particular to amplify the role of the quadratic functional in the complex methods.

1. Introduction

THE conjugate gradient method for the solution of linear algebraic systems was derived by Hestenes & Stiefel (1952). It involves the generation of a sequence of vectors $\boldsymbol{r}_j$ which span Krylov subspaces; that is to say, given a starting vector $\boldsymbol{r}_0$, then after i steps a sequence of vectors $\boldsymbol{r}_0, \boldsymbol{r}_1, \ldots, \boldsymbol{r}_i$ will have been generated so as to span the subspace

$$S_{i+1} = \{\boldsymbol{r}_0, \boldsymbol{A}\boldsymbol{r}_0, \ldots, \boldsymbol{A}^i\boldsymbol{r}_0\},$$

where $\boldsymbol{A}$ is the matrix of coefficients of the system.

In this respect, the conjugate gradient method has a close affinity to the Lanczos (1950) method for reducing a matrix to tridiagonal form with a view to finding the eigenvalues. Furthermore, both methods have the property of there being three-term recurrence relations between the vectors $\boldsymbol{r}_j$, i.e. $\boldsymbol{r}_i$ can be expressed in terms of $\boldsymbol{r}_{i-1}$ and $\boldsymbol{r}_{i-2}$. In the Lanczos method, the vectors $\boldsymbol{r}_j$ computed after $m-1$ steps (i.e. $j = 0, 1, \ldots, m-1$) form the columns of the matrix $\boldsymbol{Q}_m$ by which a tridiagonal matrix $\boldsymbol{T}_m$ can be computed, related to the original matrix via

$$\boldsymbol{A}\boldsymbol{Q}_m = \boldsymbol{Q}_m\boldsymbol{T}_m + \boldsymbol{F}_m;$$

in exact arithmetic, the first $m-1$ columns of $\boldsymbol{F}_m$ will contain only zero elements. In the CG method, the vectors $\boldsymbol{r}_j$ are residuals of the system of linear equations $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$, corresponding to approximate solutions $\boldsymbol{x}_j$ via

$$\boldsymbol{r}_j = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_j.$$

The conjugate gradient method has the property that if A is symmetric, then the vectors r_i, if computed using exact arithmetic, are linearly independent of their predecessors $r_{i-1},\ldots, r_0$. The process will continue until, for some $m \leqslant n$ (where n is the order of the matrix A), the vector $A^m r_0$ is contained in S_m. The inequality will apply if A has any multiple eigenvalues or if r_0 is deficient in one or more eigenvector directions. But, since r_m must by the conjugacy relations be orthogonal to the subspace S_m, it must therefore be identically zero.

This finite termination property leads to the view of conjugate gradients as being a direct method for the solution of systems of linear equations. However, finite termination is observed not to occur in the course of practical numerical computation; moreover there are direct methods which achieve solutions in fewer operations than the number predicted for the CG method. For these reasons the conjugate gradient method received little practical application as a direct method.

However, interest was revived by Reid's (1971) paper, which showed that the sequence of vectors $r_0,\ldots, r_i$ displayed useful and, in the case of A being symmetric, monotonic convergence towards zero, with a corresponding convergence of x_j towards the solution x. The rate of convergence can be related to the eigenvalue structure of the matrix A, and offers the prospect, in some circumstances, of being able to provide a sufficiently accurate solution in a number of steps which is much less than n, thereby making the CG method competitive with both direct methods and established iterative methods.

It is in this guise, as an iterative solution method, that the conjugate gradient method has found practical application. The particular requirements on the eigenvalue structure for rapid convergence to be achieved are frequently not met in practical problems, but procedures can be adopted in order to render such systems more amenable to solution by CG methods. The nature of these requirements and the 'pre-conditioning' procedures designed to meet them have been discussed extensively elsewhere (e.g. Meijerink & van der Vorst, 1977; also Markham, 1987).

The present paper reviews the basic iterative methods of the conjugate gradient type. Two particular themes are pursued in the discussion, the so-called 'minimization' properties of the methods and the extensions of the basic CG method, for real symmetric systems, to methods for real asymmetric and general complex systems.

2. The Conjugate Gradient Method

Consider a system of n linear equations in n unknowns, written in the form

$$Ax = b,$$

where b is the known vector of right-hand-side terms and x is the vector of unknowns. At this stage no assumptions will be made about the matrix A concerning its symmetry or eigenvalue properties, other than that it is of full rank. Given an approximate solution x_i, the corresponding residual vector can be defined by

$$r_i = b - Ax_i.$$

Suppose that an iterative method is defined as follows: given a starting vector x_0, a sequence of approximate solutions is calculated via the steps

$$x_{i+1} = x_i + \alpha_i p_i, \qquad r_{i+1} = r_i - \alpha_i A p_i, \qquad p_{i+1} = r_{i+1} + \beta_i p_i, \quad (2.1a,b,c)$$

for $i = 0, 1, \ldots,$ until some termination criterion is achieved. The initial 'search vector' p_0 is set equal to r_0. Equations (2.1) define a whole family of algorithms based on three-term recurrence; members of the family are distinguished by particular prescriptions for the scalars α_i and β_i $(i = 0, 1, \ldots)$.

The conjugate gradient method represents a particular choice of prescription for the two scalars. The conventional presentation of conjugate gradient type methods proceeds by stating the particular algorithm of interest, making certain assumptions (e.g. that A is symmetric) and then deriving properties such as conjugacy of search directions and monotonic convergence. The approach that will be adopted here is in some sense the reverse: starting from properties which might be regarded as characteristic of conjugate gradient type methods, the intention has been to derive an algorithm with the least restrictions necessary to ensure that those properties hold. In the event it will prove expedient to adopt some sufficient rather than strictly necessary restrictions, but this expositional route still has the merit of showing precisely how these restrictions are induced by requirements. Of course, it will be recognized that starting this procedure from a different set of 'required properties' would change the derivation of the algorithm; in particular a stronger set of required properties might lead to greater restrictions, or indeed to the conclusion that no such algorithm could be derived at all.

2.1 *Derivation of methods from the required conjugacy properties*

Suppose that a method is sought with particular conjugacy properties expressible in the form

$$\langle r_i, p_j \rangle = 0 \quad \text{for } i > j, \tag{2.2a}$$

$$\langle Ap_i, p_j \rangle = 0 \quad \text{for } i > j, \tag{2.2b}$$

where the product $\langle x, y \rangle$ will for the moment be assumed merely to be bilinear and A is for the moment arbitrary (with respect to symmetry and eigenvalue properties). To determine the conditions under which such a method exists, it will be supposed that they are true for $i \leqslant k$. (This inductive proof follows that of Reid (1971).) The hypotheses will be denoted

$$\text{H1}(k): \quad \langle r_i, p_j \rangle = 0 \quad \text{for } j < i \leqslant k \tag{2.3a}$$

$$\text{H2}(k): \quad \langle Ap_i, p_j \rangle = 0 \quad \text{for } j < i \leqslant k \tag{2.3b}$$

Consider

$$\langle r_{k+1}, p_j \rangle = \langle r_k, p_j \rangle - \alpha_k \langle Ap_k, p_j \rangle$$

(where it will be assumed that $j \leqslant k$). If $k \neq j$ then H1(k) and H2(k) imply that the expression is equal to zero. If $k = j$ then the products will not necessarily (and in fact will probably not) be zero; thus $\langle r_{k+1}, p_k \rangle$ can only be zero (and

H1($k+1$) thereby proved) if α_k is defined by

$$\alpha_k = \langle \boldsymbol{r}_k, \boldsymbol{p}_k \rangle / \langle \boldsymbol{A}\boldsymbol{p}_k, \boldsymbol{p}_k \rangle. \tag{2.4}$$

If the denominator vanishes, then the algorithm breaks down; the extent to which this is a problem is dependent on the way in which the product $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$ is defined and on the properties of the matrix $\boldsymbol{A}$, as discussed briefly in Section 2.2.

To reiterate, with α_k defined by (2.4), H1(k) and H2(k) imply H1($k+1$).

Now consider (still with the assumption that $j \leq k$)

$$\langle \boldsymbol{A}\boldsymbol{p}_{k+1}, \boldsymbol{p}_j \rangle = \frac{1}{\alpha_{k+1}} [\langle \boldsymbol{r}_{k+1}, \boldsymbol{p}_j \rangle - \langle \boldsymbol{r}_{k+2}, \boldsymbol{p}_j \rangle].$$

The first term inside the brackets is zero by H1($k+1$), but the second would require H1($k+2$), which in turn requires the truth of the hypothesis H2($k+1$) currently under examination. Some further restriction is therefore required, and the form of this can be seen by examining

$$\begin{aligned} \langle \boldsymbol{p}_{k+1}, \boldsymbol{A}\boldsymbol{p}_j \rangle &= \langle \boldsymbol{r}_{k+1}, \boldsymbol{A}\boldsymbol{p}_j \rangle + \beta_k \langle \boldsymbol{p}_k, \boldsymbol{A}\boldsymbol{p}_j \rangle \\ &= \frac{1}{\alpha_j} [\langle \boldsymbol{r}_{k+1}, \boldsymbol{p}_j \rangle (1 + \beta_j) - \langle \boldsymbol{r}_{k+1}, \boldsymbol{p}_{j+1} \rangle \\ &\quad - \langle \boldsymbol{r}_{k+1}, \boldsymbol{p}_{j-1} \rangle \beta_{j-1}] + \beta_k \langle \boldsymbol{p}_k, \boldsymbol{A}\boldsymbol{p}_j \rangle. \end{aligned}$$

If $j < k$, all of the products of the $\langle \boldsymbol{r}, \boldsymbol{p} \rangle$ type are zero by virtue of H1($k+1$), leaving

$$\langle \boldsymbol{p}_{k+1}, \boldsymbol{A}\boldsymbol{p}_j \rangle = \beta_k \langle \boldsymbol{p}_k, \boldsymbol{A}\boldsymbol{p}_j \rangle.$$

If $j = k$, the right-hand side takes the form

$$\frac{1}{\alpha_k} [- \langle \boldsymbol{r}_{k+1}, \boldsymbol{p}_{k+1} \rangle] + \beta_k \langle \boldsymbol{p}_k, \boldsymbol{A}\boldsymbol{p}_k \rangle,$$

and so $\langle \boldsymbol{p}_{k+1}, \boldsymbol{A}\boldsymbol{p}_k \rangle = 0$ if β_k is defined by

$$\beta_k = \frac{1}{\alpha_k} \frac{\langle \boldsymbol{r}_{k+1}, \boldsymbol{p}_{k+1} \rangle}{\langle \boldsymbol{p}_k, \boldsymbol{A}\boldsymbol{p}_k \rangle}. \tag{2.5}$$

The proof of hypothesis H2($k+1$) comes from adopting the additional constraints that

$$\langle \boldsymbol{A}\boldsymbol{p}_{k+1}, \boldsymbol{p}_j \rangle = \langle \boldsymbol{p}_{k+1}, \boldsymbol{A}\boldsymbol{p}_j \rangle \quad \text{and} \quad \langle \boldsymbol{p}_k, \boldsymbol{A}\boldsymbol{p}_j \rangle = 0, \quad \text{for all } j < k. \tag{2.6a,b}$$

If the product takes a form such that

$$\langle \boldsymbol{A}\boldsymbol{x}, \boldsymbol{y} \rangle = \langle \boldsymbol{x}, \boldsymbol{A}\boldsymbol{y} \rangle, \tag{2.7}$$

then (2.6a) is obviously true, and (2.6b) is true via H2(k); this appears to be the weakest possible condition on the bilinear form.

Furthermore, the requirement of bilinearity, as exploited throughout the derivation, means that the product can be written, without loss of generality, as

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{x}^{\mathrm{T}} \boldsymbol{L} \boldsymbol{y}. \tag{2.8}$$

Given (2.8), the condition (2.7) can now be expressed as

$$\boldsymbol{A}^{\mathrm{T}}\boldsymbol{L} = \boldsymbol{L}\boldsymbol{A}. \tag{2.9}$$

Despite the earlier stated intention to discover necessary rather than sufficient conditions for the conjugate gradient method, it will be convenient at this stage to pursue the most obvious way of ensuring that (2.9) is maintained, namely by requiring that $\boldsymbol{A}$ is symmetric and that $\boldsymbol{A}$ commutes with $\boldsymbol{L}$. The most general way of satisfying (2.9) is discussed briefly in the Appendix. If special cases such as $\boldsymbol{A}$ being diagonal are excluded, the prescription adopted above amounts to restricting attention to the cases $\boldsymbol{L} = \boldsymbol{A}^{\mu}$; the index μ could in principle be fractional, but interest will usually lie with μ integral.

With these restrictions, the conjugacy relations and definitions of α_i and β_i now assume their familiar form (see Reid, 1971; Jacobs, 1980). The full inductive proof includes a demonstration that the conjugacy relations hold for $k = 1$, a stage omitted here for brevity. The 'cumulative conjugacy' relations are usually written in the form

$$\langle \boldsymbol{r}_i, \boldsymbol{p}_j \rangle = 0 \quad \text{for } i > j, \tag{2.10a}$$

$$\langle \boldsymbol{p}, \boldsymbol{A}\boldsymbol{p}_j \rangle = 0 \quad \text{for } i \neq j, \tag{2.10b}$$

$$\langle \boldsymbol{r}_i, \boldsymbol{r}_j \rangle = 0 \quad \text{for } i \neq j, \tag{2.10c}$$

with the scalar definitions

$$\alpha_i = \frac{\langle \boldsymbol{r}_i, \boldsymbol{p}_i \rangle}{\langle \boldsymbol{p}_i, \boldsymbol{A}\boldsymbol{p}_i \rangle} = \frac{\langle \boldsymbol{r}_i, \boldsymbol{r}_i \rangle}{\langle \boldsymbol{p}_i, \boldsymbol{A}\boldsymbol{p}_i \rangle}, \tag{2.11a}$$

$$\beta_i = -\frac{\langle \boldsymbol{r}_{i+1}, \boldsymbol{A}\boldsymbol{p}_i \rangle}{\langle \boldsymbol{p}_i, \boldsymbol{A}\boldsymbol{p}_i \rangle} = \frac{\langle \boldsymbol{r}_{i+1}, \boldsymbol{r}_{i+1} \rangle}{\langle \boldsymbol{r}_i, \boldsymbol{r}_i \rangle}. \tag{2.11b}$$

(The additional relation $\langle \boldsymbol{p}_k, \boldsymbol{r}_k \rangle = \langle \boldsymbol{r}_k, \boldsymbol{r}_k \rangle$ can be demonstrated either by a simple inductive proof or by using the symmetry of $\boldsymbol{L}$ (from the Appendix).)

The scalars α_i and β_i have now been defined up to the choice of μ in the definition of the product

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}^{\mu}\boldsymbol{y}. \tag{2.12}$$

The next subsection considers the impact of various choices of μ.

2.2 *Functional forms for the residual measure*

Central to the view of the conjugate gradient method as a viable iterative scheme (Reid, 1971) is the concept of some measure of the residual vector $\boldsymbol{r}$ being reduced at each step. This iterative progression may allow the method to achieve a solution to within a satisfactory accuracy in considerably fewer steps than required to span the Krylov space.

Consider the functional $f = \boldsymbol{r}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{r}$, where $\boldsymbol{K}$ is a (fixed) matrix; without loss of generality, $\boldsymbol{K}$ can be assumed to be symmetric, since any antisymmetric part of $\boldsymbol{K}$ makes no contribution to f. Other authors have employed the notation h^2 for the functional, but since, as will be shown, the functional value may become negative

(and even complex in the case of the complex CG extensions, see Section 4) the symbol 'f' seems less confusing. At the $(k+1)$th step, f will take the form

$$f_{k+1} = (\boldsymbol{r}_k - \alpha_k \boldsymbol{A}\boldsymbol{p}_k)^{\mathrm{T}} \boldsymbol{K} (\boldsymbol{r}_k - \alpha_k \boldsymbol{A}\boldsymbol{p}_k). \tag{2.13}$$

The functional will have a turning point with respect to α_k at the point where $\partial f_{k+1}/\partial \alpha_k = 0$; thus the scalar α_k defined by the conjugate gradient method produces a turning point in f_{k+1} if

$$\alpha_k = \frac{\boldsymbol{r}_k^{\mathrm{T}} \boldsymbol{L}\boldsymbol{p}_k}{\boldsymbol{p}_k^{\mathrm{T}} \boldsymbol{L}\boldsymbol{A}\boldsymbol{p}_k} = \frac{(\boldsymbol{A}\boldsymbol{p}_k)^{\mathrm{T}} \boldsymbol{K}\boldsymbol{r}_k + \boldsymbol{r}_k^{\mathrm{T}} \boldsymbol{K}\boldsymbol{A}\boldsymbol{p}_k}{2(\boldsymbol{A}\boldsymbol{p}_k)^{\mathrm{T}} \boldsymbol{K}\boldsymbol{A}\boldsymbol{p}_k}. \tag{2.14}$$

A particular functional with the desired properties can then be generated by equating terms separately in the numerators and denominators of the two expressions for α_k, giving

$$\boldsymbol{K}\boldsymbol{A} = \boldsymbol{L}, \qquad \boldsymbol{A}^{\mathrm{T}} \boldsymbol{K} = \boldsymbol{L}. \tag{2.15a,b}$$

Since the symmetry of $\boldsymbol{L}$ has already been demonstrated (in the Appendix), the functional with

$$\boldsymbol{K} = \boldsymbol{L}\boldsymbol{A}^{-1} \tag{2.15c}$$

has the required properties. The observation that $\boldsymbol{L}\boldsymbol{A}^{-1}$ is symmetric does not represent a newly imposed restriction, as

$$\boldsymbol{L}\boldsymbol{A}^{-1} = \boldsymbol{L}(\boldsymbol{L}\boldsymbol{A})^{-1}\boldsymbol{L}$$

and (as shown in the Appendix) both $\boldsymbol{L}$ and $\boldsymbol{L}\boldsymbol{A}$ must be symmetric for the conjugacy relations to hold. There is no point in looking for other functionals $f = \boldsymbol{r}^{\mathrm{T}} \boldsymbol{K}' \boldsymbol{r} + c$, $\boldsymbol{K}' \neq \boldsymbol{K}$; there are of course other 'functionals' with turning points at α_k, but the requirements of

(i) locally quadratic behaviour in α_k,
(ii) $f = 0$ when $\boldsymbol{r} = \boldsymbol{0}$, and
(iii) $\boldsymbol{K}'$ is a constant matrix (not varying with k),

constrain the functional to within a multiplicative scalar constant of $f = \boldsymbol{r}^{\mathrm{T}} \boldsymbol{L}\boldsymbol{A}^{-1}\boldsymbol{r}$. A similar argument will be used in identifying the functional for the biconjugate gradient method (Section 3.3).

If, as pursued in the main text, $\boldsymbol{L} = \boldsymbol{A}^{\mu}$, then the functional is of the form $\boldsymbol{r}^{\mathrm{T}} \boldsymbol{A}^{\nu} \boldsymbol{r}$ (as considered by Reid, 1971), corresponding to product definitions

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{x}^{\mathrm{T}} \boldsymbol{A}^{\mu} \boldsymbol{y}, \qquad \nu = \mu - 1. \tag{2.16}$$

As before, interest will lie principally with methods with μ and ν integral, although they could in principle be fractional. The choice $\nu = 0$ minimizes $\boldsymbol{r}^{\mathrm{T}} \boldsymbol{r}$, with the attraction that the functional only vanishes at $\boldsymbol{r} = \boldsymbol{0}$, i.e. at the solution to the system $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$. However, this so-called 'minimum residual method' (henceforth denoted the MR method in this paper) appears to involve more work at each iteration, in that product calculations appear to require a matrix–vector product to be performed. (However, it is possible to circumvent this additional computational burden by a rearrangement of the algorithm; see Markham

(1987).) Nevertheless, on a variety of grounds Reid advocates the use of $\nu = -1$, i.e.

$$f = \boldsymbol{r}^{\mathrm{T}}\boldsymbol{A}^{-1}\boldsymbol{r} \quad \text{and} \quad \langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{y}, \tag{2.17}$$

and many authors (e.g. Jackson & Robinson, 1982) refer to this choice as the 'conjugate gradient method' proper. Since this choice reduces the product definition (2.16) to the inner product of the two vectors, the method will henceforth be denoted the IP method in this paper.

Note that

$$f_k - f_{k+1} = \alpha_k \langle \boldsymbol{r}_k, \boldsymbol{p}_k \rangle = \alpha_k^2 \langle \boldsymbol{p}_k, \boldsymbol{A}\boldsymbol{p}_k \rangle; \tag{2.18}$$

if $\boldsymbol{A}$ is real and symmetric, the search vector $\boldsymbol{p}_k$ can be decomposed into a linear sum of linearly independent eigenvectors, $\boldsymbol{p}_k = \sum a_k^i \boldsymbol{u}^i$, with associated eigenvalues λ_i, and (2.18) can be rewritten as

$$f_k - f_{k+1} = \alpha_k^2 \sum_i (a_k^i)^2 (\lambda^i)^{\mu+1}. \tag{2.19}$$

The scalars α_k and a_k^i are always real. With the MR method, it is possible thus to conclude that the functional f must be reduced at each step. Since $f \geqslant 0$ with $\nu = \mu - 1 = 0$ and given, in addition, the finite termination property ($\boldsymbol{r} = 0$ after at most n steps), it is possible to speak of 'monotonic convergence towards the solution'.

In the IP case (favoured by Reid (1971)), $\boldsymbol{A}$ must be positive-definite in order both to show that $f \geqslant 0$ and that f reduces at each step. This observation, together with the fact that the denominator $\langle \boldsymbol{p}_k, \boldsymbol{A}\boldsymbol{p}_k \rangle$ in the definition of α_k might become zero in the IP algorithm if $\boldsymbol{A}$ were not positive-definite, has apparently led to the belief that the conjugate gradient method is strictly applicable only to positive-definite matrices (e.g. Jacobs, 1980). This is indeed correct if attention is focused on the IP algorithm and the $f > 0$ and 'monotonic convergence' properties are required. However, if the requirements on the functional behaviour are relaxed, then the only difficulty in applying the IP method to non-positive-definite systems is the possibility of the α_k denominator vanishing. The extent to which this is a serious difficulty in the practical context of finite-length arithmetic remains to be determined.

In the MR case, no reference has yet been made to the positive-definiteness or otherwise of $\boldsymbol{A}$. Provided that $\boldsymbol{p}_k \neq \boldsymbol{0}$ and $\boldsymbol{A}$ is of full rank, the α_k denominator can never be zero; but under the same conditions the numerator $\langle \boldsymbol{r}_k, \boldsymbol{r}_k \rangle$ *could* vanish, and no further iterative progress would be made. Similar comments to those above on finite-length arithmetic apply here.

In the positive-definite case, the term $\langle \boldsymbol{p}_k, \boldsymbol{A}\boldsymbol{p}_k \rangle$ is always positive for either of the integral-μ (MR or IP) choices at (2.16), hence the usual statement that the 'conjugate gradient method minimizes the functional at each step' (the second derivative of f with respect to α being positive). It will be noted that in the indefinite case, and for the IP method, the term $\langle \boldsymbol{p}_k, \boldsymbol{A}\boldsymbol{p}_k \rangle$ being negative actually implies that f is maximized at the value of α_k given by the conjugate gradient method.

However, before it is concluded that the IP algorithm is totally inappropriate

unless A is positive definite, two additional points should be observed. If the search directions p_k are close to the steepest descent directions r_k, then the functional f_{k+1} is likely to be maximized ($p_k^T A p_k$ negative) at α_k on precisely those steps when $f_k = r_k^T A^{-1} r_k$ is negative. In other words, if f_k is negative, then f_{k+1} is likely to be greater than f_k, and not 'more negative', as would be the case if the functional was indeed strictly minimized at each step. Hence the solution will usually, though not invariably, move off from x_k towards x_{k+1} in a direction in which the modulus of f is decreasing.

Secondly, although it is undoubtedly true that the functional f may change sign on a given step, the two points on the locus $x_k + \alpha p_k$ at which f is then zero have no particular significance as regards solving the system of equations ($r \neq 0$). At a true solution $r_{k+1} = 0$ the functional f_{k+1} must have a turning point in α_k, and it is this turning point which is located by the conjugate gradient method regardless of the sign of f_{k+1}.

3. The polynomial formulation for real systems

The previous section has considered the conjugate gradient method and its application to real symmetric systems. In fact it has been shown that the requirement (2.9) strictly requires only that LA is symmetric, with the matrix L, occurring in the product definition, itself also symmetric. Nevertheless it is clear that an extension of the method would be required in order to cope with general asymmetric systems, and this is the subject of the present section. However, the discussion starts by reformulating the conjugate gradient method in terms of polynomials.

The expression of the conjugate gradient method in terms of polynomials, either in A or in the eigenvalues of A, has been used by a number of authors (e.g. Jennings, 1977). The polynomial formulation will prove useful in the description of CG-type methods for real asymmetric (Section 3.2) and complex (Section 4) systems, and is essential for the derivation of the CG-squared method (Sonneveld, Wesseling, & Zeeuw, 1985).

3.1 *The Polynomial Formulation for Real Symmetric Systems*

The present section follows the notation of Sonneveld, whose starting point is the observation that, after k steps of the conjugate gradient method applied to a symmetric system, the vectors r_k and p_k can be written

$$r_k = \phi_k(A) r_0 \quad \text{and} \quad p_k = \theta_k(A) r_0, \tag{3.1}$$

where ϕ_k and θ_k are both kth order polynomials.

The conjugate gradient algorithm for real symmetric matrices can now be written in the polynomial form

$$\phi_{i+1} = \phi_i - \alpha_i \psi \theta_i, \qquad \theta_{i+1} = \phi_{i+1} + \beta_i \theta_i, \tag{3.2a,b}$$

where $\psi(A) \equiv A$.

The conjugacy requirements (2.2) can now be expressed as requirements on the products $[\gamma, \zeta]$ of such polynomials:

$$[\phi_i, \theta_j] = 0 \quad \text{for } i > j, \tag{3.3a}$$

$$[\psi\theta_i, \theta_j] = 0 \quad \text{for } i > j, \tag{3.3b}$$

where

$$[\gamma, \zeta] = \langle \gamma(\boldsymbol{A})\boldsymbol{r}_0, \zeta(\boldsymbol{A})\boldsymbol{r}_0 \rangle. \tag{3.3c}$$

The additional constraint (2.4) now takes the form

$$\alpha_k = \frac{[\phi_k, \theta_k]}{[\psi\theta_k, \theta_k]} \tag{3.4}$$

(and analogously for β_k (2.5)) and finally the additional constraints (2.6a) and (2.6b) take the form

$$[\theta_{k+1}, \theta_j] = [\theta_{k+1}, \psi\theta_j] \quad \text{for } j \leqslant k, \tag{3.5a}$$

$$[\theta_k, \psi\theta_j] = 0 \quad \text{for } j < k, \tag{3.5b}$$

which requirements are met by

$$[\gamma, \zeta] = \langle \gamma(\boldsymbol{A})\boldsymbol{r}_0, \zeta(\boldsymbol{A})\boldsymbol{r}_0 \rangle = \boldsymbol{r}_0^{\mathrm{T}}\gamma(\boldsymbol{A}^{\mathrm{T}})\boldsymbol{L}\zeta(\boldsymbol{A})\boldsymbol{r}_0 \tag{3.6}$$

if $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{L} = \boldsymbol{L}\boldsymbol{A}$; as in Section 2.1, the simplest (but not necessary) prescription will be that $\boldsymbol{A}$ is symmetric and must commute with $\boldsymbol{L}$.

3.2 *Real Asymmetric Systems*

It will be shown in this subsection how a generalization of the product definition (3.6) leads straight to the biconjugate gradient method (Fletcher, 1976) for real asymmetric systems. Suppose that, instead of (3.6), the product took the form

$$[\gamma, \zeta] = \boldsymbol{r}_0^{\mathrm{T}}\gamma(\boldsymbol{A})\boldsymbol{L}\zeta(\boldsymbol{A})\boldsymbol{r}_0 \tag{3.7}$$

$$\alpha_k = \frac{[\phi_k, \theta_k]}{[\theta_k, \psi\theta_k]}, \qquad \beta_k = \frac{[\phi_{k+1}, \theta_{k+1}]}{[\phi_k, \theta_k]}, \tag{3.8a,b}$$

where $\gamma(\boldsymbol{A}^{\mathrm{T}})$ has been replaced by $\gamma(\boldsymbol{A})$. The conditions (3.5) are now met by a simple commutativity relation

$$\boldsymbol{A}\boldsymbol{L} = \boldsymbol{L}\boldsymbol{A},$$

but without reference to any other properties of $\boldsymbol{A}$ and $\boldsymbol{L}$. The products in the definitions of α and β now take the form

$$[\phi_k, \theta_k] = \langle \phi_k(\boldsymbol{A}^{\mathrm{T}})\boldsymbol{r}_0, \theta_k(\boldsymbol{A})\boldsymbol{r}_0 \rangle, \tag{3.9a}$$

$$[\theta_k, \psi\theta_k] = \langle \theta_k(\boldsymbol{A}^{\mathrm{T}})\boldsymbol{r}_0, \boldsymbol{A}\theta_k(\boldsymbol{A})\boldsymbol{r}_0 \rangle. \tag{3.9b}$$

$\phi_k(\boldsymbol{A}^{\mathrm{T}})\boldsymbol{r}_0$ and $\theta_k(\boldsymbol{A}^{\mathrm{T}})\boldsymbol{r}$ are vectors in a biresidual system; if $\boldsymbol{r}$ and $\boldsymbol{p}$ are obtained by evaluating the polynomial system (3.2) at $\boldsymbol{A}$, then these new vectors

$$\bar{\boldsymbol{r}}_k = \phi_k(\boldsymbol{A}^{\mathrm{T}})\boldsymbol{r}_0 \quad \text{and} \quad \bar{\boldsymbol{p}}_k = \theta_k(\boldsymbol{A}^{\mathrm{T}})\boldsymbol{r}_0 \tag{3.10}$$

are obtained by evaluating at A^T. Thus at each step the vectors in a biresidual system

$$\bar{r}_{k+1} = \bar{r}_k - \alpha_k A^T \bar{p}_k, \qquad \bar{p}_{k+1} = \bar{r}_k + \beta_k \bar{p}_k \tag{3.11a,b}$$

(as well as r_{k+1} and p_{k+1}) will have to be calculated, in order that the scalars α_{k+1} and β_{k+1} can be computed.

Note finally that the conjugacy relations can be written in terms of the slightly more general product

$$[\gamma, \zeta] = \bar{r}_0^T \gamma(A) L \zeta(A) r_0, \tag{3.12}$$

with $\bar{r}_k$ and $\bar{p}_k$ appropriately defined in terms of some initial biresidual vector $\bar{r}_0$, without implying any particular relationship between r_0 and $\bar{r}_0$. The method obtained is the biconjugate gradient method of Fletcher (1976). The commutativity of A and L also ensures that

$$[\gamma, \zeta] = [\zeta, \gamma],$$

which will be utilized in the following subsection, which is concerned with the choice of L and the form of the functional f when the biconjugate gradient method is applied to asymmetric systems.

3.3 *The Functional Form for Real Asymmetric Systems*

Consider the functional $f = u^T K' v$, where u and v are linear in α, i.e.

$$u_{k+1} = u_k - \alpha_k s_k, \qquad v_{k+1} = v_k - \alpha_k t_k. \tag{3.13a,b}$$

The functional has a turning point with respect to α_k at the point where

$$\alpha_k = \frac{s_k^T K' v_k + u_k^T K' t_k}{2 s_k^T K' t_k}. \tag{3.14}$$

Comparing this with the form for α_k in the biconjugate gradient method (3.8a),

$$\alpha_k = \frac{\bar{r}_k^T L p_k}{\bar{p}_k^T L A p_k},$$

a functional can be sought by requiring the two forms to be equal. Following the same approach as in Section 2.2 and separately identifying the numerators and the denominators, the latter equivalence yields

$$S^T K' T = LA, \tag{3.15a}$$

where the new (fixed) matrices S and T are defined by

$$s_k = S \bar{p}_k, \qquad t_k = T p_k. \tag{3.15b}$$

Substituting these relations into the numerator gives

$$\bar{p}_k^T L A T^{-1} v_k + u_k (S^T)^{-1} L A p_k = 2 \bar{r}_k^T L p_k.$$

Rewriting the right-hand side as

$$\bar{p}_k^T L r_k + \bar{r}_k^T L p_k \quad \text{(using } [\gamma, \zeta] = [\zeta, \gamma]\text{)}$$

and separately identifying the two terms gives

$$\boldsymbol{A}\boldsymbol{T}^{-1}\boldsymbol{v}_k = \boldsymbol{r}_k \quad \text{and} \quad \boldsymbol{u}_k^{\mathrm{T}}(\boldsymbol{S}^{\mathrm{T}})^{-1}\boldsymbol{A} = \bar{\boldsymbol{r}}_k^{\mathrm{T}}.$$

In other words the functional f_k can be written as

$$\boldsymbol{u}_k^{\mathrm{T}}\boldsymbol{K}\boldsymbol{v}_k = \bar{\boldsymbol{r}}_k^{\mathrm{T}}\boldsymbol{A}^{-1}\boldsymbol{L}\boldsymbol{R}_k = \bar{\boldsymbol{r}}_k^{\mathrm{T}}\boldsymbol{K}\boldsymbol{r}_k, \tag{3.16}$$

where $\boldsymbol{K} = \boldsymbol{A}^{-1}\boldsymbol{L}$. In the case where $\bar{\boldsymbol{r}}_0 = \boldsymbol{r}_0$, $\boldsymbol{A}$ is symmetric and $\boldsymbol{L} = \boldsymbol{A}^{\mu}$, this reduces to the conjugate gradient method described previously. In the general asymmetric case, as in Section 2.2, there is no need to look for more general functionals than (3.16).

In looking at the nature of the turning point, it is seen that

$$\frac{\partial^2 f_{k+1}}{\partial \alpha_k^2} = 2\langle \bar{\boldsymbol{p}}_k, \boldsymbol{A}\boldsymbol{p}_k \rangle \quad \text{and} \quad f_k - f_{k+1} = \alpha_k^2 \langle \bar{\boldsymbol{p}}_k, \boldsymbol{A}\boldsymbol{p}_k \rangle. \tag{3.17a,b}$$

The equations of (3.16) and (3.17) can be rewritten if it is assumed for the moment that there are complete sets of left and right eigenvectors of $\boldsymbol{A}$, of which $\boldsymbol{e}_i$ (left) and $\boldsymbol{u}_i$ (right) are the eigenvectors corresponding to the distinct eigenvalue λ_i. The two sets of vectors have the property that

$$\boldsymbol{e}_i^{\mathrm{T}}\boldsymbol{u}_j = \delta_{ij}. \tag{3.18}$$

Note that, in the symmetric case, the left and right sets will be identical and will form an orthogonal basis. If

$$\bar{\boldsymbol{r}}_0 = \sum_i a_i \boldsymbol{e}_i \quad \text{and} \quad \boldsymbol{r}_0 = \sum_j b_j \boldsymbol{u}_j, \tag{3.19}$$

and $\boldsymbol{L} = \boldsymbol{A}^{\mu}$ (as the most generally applicable way of meeting the commutativity requirement) and $\boldsymbol{K} = \boldsymbol{A}^{\nu}$ (from (3.15)), then the equations (3.16) and (3.17) can be expressed in the form

$$f_{k+1} = \sum_i a_i b_i \phi_{k+1}^2(\lambda_i)\lambda_i^{\nu} \tag{3.20a}$$

$$f_k - f_{k+1} = \alpha_k^2 \sum_i a_i b_i \theta_{k+1}^2(\lambda_i)\lambda_i^{\mu+1} \tag{3.20b}$$

$$\frac{\partial^2 f_{k+1}}{\partial \alpha_k^2} = 2\sum_i a_i b_i \theta_{k+1}^2(\lambda_i)\lambda_i^{\mu+1}. \tag{3.20c}$$

In general the coefficients a_i will not equal b_i even if $\bar{\boldsymbol{r}}_0 = \boldsymbol{r}_0$, so monotonic convergence cannot be proved for the biconjugate gradient method applied to asymmetric systems. (In the symmetric case, with $\bar{\boldsymbol{r}}_0 = \boldsymbol{r}_0$, the coefficients will be equal and the results of Section 2.2 are recovered.) Indeed the situation is much more like the $\boldsymbol{r}^{\mathrm{T}}\boldsymbol{A}^{-1}\boldsymbol{r}$ IP method applied to real symmetric indefinite systems, although f can now potentially be negative even if $\mu = 1$, $\nu = 0$ is selected (the choices corresponding to the MR method of Section 2.2) and/or the eigenvalues of $\boldsymbol{A}$ are all positive.

The functional f is of course still quadratic in α_k; if for a given search direction f may change sign, then there are two points at which $f = 0$. At these points $\bar{\boldsymbol{r}}$ and $\boldsymbol{K}\boldsymbol{r}$ must either be orthogonal, or one or both of the vectors are zero. It is no

longer the case that, if a true solution $\boldsymbol{r}^* = \boldsymbol{0}$ lies on the locus of points $\boldsymbol{x}^* = \boldsymbol{x}_k + s\boldsymbol{p}_k$, then the functional must of necessity have its turning point at that solution; for that to be the case, $\bar{\boldsymbol{r}}_{k+1}$ would have either to be zero at the same point or lying in the space orthogonal to $\boldsymbol{K}\boldsymbol{r}_{k+1}$ as $\boldsymbol{r}_{k+1} \to 0$.

A number of other results relating to the biconjugate gradient method applied to real asymmetric systems can be derived as special cases from the analysis of the general complex case which follows.

4. The general complex case

4.1 *Two Possible Formulations*

The biconjugate gradient method can be applied in precisely the form of Section 3 to general complex systems (for which $\boldsymbol{A}$, $\boldsymbol{x}$, and $\boldsymbol{b}$ may be complex, and no particular symmetry of $\boldsymbol{A}$ is assumed), simply by allowing all the quantities to become complex. This is the complex generalization of the biconjugate gradient method used by Cullum and Willoughby, as reported by Jacobs (private communication). Of particular interest in the present section is the behaviour of the complex functional $f = \bar{\boldsymbol{r}}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{r}$ in the region of $\alpha = \alpha_k$.

It should be noted immediately that the Jacobs (1980) generalization of the biconjugate gradient method employed a functional $f = \hat{\boldsymbol{r}}^{\mathrm{H}}\boldsymbol{K}\boldsymbol{r}$, with the biresidual system $(\hat{\boldsymbol{r}}, \hat{\boldsymbol{p}})$ employing the formulae

$$\hat{\boldsymbol{r}}_{i+1} = \hat{\boldsymbol{r}}_i - \alpha_i^* \boldsymbol{A}^{\mathrm{H}}\hat{\boldsymbol{p}}_i, \qquad \hat{\boldsymbol{p}}_{i+1} = \hat{\boldsymbol{r}}_{i+1} + \beta_i^* \hat{\boldsymbol{p}}_i, \tag{4.1a,b}$$

$$\alpha_i = \frac{\langle \hat{\boldsymbol{r}}_i, \boldsymbol{p}_i \rangle}{\langle \hat{\boldsymbol{p}}_i, \boldsymbol{A}\boldsymbol{p}_i \rangle} = \frac{\hat{\boldsymbol{r}}_i^{\mathrm{H}}\boldsymbol{L}\boldsymbol{p}_i}{\hat{\boldsymbol{p}}_i^{\mathrm{H}}\boldsymbol{L}\boldsymbol{A}\boldsymbol{p}_i}, \tag{4.2a}$$

$$\beta_i = \frac{\langle \hat{\boldsymbol{r}}_{i+1}, \boldsymbol{p}_{i+1} \rangle}{\langle \hat{\boldsymbol{r}}_i, \boldsymbol{p}_i \rangle} = \frac{\hat{\boldsymbol{r}}_{i+1}^{\mathrm{H}}\boldsymbol{L}\boldsymbol{p}_{i+1}}{\hat{\boldsymbol{r}}_i^{\mathrm{H}}\boldsymbol{L}\boldsymbol{p}_i}, \tag{4.2b}$$

where $^{\mathrm{H}}$ denotes Hermitian and * denotes complex conjugate. It is immediately obvious that this system is equivalent to the $\bar{\boldsymbol{r}}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{r}$ system above if $\hat{\boldsymbol{r}}_0 = \bar{\boldsymbol{r}}_0^*$. In other words any algorithm expressed in one system has an exactly equivalent algorithm expressible in the other, the link being established via the biresidual starting vectors.

Part of the rationale behind favouring the $\hat{\boldsymbol{r}}^{\mathrm{H}}\boldsymbol{K}\boldsymbol{r}$ system comes from the desire to keep f as real and positive as possible. If $\boldsymbol{K} = \boldsymbol{I}$ and $\boldsymbol{A}$ is Hermitian, then the choice $\hat{\boldsymbol{r}}_0 = \boldsymbol{r}_0$ gives

$$f = [\phi(\boldsymbol{A})\boldsymbol{r}_0]^{\mathrm{H}}\phi(\boldsymbol{A})\boldsymbol{r}_0, \tag{4.3}$$

which will be real; thus departures from the positive real axis signify departures of $\boldsymbol{A}$ from a Hermitian matrix. Note that in the complex symmetric, as opposed to the Hermitian, case, the choice $\bar{\boldsymbol{r}}_0 = \boldsymbol{r}_0$ in the $\bar{\boldsymbol{r}}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{r}$ system gives

$$f = [\phi(\boldsymbol{A})\boldsymbol{r}_0]^{\mathrm{T}}\phi(\boldsymbol{A})\boldsymbol{r}_0, \tag{4.4}$$

which does not in general lie on the real axis. So the $\hat{\boldsymbol{r}}^{\mathrm{H}}\boldsymbol{K}\boldsymbol{r}$ system appears to have an advantage over its rival in that for a particular class of complex matrices the functional will be real.

However, it is important to consider whether or not having f near the real axis has any particular merit. In the next subsection it will be assumed that the method has a functional and inner product,

$$f = \bar{\boldsymbol{r}}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{r}, \qquad (\bar{\boldsymbol{x}}, \boldsymbol{y}) = \bar{\boldsymbol{x}}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{y}, \tag{4.5a,b}$$

in the knowledge that results will carry over to the $\hat{\boldsymbol{r}}^{\mathrm{H}}\boldsymbol{K}\boldsymbol{r}$ case via $\hat{\boldsymbol{r}}_0 = \bar{\boldsymbol{r}}_0^*$.

Note also that in the polynomial formulation the $\bar{\boldsymbol{r}}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{r}$ system corresponds to starting from a product definition of the form

$$[\gamma, \zeta] = \bar{\boldsymbol{r}}_0^{\mathrm{T}}\gamma(\boldsymbol{A})\boldsymbol{L}\zeta(\boldsymbol{A})\boldsymbol{r}_0 \tag{4.6a}$$

(cf. (3.7) in the real case), whereas the $\hat{\boldsymbol{r}}^{\mathrm{H}}\boldsymbol{K}\boldsymbol{r}$ system starts from

$$[\gamma, \zeta] = \hat{\boldsymbol{r}}_0^{\mathrm{H}}\gamma(\boldsymbol{A})\boldsymbol{L}\zeta(\boldsymbol{A})\boldsymbol{r}_0. \tag{4.6b}$$

Clearly there are other ways of defining products but they do not yield methods of interest; for example

$$[\gamma, \zeta] = \boldsymbol{r}_0^{\mathrm{H}}\gamma(\boldsymbol{A}^{\mathrm{H}})\boldsymbol{L}\zeta(\boldsymbol{A})\boldsymbol{r}_0 \tag{4.6c}$$

would lead to a 'conjugate gradient method' with a real functional and no separate biresidual system; however, the method would not possess any conjugacy properties (cf. (3.6)) unless $\boldsymbol{A}$ was Hermitian.

4.2 *The Complex Functional at the Turning Point*

Consider $f_{k+1} = \bar{\boldsymbol{r}}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{r}$ as a function of both α_{r} and α_{i}, the real and imaginary parts of α (with the subscripts k and $k+1$ now dropped except when explicitly necessary); f_{k+1} itself has real and imaginary parts f_{r} and f_{i} (which may be negative). Writing $(\bar{\boldsymbol{p}}, \boldsymbol{A}\boldsymbol{p})$ as σ and $(\bar{\boldsymbol{r}}, \boldsymbol{p})$ as ρ, the derivatives of f_{k+1} with respect to the two components of α are given by

$$\partial f_{\mathrm{r}}/\partial\alpha_{\mathrm{r}} = 2(\alpha\sigma - \rho), \qquad \partial f_{\mathrm{i}}/\partial\alpha_{\mathrm{i}} = 2(\alpha\sigma - \rho). \tag{4.7a,b}$$

It can be shown that both real and imaginary components f_{r} and f_{i} have turning points with respect to α_{r} and α_{i} at $\alpha = \rho/\sigma$.

Also the following second derivatives are obtained:

$$\partial^2 f_{\mathrm{r}}/\partial\alpha_{\mathrm{r}}^2 = 2\sigma_{\mathrm{r}}, \qquad \partial^2 f_{\mathrm{i}}/\partial\alpha_{\mathrm{r}}^2 = 2\sigma_{\mathrm{i}}, \tag{4.8a,b}$$

$$\partial^2 f_{\mathrm{r}}/\partial\alpha_{\mathrm{i}}^2 = -2\sigma_{\mathrm{r}}, \qquad \partial^2 f_{\mathrm{i}}/\partial\alpha_{\mathrm{i}}^2 = -2\sigma_{\mathrm{i}}, \tag{4.8c,d}$$

$$\partial^2 f_{\mathrm{r}}/\partial\alpha_{\mathrm{r}}\partial\alpha_{\mathrm{i}} = -2\sigma_{\mathrm{i}}, \qquad \partial^2 f_{\mathrm{i}}/\partial\alpha_{\mathrm{r}}\partial\alpha_{\mathrm{i}} = 2\sigma_{\mathrm{r}}. \tag{4.8e,f}$$

The striking feature here is that if σ_{r} is positive, say, then f_{r} is minimized with respect to α_{r} but maximized with respect to α_{i}. Thus, whereas in the real cases (Sections 2 and 3) the real functional f would have either a local minimum or a local maximum at $\alpha = \rho/\sigma$, in the complex case the real components of the functional, f_{r} and f_{i}, always exhibit more complicated behaviour in the region of $\alpha = \rho/\sigma$: if a component is a maximum with respect to α_{r} then it must be a minimum with respect to α_{i}, and vice versa.

The functional components will be returned to in due course, but for the

moment consider a new functional $\|f\|^2$, where

$$\|f\|^2 = (f_r)^2 + (f_i)^2; \tag{4.9}$$

this has the distinct advantage of being a real nonnegative functional. It can be shown, not unnaturally, that $\|f\|^2$ has turning points with respect to α_r and α_i at $\alpha = \rho/\sigma$ and $f = 0$. (In general $f = 0$ will not be a turning point in f any more than it was in the real asymmetric case of Section 3.)

This completes the analysis necessary to show that the f turning point at $\alpha = \rho/\sigma$ has exactly the same status as in the real asymmetric case. There is no particular requirement to keep f near to the real positive axis, as the turning point in $\|f\|^2$ shows.

4.3 *Concluding Remarks on the Complex Functional*

Expanding f_r and f_i about $\alpha - \rho/\sigma$ in terms of the components of $\alpha' = \alpha - \rho/\sigma = \xi + i\eta$, it can be shown that the two curves $f_r = 0$ and $f_i = 0$ are rectangular hyperbolae in the Argand plane ξ, η. The quadrants in which $f = 0$ solutions will occur will then depend on the signs of F_r and F_i, the real and imaginary parts of f evaluated at $\alpha = \rho/\sigma$.

However, because of an offset of $\pi/4$ between the two sets of major axes, there are always two points in the Argand plane ξ, η at which both f_r and f_i are zero, i.e. at which the functional $f = \bar{\boldsymbol{r}}^T \boldsymbol{K} \boldsymbol{r}$ disappears. The possibility of there being two vanishing points, in the real case, has become obligatory in the complex case, with the turning point at ρ/σ being a saddle point with respect to α_r and α_i for both real and imaginary parts of the complex functional.

To summarize, the standard conjugate gradient method for real symmetric systems is conventionally regarded as minimizing a functional which is regarded as a good residual measure, e.g. '$h^2 = \boldsymbol{r}^T \boldsymbol{r}$'. It was shown in Section 2 that in some circumstances the functional might become negative, and that one might have to speak of turning points of the functional as opposed to just minima. It was further shown in Section 3 that in the real asymmetric case the biconjugate gradient method may exhibit similar functional behaviour (value changing sign, maxima as well as minima occurring) regardless of the choice of algorithm or the definiteness or otherwise of the matrix.

The analysis of Section 4 has shown a further weakening of this role in the complex case. The functional $f = \bar{\boldsymbol{r}}^T \boldsymbol{K} \boldsymbol{r}$ is itself a complex function, and the real functional $\|f\|^2$ exhibits behaviour quite unlike that of a 'good residual measure' such as $\boldsymbol{r}^H \boldsymbol{r}$. The prescription for α_k in the complex biconjugate gradient method now identifies a saddle point of $\|f\|^2$ with respect to the two component directions (α_r, α_i) and the directions of steepest descent away from this point will lead to points in the Argand plane at which $\|f\|^2$ vanishes completely. As well as suggesting that functional behaviour does not provide a good description of the workings of the complex biconjugate gradient method, this may also have implications for any potential use of complex CG-type methods in minimization problems.

The comments on the choice of functional f for the complex biconjugate

gradient method carry over to the complex extension of the CG-squared method (Sonneveld, Wesseling, & Zeeuw, 1985). The method is implemented simply by allowing A, x, and b to be complex in the standard algorithm, changes in the functional f being effected via $\bar{r}_0$.

5. Acknowledgement

This paper is published by permission of the Central Electricity Generating Board. The work was performed at the Central Electricity Research Laboratories, Leatherhead. The author would like to thank the two anonymous referees for their detailed comments and helpful suggestions on earlier drafts of this paper.

References

Concus, P., Golub, G. H., & O'Leary, D. P. 1976 A generalized conjugate gradient method for the numerical solution of elliptic partial differential equations. In: *Sparse Matrix Computations* (J. R. Bunch and D. J. Rose, Eds). Academic Press.

Faber, V., & Manteuffel, T. 1984 Necessary and sufficient conditions for the existence of a conjugate gradient method. *SIAM J. Numer. Anal.* **21,** 352–362.

Fletcher, R. 1976 Conjugate gradient methods for indefinite systems. In: *Proc. of Dundee Conference on Numerical Analysis, 1975.* (G. A. Watson, Ed.). Springer-Verlag.

Hestenes, M. R., & Stiefel, E. 1952 Methods of conjugate gradients for solving linear systems. *J. Res. Nat. Bur. Standards* **49,** 409–436.

Jackson, C. P., & Robinson, P. C. 1982 A numerical study of various algorithms related to the preconditioned conjugate gradient method. AERE Harwell Report No. TP 950. (Published in revised form in *Int. J. for Num. Meth. in Eng.* **21,** 1315–1338 (1985).)

Jacobs, D. A. H. 1980 Generalizations of the conjugate gradient method for solving non-symmetric and complex systems of algebraic equations. CERL Note No. RD/L/N 70/80.

Jennings, A. 1977 Influence of the eigenvalue spectrum on the convergence rate of the conjugate gradient method. *J. Inst. Maths. Applics.* **20,** 61–72.

Lanczos, C. 1950 An iteration method for the solution of the eigenvalue problem of linear differential and integral operators. *J. Res. Nat. Bur. Standards* **45,** 225–282.

Markham, G. 1987 Preconditioned conjugate-gradient-type methods for robust and efficient iterative solution packages. CERL Report No. TPRD/L/3217/R87.

Meijerink, J. A., & van der Vorst, H. A. 1977 An iterative solution method for systems of which the coefficient matrix is a symmetric M-matrix, *Maths. Comp.* **31,** 148–162.

Reid, J. K. 1971 On the method of conjugate gradients for the solution of large sparse systems of equations. In: *Large Sparse Sets of Linear Equations* (J. K. Reid, Ed.). Academic Press.

Sonneveld, P., Wesseling, P., & Zeeuw, P. M. 1985 Multigrid and conjugate gradient methods as convergence acceleration techniques. In: *Multigrid Methods for Integral and Differential Equations.* (D. J. Paddon and H. Holstein, Eds). Oxford University Press.

Appendix

In the main body of the text, the requirement

$$A^{\mathrm{T}}L = LA \tag{2.9}$$

was satisfied by the prescription that $\boldsymbol{A}$ is symmetric and that $\boldsymbol{A}$ commutes with $\boldsymbol{L}$. It is easy to determine such sufficient conditions for (2.9) to be satisfied; but what are the weakest (i.e. necessary) conditions?

Starting from the relation $\boldsymbol{A}^{\mathsf{T}} = \boldsymbol{L}\boldsymbol{A}\boldsymbol{L}^{-1}$, it can be shown that if the left and right eigenvectors for the ith eigenvalue of $\boldsymbol{A}$ are $\boldsymbol{x}^{i\mathsf{T}}$ and $\boldsymbol{y}^{i}$, then

$$\boldsymbol{y}^{i} = \boldsymbol{L}^{-\mathsf{T}}\boldsymbol{x}^{i} \quad \text{and} \quad \boldsymbol{y}^{i\mathsf{T}}\boldsymbol{L}^{\mathsf{T}} = \boldsymbol{x}^{i\mathsf{T}}, \tag{A.1}$$

i.e. $\boldsymbol{L}$ must be symmetric.

If $\boldsymbol{L}$ is symmetric, then (2.9) implies immediately that $\boldsymbol{L}\boldsymbol{A}$ is itself symmetric, i.e.

$$\boldsymbol{A} = \boldsymbol{L}^{-1}\hat{\boldsymbol{A}}, \tag{A.2}$$

where $\hat{\boldsymbol{A}}$ is a symmetric matrix. To recapitulate, (2.9) requires that $\boldsymbol{L}$ is symmetric and thereby that $\boldsymbol{A}$ is related to a symmetric matrix via (A.2).

If, in addition, $\boldsymbol{L}$ is real, symmetric, and positive-definite, then it may be factored into

$$\boldsymbol{L} = \boldsymbol{C}_1^{\mathsf{T}}\boldsymbol{C}_1, \tag{A.3}$$

and it follows that

$$\boldsymbol{A} = \boldsymbol{C}_1^{-1}\tilde{\boldsymbol{A}}\boldsymbol{C}_1, \tag{A.4}$$

where $\tilde{\boldsymbol{A}}$ is a symmetric matrx. This corresponds to the result given by Faber and Manteuffel (1984). However, they start from a requirement that $\boldsymbol{L}$ is positive-definite, in order that their error norm is an inner product, and so reach a condition which in the present context is sufficient rather than necessary.

In the case (A.4), Faber and Manteuffel note that, with a change of basis, a CG method applied to the matrix (A.4) can be expressed as a method of the type of Concus, Golub, & O'Leary (1976) applied to $\tilde{\boldsymbol{A}}$.

Whilst the result in this Appendix does enable novel methods to be constructed for matrices $\boldsymbol{A}$ with a particular structure (for example, the product of two symmetric matrices), for general matrices $\boldsymbol{A}$ the satisfaction of (2.9) will reduce to the prescription in the main text for computationally interesting methods. Furthermore, unless it is known that $\boldsymbol{A} = \boldsymbol{S}^{\kappa}$, where κ is some integer and $\boldsymbol{S}$ is explicitly available, useful methods will have integral μ in the relation $\boldsymbol{L} = \boldsymbol{A}^{\mu}$.