A Theoretical Introduction to Numerical Analysis

Problem (6.75) is a typical problem of solving an overdetermined system of linear algebraic equations in the sense of the least squares. Indeed, the matrix $H_p$ has $p + 1$ rows and $p$ columns, i.e., there are $p + 1$ equations and only $p$ unknowns. Solutions of such systems can, generally speaking, only be found in the weak sense, in particular, in the sense of the least squares. The concept of weak, or generalized, solutions, as well as the methods for their computation, are discussed in Chapter 7.

In the meantime, let us mention that if the Arnoldi orthogonalization process (6.71) does not terminate on or before $k = p$, then the minimization problem (6.75) has a unique solution. As shown in Section 7.2, this is an implication of the matrix $H_p$ being full rank. The latter assertion, in turn, is true because according to formula (6.72), equation number $k$ of (6.71) can be recast as follows:

$$Au_k = v_k + \sum_{j=1}^{k} h_{jk}u_j = u_{k+1} \|v_k\| + \sum_{j=1}^{k} h_{jk}u_j = \sum_{j=1}^{k+1} h_{jk}u_j,$$

which means that $h_{k+1,k} = \|v_k\| \neq 0$. As such, all columns of the matrix $H_p$ are linearly independent since every column has an additional non-zero entry $h_{k+1,k}$ compared to the previous column. Consequently, the vector $w^{(p)}$ can be obtained as a solution to the linear system (see Theorem 7.1 on page 215):

$$H_p^T H_p w^{(p)} = -H_p^T q^{(p+1)}.$$  

The solution $w^{(p)}$ of system (6.76) is unique because the matrix $H_p^T H_p$ is non-singular (Exercise 2). In practice, one does not normally reduce the least squares minimization problem (6.75) to linear system (6.76) since this reduction may lead to the introduction of large additional errors (amplification of round-off). Instead, problem (6.75) is solved using the $QR$ factorization of the matrix $H_p$, see Section 7.2.2.

Note that in the course of the previous analysis we assumed that the dimension of the Krylov subspaces $K_p(A,r^{(0)})$ would increase monotonically as a function of $p$. Let us now see what happens if the alternative situation takes place, i.e., if the Arnoldi process terminates prematurely.

**THEOREM 6.5**

Let $p$ be the smallest integer number for which the Arnoldi process (6.71) terminates:

$$Au_p - \sum_{j=1}^{p} (u_j, Au_p)u_j = 0.$$

Then the corresponding iterate yields the exact solution:

$$x^{(p)} = x = A^{-1}f.$$

**PROOF** By hypothesis of the theorem, $Au_p \in K_p$. Consequently, $AK_p \subset K_p$. This implies [cf. formula (6.72)]:

$$AU_p = U_p \hat{H},$$  

(6.77)
where \( \tilde{H} \) is a \( p \times p \) matrix. This matrix is non-singular because otherwise it would have had linearly dependent rows. Then, according to formula (6.77), each column of \( AU_p \) could be represented as a linear combination of only a subset of the columns from \( U_p \) rather than as a linear combination of all of its columns. This, in turn, means that the Arnoldi process terminates earlier than \( k = p \), which contradicts the hypothesis of the theorem.

For the norm of the residual \( r^{(p)} = Ax^{(p)} - f \) we can write:

\[
\|r^{(p)}\| = \|Ax^{(p)} - f\| = \|A(x^{(p)} - x^{(0)}) + r^{(0)}\|.
\] (6.78)

Next, we notice that since \( x^{(p)} \in N_p \), then \( x^{(p)} - x^{(0)} \in K_p(A,r^{(0)}) \), and consequently, \( \exists w \in \mathbb{R}^p : x^{(p)} - x^{(0)} = U_p w \), because the columns of the matrix \( U_p \) provide a basis in the space \( K_p(A,r^{(0)}) \). Let us also introduce a \( p \)-dimensional vector \( q^{(p)} \) with real components:

\[
q^{(p)} = \left[ \|r^{(0)}\|, 0, 0, \ldots, 0 \right]^T,
\]

so that \( U_p q^{(p)} = r^{(0)} \). Then, taking into account equality (6.77), as well as the orthonormality of the columns of the matrix \( U_p \): \( U_p^T U_p = I_p \), we obtain from formula (6.78):

\[
\|r^{(p)}\| = \|U_p(q^{(p)} + Hw)\| = \|q^{(p)} + \tilde{H}w\|.
\] (6.79)

Finally, we recall that on every iteration of GMRES we minimize the norm of the residual: \( \|r^{(p)}\| \rightarrow \min \). Then, we can simply set \( w = -\tilde{H}^{-1}q^{(p)} \) in formula (6.79), which immediately yields \( \|r^{(p)}\| = 0 \). This is obviously a minimum of the norm, and it implies \( r^{(p)} = 0 \), i.e., \( Ax^{(p)} = f \Rightarrow x^{(p)} = A^{-1}f = x \).

We can now summarize two possible scenarios of behavior of the GMRES iteration. If the Arnoldi process terminates prematurely at some \( p < n \) (\( n \) is the dimension of the space), then, according to Theorem 6.5, \( x^{(p)} \) is the exact solution to \( Ax = f \). Otherwise, the maximum number of iterations that the GMRES can perform is equal to \( n \). Indeed, if the Arnoldi process does not terminate prematurely, then \( U_n \) will contain \( n \) linearly independent vectors of dimension \( n \) and consequently, \( K_n(A,r^{(0)}) = \mathbb{R}^n \). As such, the last minimization of the residual in the sense of (6.73) will be performed over the entire space \( \mathbb{R}^n \), which obviously yields the exact solution \( x = A^{-1}f \). Therefore, technically speaking, the GMRES can be regarded as a direct method for solving \( Ax = f \), in much the same way as we regarded the method of conjugate gradients as a direct method (see Section 5.6).

In practice, however, the GMRES is never used in the capacity of a direct solver, it its only used as an iterative scheme. The reason is that for high dimensions \( n \) it is feasible to perform only very few iterations, and one should hope that the approximate solution obtained after these iterations will be sufficiently accurate in a given context. The limitations for the number of iterations come primarily from the large storage...
requirements for the Krylov subspace basis $U_p$, as well as from the increasing computational costs associated with solving the sequence of the least squares problems (6.75) for $p = 1, 2, \ldots$. Note that the method of conjugate gradients does not entail this type of limitations because its descent directions are automatically $A$-orthogonal. These additional constraints that characterize the GMRES are the "price to pay" for its broader applicability and ability to handle general matrices $A$, as opposed to only symmetric positive definite matrices, for which the method of conjugate gradients works. However, another inherent limitation of the GMRES fully translates to the method of conjugate gradients (or the other way around). Indeed, the exact solution of $Ax = f$ can be obtained by means of the GMRES only if the computations are conducted with infinite precision. On a finite precision computer the method is prone to numerical instabilities. No universal cure is available for this problem; some partial remedies, such as restarts, are discussed, e.g., in [Saa03].

Exercises

1. Prove that the Arnoldi process (6.71) indeed yields an orthonormal system of vectors: $u_1, u_2, \ldots$.
2. Prove that the system matrix $H_p^T H_p$ in (6.76) is symmetric positive definite.

6.4 Multigrid Iterations

We have seen previously that in many cases numerical methods with superior performance can be developed at the expense of narrowing down the class of problems that they are designed to solve. In the framework of direct methods, examples include the tri-diagonal elimination (Section 5.4.2), as well as the methods that exploit the finite Fourier series and the FFT (Section 5.7). In the framework of iterative methods, a remarkable example of that kind is given by multigrid.

Multigrid methods have been originally developed for solving elliptic boundary value problems discretized by finite differences (Chapter 12). A key distinctive characteristic of these methods is that the number of iterations required for reducing the initial error by a prescribed factor does not depend on the dimension of the grid at all. Accordingly, the required number of arithmetic operations is directly proportional to the grid dimension $N^2$, where $N$ is the number of grid nodes along one coordinate direction and $n$ is the dimension of the space $\mathbb{R}^n$. This is clearly an asymptotically unimprovable behavior, because the overall number of quantities to be computed (solution values on the grid) is also directly proportional to the grid dimension. As the grid dimension determines the condition number of the corresponding matrix,$^2$

$^2$The latter is typically inversely proportional to the square of the grid size: $\mu = \mathcal{O}(h^{-2})$, see formula (5.115), i.e., $\mu = \mathcal{O}(N^2)$. 
