Projection methods for solving sparse linear systems

By R. P. Tewarson*

Some methods of successive approximation for the solution of simultaneous linear equations are discussed. The coefficient matrix A of the linear system is assumed to be sparse. It is shown that savings in the computer storage and the computing time are possible, if there exists a subset of the rows (columns) of A, consisting of only orthogonal rows (columns). Such savings are also possible, if for some permutation matrices P and Q, PAQ has a particular structure, viz., singly bordered block diagonal form. It is shown that the set of orthogonal rows (columns) of A, as well as P and Q can be determined by using some results from graph theory (e.g., incidence matrices, row and column graphs, points of attachment). Geometrical interpretations of the methods and their inter-relatiohip are given.

(First received December 1967 and in revised form February 1968)

1. Introduction

Let us denote a system of simultaneous linear equations by

$$Ax = b, \tag{1.1}$$

where A is an $n \times n$ sparse matrix and both x and b are n element column vectors. Evidently the exact solution of (1.1) is $x = A^{-1}b$. Let us consider the following. Given an initial approximate solution x_0 of (1.1), form a sequence of approximations

$$x_{k+1} = x_k + C_k r_k, \; k = 0, 1, 2, \ldots, \tag{1.2}$$

where C_k is some matrix, and

$$s_k = x - x_k, \; r_k = b - Ax_k. \tag{1.3}$$

From (1.2) and (1.3), we have

$$s_{k+1} = (I - C_k A)s_k. \tag{1.4}$$

If each $I - C_k A$ is a projector (Hermitian and idempotent), then (1.2) is a method of projection. It is shown in (Householder, 1964, p. 98) that, for a given x_k, if we minimise the quantity $|s_{k+1}|^2$, then (1.2) can be written as

$$x_{k+1} = x_k + \frac{v_k^T r_k}{v_k^T A A^T v_k} A^T v_k, \tag{1.5}$$

where v_k is an n element column vector.

2. The Kaczmarz method

In (1.5), if we take $v_k = e_i$ (the ith column of the identity matrix), then we get

$$x_{k+1} = x_k + \frac{r_k^i}{|A_i|^2} A_i^T, \tag{2.1}$$

where r_k^i denotes the ith element of r_k and A_i the ith row of A. The scheme given by (2.1) is due to Kaczmarz (1937).

It is easy to give a geometrical interpretation of the Kaczmarz method (Bodewig, 1959, p. 186; Tomkins, 1956, p. 454) as follows. Let x_k be a given approximation to x. The system (1.1) can also be written as

$$A_i x = b_i, \; i = 1, 2, \ldots, n, \tag{2.2}$$

where b_i is the ith component of b. Each of the above n equations represents an n–1 dimensional hyperplane in the n dimensional Euclidian space E^n. The solution x is the common point of intersection of all such hyperplanes. Let x_{k+1} be the projection of x_k on $A_i x = b_i$. Since A_i^T as well as $x_{k+1} - x_k$ are both perpendicular to the hyperplane, $A_i x = b_i$,

therefore

$$x_{k+1} - x_k = \lambda A_i^T. \tag{2.3}$$

The fact that x_{k+1} lies on $A_i x = b_i$ gives

$$A_i x_{k+1} = b_i. \tag{2.4}$$

Premultiplying (2.3) by A_i, we have

$$A_i x_{k+1} - A_i x_k = \lambda A_i A_i^T.$$

Using (2.4) and the fact that $A_i A_i^T = |A_i|^2$, we get

$$\lambda = \frac{b_i - A_i x_k}{|A_i|^2} = \frac{r_k^i}{|A_i|^2},$$

which, on substitution in (2.3), yields (2.1). It is evident that x_{k+1} is closer to x than x_k; thus the convergence is assured.

3. Structure of A and the Kaczmarz method

If the matrix A is sparse, then it is usually possible to find a permutation matrix P, such that

$$PA = \begin{bmatrix} R \\ N \end{bmatrix}, \tag{3.1}$$

where R is $m \times n$ and N is $(n - m) \times n$; furthermore, the rows of R are orthogonal. Since finding all the mutually orthogonal rows of A involves a large amount of computational effort, we shall restrict ourselves to the following: If any pair of rows of A do not have any non-zero elements in the same column of A, then we shall call them 'disjoint' (Tewarson, 1967*b*). Clearly, disjoint rows are orthogonal. In sparse matrices, almost all of the rows which are orthogonal, are usually disjoint. Thus, we shall assume that all the rows of R are disjoint. If a pair of rows is orthogonal, but not disjoint, then one of them should be moved to N. An algorithm for enumerating the disjoint rows of A (viz.,

---

* *State University of New York at Stony Brook, N.Y.* This research was supported in part by National Aeronautics and Space Administration, Washington, D.C. Grant No. NGR-33-015-016.

the determination of P) is given in (Tewarson, 1967a). The algorithm makes use of the matrix BB^T, where B is the incidence matrix associated with A, and Boolean addition is used in the matrix multiplication. An alternative interpretation of the disjoint rows of A can be given as follows. Consider the row graph of the matrix A. The nodes are the rows of A and any two nodes are said to be directly connected by an edge iff the corresponding rows are not disjoint (Tewarson, 1967c). In view of the above-mentioned interpretation, the rows of R are not connected directly to each other; viz., there do not exist paths of unit length between any two nodes associated with the rows of R. The graph theoretic interpretation opens up the possibility of using the vast literature of graph theory. In any case, having determined P, we have from (1.1) and (3.1)

$$PA\,x = \begin{bmatrix} R \\ N \end{bmatrix} x = Pb = \begin{bmatrix} c \\ d \end{bmatrix} \text{ (say)}, \tag{3.2}$$

which is equivalent to

$$Rx = c \text{ and } Nx = d. \tag{3.3}$$

Evidently, c is $m \times 1$ and d is $(n - m) \times 1$. Let the ith row of R be denoted by R_i. Then the Kaczmarz method can be stated as follows:

Theorem 3.1. Let D be a diagonal matrix with $1/|R_i|$; $i = 1, 2, \ldots, m$ as its diagonal elements. If the scheme (2.1) is applied m times to the system (3.3), using the rows of R, and x_k as the initial approximation, then

$$x_{k+m} = x_k + R^T D^2(c - Rx_k). \tag{3.4}$$

Proof. First, let us normalise the rows of R as follows. From (3.3) we have $DRx = Dc$ or $Fx = g$ (where $F = DR$ and $Dc = g$). Evidently, $F_i F_j^T = 0$, $i \neq j$ and $F_i F_i^T = 1$. Now let

$$x_{k+j} = x_k + \sum_{i=1}^{j} (g_i - F_i x_k) F_i^T. \tag{3.5}$$

If we use the above equation in (2.1) with F_{j+1}, we have

$$x_{k+j+1} = x_{k+j} + (g_{j+1} - F_{j+1} x_{k+j}) F_{j+1}^T.$$

Now, if we substitute the value of x_{k+j} given by (3.5) in the above equation, then in view of the fact that $F_{j+1}^T F_i = 0 \;\; i \leqslant j$, we have

$$x_{k+j+1} = x_{k+j} + \sum_{k=1}^{j+1} (g_i - F_i x_k) F_i^T.$$

If x_k and F_1 are used in (2.1), we get

$$x_{k+1} = x_k + (g_i - F_i x_k) F_1^T.$$

Thus, (3.5) holds for $j = 1$ and whenever it holds for j, it holds for $j + 1$, hence

$$x_{k+m} = x_k + \sum_{i=1}^{m} (g_i - F_i x_k) F_i^T = x_k + F^T(g - Fx_k)$$
$$= x_k + R^T D^T (Dc - DRx_k).$$

Since $D = D^T$, the above equation is the same as (3.4), which completes the proof of the theorem.

From the above theorem, we see that considerable saving in the computing time is possible, if for the orthogonal (disjoint) rows, instead of equation (2.1), (3.4) is used. Of course, we have to use (2.1) for the remaining rows. It is possible to extend theorem 3.1 to include the case when A has disjoint submatrices. Suppose there exist permutation matrices P and Q, such that

$$PAQ = \begin{bmatrix} R^{(1)} & 0 & \ldots & 0 \\ 0 & R^{(2)} & \ldots & 0 \\ \cdot & \cdot & \ldots & \cdot \\ \cdot & \cdot & \ldots & \cdot \\ \cdot & \cdot & \ldots & \cdot \\ 0 & 0 & \ldots & R^{(t)} \\ N^{(1)} & N^{(2)} & \ldots & N^{(t)} \end{bmatrix}, \tag{3.6}$$

where $R^{(i)}$ is $m_i \times n_i$, $N^{(i)}$ is $w \times n_i$: $i = 1, 2, \ldots, t$, and $w = n - \sum_{i=1}^{t} m_i$.

The non-singularity of A implies that $n_i \geqslant m_i$. Let $N = [N^{(1)}, N^{(2)}, \ldots, N^{(t)}]$.

The matrices P and Q can be determined by using techniques similar to those given in (Mayoh, 1965); viz., modified to include rectangular matrices, as the $R^{(i)}$'s in our case can be rectangular. The right-hand side of (3.6) is called a singly bordered block diagonal matrix. The attachment set defined in (Mayoh, 1965) is in the set of rows in N. Having determined P and Q as above, or otherwise, we can write (1.1) as

$$PAQQ^{-1}x = Pb = f \text{(say)}.$$

Since $Q^{-1} = Q^T$ and if we let $y = Q^T x$, then we have

$$PAQy = f. \tag{3.7}$$

Let x_k be an approximate solution of (3.7) and $x_k^{(i)}$ be a column vector consisting of the n_ith through the $(n_{i+1} - 1)$th elements of x_k. Let ϕ_i be the orthogonal projection of x_k on the column space of

$$(0, \ldots, 0, R^{(i)}, 0, \ldots, 0)^T = U_i^T \text{ (say)}.$$

Then $\qquad \phi_i = U_i^T (U_i U_i^T)^{-1} U_i x_k$

$$= U_i^T [R^{(i)} R^{(i)^T}]^{-1} R^{(i)} x_k^{(i)},$$

or $\qquad \phi_i = [0, \ldots, 0, \eta_i^T, \ldots 0]^T, \qquad (3.8)$

where $\qquad \eta_i = R^{(i)^T} [R^{(i)} R^{(i)^T}]^{-1} R^{(i)} x_k^{(i)}.$

Now x_k can be expressed as

$$x_k = \sum_{i=1}^{t} \phi_i + \theta, \text{ where } U_i \theta = R^{(i)} \theta = 0, i = 1, 2, \ldots, t. \tag{3.9}$$

We can extend theorem 3.1 as follows:

Theorem 3.2. If $\eta_i^{(m_i)}$ denotes the value of η_i after m_i applications of (2.1) using the rows of $R^{(i)}$ and $\phi_i^{(m_i)}$ the corresponding value of ϕ_i, then

$$x_{k+n-w} = x_k + \sum_{i=1}^{t} [\phi_i^{(m_i)} - \phi_i]. \tag{3.10}$$

Proof. For $i \neq j$, we see from (3.6) that $U_i U_j^T = 0$; and by construction ϕ_j lies in the column space of U_j^T; therefore it follows that $U_i \phi_j = 0$. Thus in the Kaczmarz method (2.1), when using the rows of $R^{(i)}$ (rows of U_i) only ϕ_i will change, but θ, ϕ_j's and $\phi_j^{(p)}$'s $(i \neq j)$ will remain the same. The reason for $\phi_j^{(p)}$'s remaining the same is as follows. From (2.1) we observe that if $\phi_j^{(p)}$ lies in the column space of U_j^T, then $\phi_j^{(p+1)}$ also lies in the same space, if any row of U_i is used. In view of the above facts and (3.9), we have $x_{k+n-w} = \sum_{i=1}^{t} \phi_i^{(m_i)} + \theta$, which gives (3.10) on substituting the value of θ from (3.9).

It should be noted that if $m_i = n_i$ for some i, then $\eta_i = R^{(i)^T}[R^{(i)}R^{(i)^T}]^{-1}R^{(i)}x_k^{(i)} = x_k^{(i)}$. Of course, in this case, the matrix $R^{(i)}$ can be inverted to give immediately the value of $x_k^{(i)}$ in (3.7), thus decreasing the order of the system by m_i. The use of theorem 3.2 leads to significant savings in storage and computing time, because when using the Kaczmarz method with the row of A belonging to the $R^{(i)}$'s; viz., computing $\phi_i^{(m_i)}$'s we only use small submatrices and their rows. Let us define the disjoint columns of A as the disjoint rows of A^T. If some of the columns of A are disjoint (and therefore necessarily orthogonal), then we can also make use of this fact in the Kaczmarz method as follows. Let Q be a permutation matrix such that

$$AQ = \begin{bmatrix} L_1 & M_1 \\ L_2 & M_2 \end{bmatrix},$$

where L_1 is $m \times m$ and $[L_1^T, L_2^T]\begin{bmatrix} L_1 \\ L_2 \end{bmatrix} = D$, a diagonal matrix. Then from (1.1) we have

$$AQQ^Tx = b \text{ or } AQy = b.$$

Thus

$$\begin{bmatrix} L_1 & M_1 \\ L_2 & M_2 \end{bmatrix} y = b. \tag{3.11}$$

Premultiplying (3.11) by the matrix

$$\begin{bmatrix} L_1^T & L_2^T \\ 0 & I \end{bmatrix}$$

we get

$$\begin{bmatrix} D & V \\ L_2 & M_2 \end{bmatrix} y = \begin{bmatrix} c \\ d \end{bmatrix}, \tag{3.12}$$

where $V = L_1^TM_1 + L_2^TM_2$ and c is $m \times 1$. Let $\begin{bmatrix} L_1 \\ L_2 \end{bmatrix} = L$, and if the ith column of L is denoted by $L^{(i)}$, and the ith diagonal element of D by σ_{ii}, then $\sigma_{ii} = L^{(i)^T}L^{(i)}$. Now, if V_i denoted the ith row of V, c_i and x_{k+1}^i, the ith element of c and x_{k+i} respectively, and $\hat{x}_{k+i}$, the last $n - m$ elements of x_{k+i}; then we have the following theorem:

Theorem 3.3. If the scheme (2.1) is applied to (3.12), with x_k as the initial approximation and using the first m rows, then for $i = 1, 2, \ldots, m$

$$x_{k+i+1} = x_{k+i} + \frac{c_i - \sigma_{ii}x_{k+i}^i - V_i\hat{x}_{k+i}}{\sigma_{ii}^2 + |V_i|^2}\begin{bmatrix} \sigma_{ii} & e_i \\ & V_i^T \end{bmatrix}. \tag{3.13}$$

Proof. The ith row of the coefficient matrix in (3.12) is $[e_i^T\sigma_{ii}, V_i]$. Let x_{k+i} be an approximation to (3.12), using the ith row and x_{k+i} in (2.1), we get

$$\begin{aligned} r_i &= c_i - [e_i^T\sigma_{ii}, V_i]x_{k+i} \\ &= c_i - \sigma_{ii}x_{k+1}^i - V_i\hat{x}_{k+i}, \end{aligned}$$

and $|e_i^T\sigma_{ii}, V_i|^2 = \sigma_{ii}^i + |V_i|^2$.

Substituting the above in (2.1) gives (3.13). It is evident that savings in time and storage will result if (3.13) is used when any one of the first m rows of the matrix is chosen for use in (2.1). Notice that the work involved in getting (3.12) is done only once, while the rows of its coefficient matrix will be used repeatedly.

In this section we have given three ways in which the structure A can be used to decrease the computing time (as well as storage) in the Kaczmarz method. In addition to the above methods, we could use some technique for the acceleration of convergence of the Kaczmarz method itself. Two such heuristic techniques are given in (Dyer, 1967), where it is mentioned that the computational experiments showed significant improvement in convergence. However, it is cautioned that the methods may fail to yield a solution at all. Essentially, the two techniques (which are non-linear relaxations) are:

$$x_{k+1}^* = x_k - \frac{x_k^TA^TA(x_{k+1} - x_k)}{|A(x_{k+1} - x_k)|^2}(x_{k+1} - x_k)$$

and

$$\begin{aligned} x_{k+2}^* &= x_{k+1} \\ &\quad - \frac{(x_{k+1} - x_k)^T(x_{k+1} - x_{k+2})}{(x_{k+2} - 2x_{k+1} + x_k)(x_{k+1} - x_{k+2})}(x_{k+2} - x_{k+1}), \end{aligned}$$

where x_k, x_{k+1} and x_{k+2} are three successive iterates obtained from the Kaczmarz method.

4. Other methods

An interesting formula is derived in (Raytheon, 1966–67) by using a geometrical interpretation. We shall now show that it can also be obtained from (1.5). Any system of linear equations, e.g. (1.1) can be written such that each element of the right-hand side vector is unity, viz., $b_i = 1, i = 1, 2, \ldots, n$. Because, if $b_i \neq 0$ or 1, we can divide the ith row of the system by b_i; on the other hand, if $b_i = 0$, we can add another row to the ith row to make $b_i \neq 0$ and then divide to make it unity. Therefore, there is no loss of generality if, instead of (1.1), we consider the system

$$A_ix = 1, i = 1, 2, \ldots, n. \tag{4.1}$$

The following theorem shows how the formula given in (Raytheon, 1966–67) can be obtained from (1.5):

Theorem 4.1. In (1.5), putting

$$z_k = x_k/|x|^2, z_{k+1} = x_{k+1}/|x|^2 \text{ and } A^Tv_k = z_k - A_i^T,$$

gives $$z_{k+1} = z_k - \frac{z_k^T(z_k - A_i^T)}{|z_k - A_i^T|^2}(z_k - A_i^T). \tag{4.2}$$

Proof. Substituting $x_k = |x|^2z_k$, $x_{k+1} = |x|^2z_{k+1}$ and $A^Tv_k = z_k - A_i^T$ in (1.5) and dividing the resulting equation by $|x|^2$, we have

$$z_{k+1} = z_k + \frac{(z_k^T - A_i)\left(\frac{x}{|x|^2} - z_k\right)}{|z_k - A_i^T|^2}(z_k - A_i^T). \tag{4.3}$$

If $(z_k^T - A_i)\frac{x}{|x|^2} = 0$, then (4.3) will become (4.2); since $z_k^T(z_k - A_i^T) = (z_k^T - A_i)z_k$. We shall now prove that $(z_k^T - A_i)x = 0$. We have

$$z_k = A^Tv_k + A_i^T = A^Tw^{(k)} \text{ (say)}. \tag{4.4}$$

Hence

$$(z_k^T - A_i)x = (w^{(k)^T}A - A_i)x$$

$$= \sum_{i=1}^{n} w_1^{(k)} - 1, \text{ using (4.1).}$$

$$= 0, \text{ if} \sum_{i=1}^{n} w_i^{(k)} = 1.$$

Therefore, if we choose $z_k = A^T w^{(k)}$, where $\sum_{i=1}^{n} w_i^{(k)} = 1$, then (4.3) becomes (4.2). In (Raytheon, 1966–67), it is shown that if z_0 is chosen on the hyperplane G passing through the points $A_1^T, A_2^T, \ldots, A_n^T$, viz.,

$$\sum_{i=1}^{n} w_i^{(0)} = 1, \; w_i^{(0)} \geqslant 0,$$

then for all $z_k, \sum_{i=1}^{n} w_i^{(k)} = 1, w_i^{(k)} \geqslant 0$, which completes the proof of the theorem.

The geometrical interpretation of (4.2) is as follows: From z_k, subtract its projection on the vector $z_k - A_i^T$ to give z_{k+1}. The point $z = \frac{x}{|x|^2}$ is the foot of the perpendicular from the origin on H. Thus z_{k+1} is the point, on the line through z_k and A_i^T, which is closest to z. Thus convergence is assured. For the starting solution z_0 of (4.2), we can take $w_i^{(0)} = \frac{1}{n}$.

Finally, we give a matrix formulation of Cimmino's method (Cimmino, 1938). Its geometrical interpretation is given in (Bodewig, 1959, p. 187). Let (1.1) be normalised such that $|A_i| = 1$, for all i and let x_0 be an approximate solution. Let D be a diagonal matrix with all positive diagonal elements m_i and having a trace equal to 2. Then x_1, the next approximation, is given by

$$x_1 = x_0 + (DA)^T r_0. \qquad (4.5)$$

As is well known, (4.5) will converge if $(DA)^T$ has a norm less than one.

But $$||DA||_1 \leqslant ||D||_1 ||A||_1 \leqslant 2 \max_i m_i . \max_i |A_i|$$

$$\leqslant 2 \max_i m_i, \text{ since } |A_i| = 1.$$

Hence (4.5) will converge if $\max_i m_i < \frac{1}{2}$.

References

Bodewig, E. (1959). *Matrix Calculus*, Amsterdam: North Holland Publishing Co.

Cimmino, G. (1938). Calcolo Approssimato per le Soluzioni di Sistemi di Equazioni Lineari, *Ricerca Sci.* II, Vol. 9, I, pp. 326–333.

Dyer, J. (1965). Acceleration of the Convergence of the Kaczmarz Method and Iterated Homogenous Transformations (Doctoral Thesis, UCLA).

Householder, A. S. (1964). *The Theory of Matrices in Numerical Analysis*, New York: Blaisdell Publishing Co.

Kaczmarz, S. (1937). Angenäherte Auflösung von Systemen linearer Gleichungen, *Bull. Internat. Acad. Polon. Sci.* Cl. A., pp. 335–357.

Mayoh, B. H. (1965). A Graph Technique for Inverting Certain Matrices, *Math. Comp.*, Vol. 19, pp. 644–646.

Raytheon Co. (1966–67). Research on Linear Systems of a Very Large Size, National Aeronautics and Space Administration, Washington, D.C., Report Nos. N67–16086 (1966) and N67–25495.

Tewarson, R. P. (1967a). The Elimination and the Orthogonalisation Methods for the Inversion of Sparse Matrices, (to appear in) *The Proceedings of the Operations Research Around the World Meetings*, New Delhi: Operations Research Society of India.

Tewarson, R. P. (1967b). Solution of a System of Simultaneous Linear Equations with a Sparse Coefficient Matrix by Elimination Methods, *B.I.T.* Vol. 7, pp. 226–239.

Tewarson, R. P. (1967c). The Product Form of Inverses of Sparse Matrices and Graph Theory, *SIAM Rev.*, Vol. 9, No. 1, pp. 91–99.

Tompkins, C. B. (1956). Methods of Steep Descent, in *Modern Mathematics for the Engineer*, E. F. Beckenbach, Ed., New York: McGraw-Hill Book Co.