ON THE SOLUTION OF LARGE, STRUCTURED LINEAR COMPLEMENTARITY PROBLEMS: III

by

Richard W. Cottle
Gene H. Golub
Richard S. Sacher

STAN-CS-74-439
AUGUST 1974

COMPUTER SCIENCE DEPARTMENT
School of Humanities and Sciences
STANFORD UNIVERSITY

ON THE SOLUTION OF LARGE, STRUCTURED LINEAR COMPLEMENTARITY PROBLEMS: III

by

Richard W. Cottle[†]

Gene H. Golub[‡]

and

Richard S. Sacher

TECHNICAL REPORT 74-7

June 1974

DEPARTMENT OF OPERATIONS RESEARCH

Stanford University
Stanford, California

[†]Research and reproduction of this report was partially supported by the Office of Naval Research under contract N-00014-67-A-0112-0011; U.S. Atomic Energy Commission Contract AT(04-3)-326 PA #18.

[‡]Research and reproduction of this report was partially supported by U.S. Atomic Energy Commission Contract AT(04-3)-326 PA #30; and National Science Foundation Grant GJ 35135X, Department of Computer Science, Stanford University.

This report was also issued as Report STAN-CS-74-439 of the Computer Science Department, Stanford University.

SECTION 1

BACKGROUND

1.1. Introduction to the Problem.

In previous papers,[1/,2/] R. W. Cottle and R. S. Sacher have discussed three algorithms for the solution of large-scale linear complementarity problems. For a given matrix $M \in R^{n\times n}$ and a given vector $q \in R^n$, the linear complementarity problem is that of finding a solution z to the system

$$q + Mz \geqq 0$$

$$z \geqq 0$$

$$z^T(q + Mz) = 0 \ .$$

The methods required that M be a tridiagonal, Minkowski matrix. This means $M = (m_{ij})$ satisfies the following conditions:

(i) $m_{ij} \leqq 0$ if $i \neq j$

(ii) $m_{ij} = 0$ if $|i-j| > 1$

(iii) M has positive principal minors.

The three algorithms may be briefly described. Algorithm I is a modification of the principal pivoting method [13]. Algorithm II is a specialization of a method proposed by Chandrasekaran [9] and employs

1/ R. W. Cottle and R. S. Sacher, "On the Solution of Large, Structured Linear Complementarity Problems: I," Technical Report 73-4, Department of Operations Research, Stanford University, 1973.

2/ R. S. Sacher, "On the Solution of Large, Structured Linear Complementarity Problems: II," Technical Report 73-5, Department of Operations Research, Stanford University, 1973.

LU factorizations. The algorithm is "adaptive" in the sense that each iteration exploits the factorization associated with its predecessor. Algorithm III is a modification of the point successive overrelaxation technique.

In this paper, we consider the more general linear complementarity problem in which the matrix is no longer necessarily tridiagonal but may be block tridiagonal. We still assume it to be Minkowski, however. This means we may partition M into submatrices M_{ij} $(i, j = 1,2,\ldots,m)$ such that

(i) M_{ii} is a Minkowski matrix of order $n_i = 1, 2, \ldots, m$,

(ii) $M_{ij} \leq 0$ (elementwise) if $i \neq j$,

(iii) M has positive principal minors,

(iv) $M_{ij} = 0$ if $|i-j| > 1$.

(Matrices satisfying condition (iv) alone are known as block tridiagonal matrices.) We also require that M be positive definite and the diagonal blocks, M_{ii}, be symmetric and tridiagonal. (With this last assumption, we may vastly increase the efficiency of the algorithm we propose in Section 2.4 by incorporating the techniques described in the previously cited paper by R. S. Sacher.) Such matrices include block tridiagonal Stieltjes matrices (see [58, p.85]) whose diagonal blocks are tridiagonal. These occur frequently in the discretization of elliptic partial differential equations. In fact, it is in this connection that an important application of the linear complementarity problem is discussed in Section 3.

The convergence of the algorithm we propose in Section 2.4 requires only that M be positive definite and that the diagonal blocks,

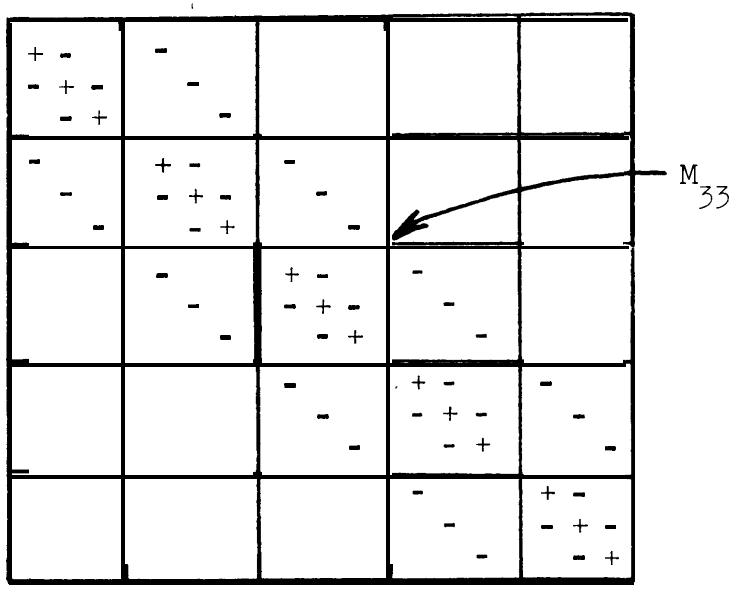
$M_{ii'}$ be symmetric. (That is, in proving convergence, we drop the assumptions of block tridiagonal structure and nonpositive off-diagonal entries.) The method is consequently stated in full generality.

1.2. Drawbacks of the Generalization of Algorithm I.

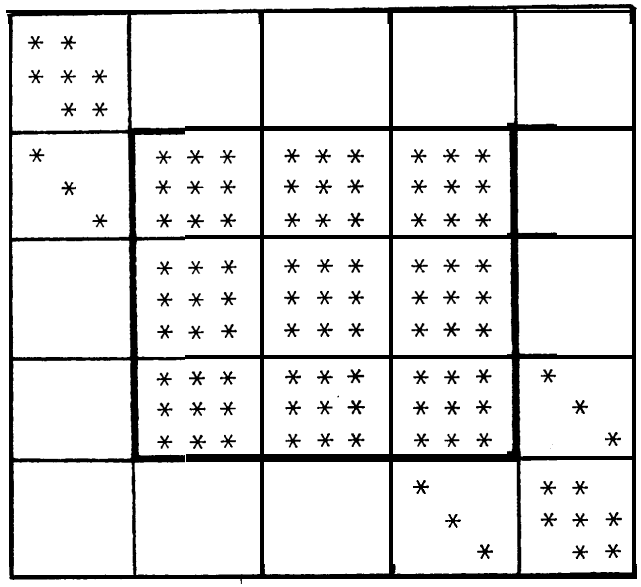
The success of Algorithms I and II for the tridiagonal case suggests that they may be profitably applied to the block tridiagonal case. The purpose of this section is to show why the benefits of those techniques are lost in their extensions.

Consider first the modified principal pivoting algorithm. Certain structural properties of the tableaux under principal pivoting when M is block tridiagonal are analogous to those when M is a tridiagonal matrix. Consequently, an immediate extension of Algorithm I may create a prohibitive number of nonzero entries to be stored as the algorithm progresses. The following two examples illustrate this remark.

Example 1. In Figure 1, the lighter lines indicate the partition of M. We assume that the matrix is block tridiagonal and Minkowski, M_{ii} is tridiagonal of order $n_i = n = 3$ and the off-diagonal blocks are diagonal matrices. The innermost block is M_{33}, the pivot block. The locations of possibly nonzero entries in the pivotal transform $\bar{M}$ are indicated by the asterisk symbol. The main significance of this example is that with the given pattern of zeroes, the principal block pivot on M_{33} may create complete fill-in within the dark border. (For notational convenience, we refer to the entries outside the



M (before pivot)



$\overline{M}$ (after pivot)

Figure 1. Effect of Principal Pivoting on a Block Tridiagonal Minkowski Matrix

pivot block but inside the dark lines as the frame of the pivot, n this case, the frame is of width n.)

Example 2. Figure 2 portrays the effect of pivoting on the blocks $M_{22}, M_{44}, \ldots, M_{m-1,m-1}$ where m is odd. If m equals n, one can easily show that even if (as in Algorithm I) we discard the transformed tableau entries in columns where a pivot has occurred, the number of nonzeroes which must be stored in the transformed tableau is $(\frac{5}{2}(n-1)+1)n^2 = \frac{5}{2}n^3 - \frac{3}{2}n^2$. Compare this with the number of initial nonzero entries in M, i.e., $n(3n-2) + 2(n-1)n = 5n^2 - 4n$. (It is not uncommon [19] for n to equal 100 and thus to have an approximate increase in the number of nonzeroes which must be stored from 50,000 to 2,500,000!)

6.3. Drawbacks of the Generalization of Algorithm II.

Recall that Algorithm II, the modification of Chandrasekaran's method using factorization, requires the solution of a sequence of systems of linear equations by LU decomposition. The order of the final system solved is equal to the cardinality of the set of positive z-variables in the solution to the linear complementarity problem. From [9], we know that if $M \in R^{p \times p}$ is a Minkowski matrix and if $q < 0$, then the solution is the positive vector $z = -M^{-1}q \in R^p$ and we are required to solve a linear system of order p. If we use the example corresponding to Figure 1, we have $p = mn$. We would like to factor M in a way that exploits its structure and sparsity as much as possible. If we were to view it as a band matrix of width n, we would use LU or Cholesky ($\hat{L}\hat{L}^T$) factorization since they both

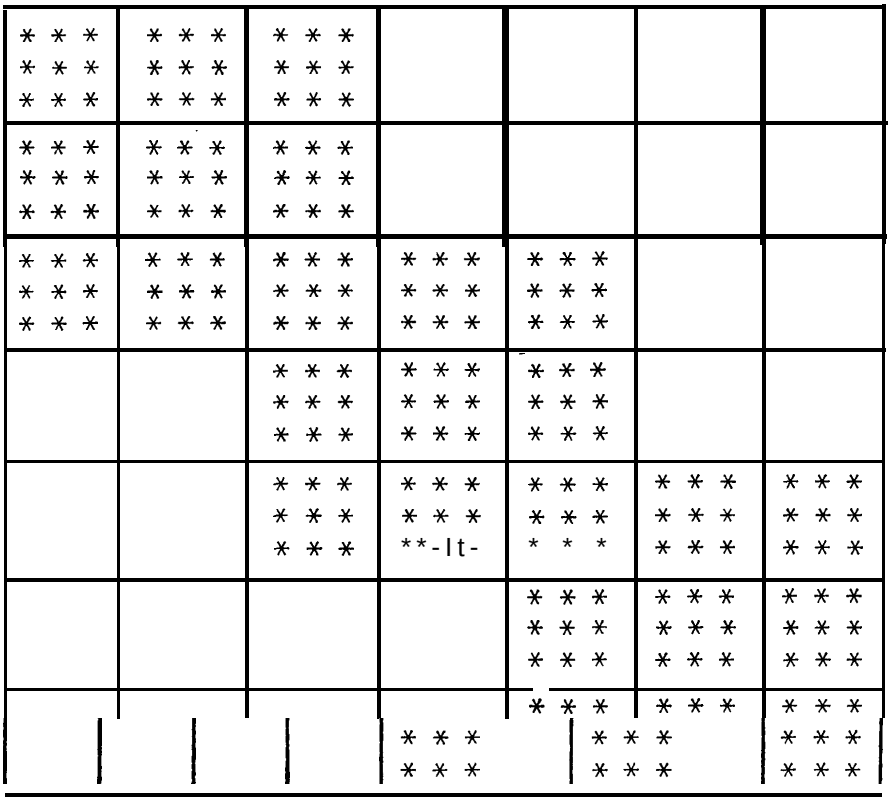


Figure 2. Matrix $\bar{M}$ of Transformed Tableau After Principal Pivots on $M_{22}, M_{44}, \ldots, M_{m-1,m-1}$ ($m = 7$). Original Matrix is Block Tri-Diagonal and Minkowski.

preserve the bandwidth [36]. Unfortunately, neither method of decomposition will preserve the sparseness of the original data. This is illustrated by the example in Figure 3 where the matrix M corresponds to the finite difference equations representing the Laplace equation: M_{11} is tridiagonal with diagonal entries equal to 4 and off-diagonal entries equal to -1; furthermore, $M_{i,i-1}$ and $M_{i,i+1}$ are negative identity matrices. Since the bandwidth of such a matrix cannot be reduced any further, the Cholesky ($\hat{L}\hat{L}^T$) factorization requires storage for almost $\sum_{i=(m-2)n}^{mn} i = 2mn^2 + mn - 2n^2 - n$ nonzero matrix entries. For instance, if $m = n = 100$, then the initial number of nonzeroes in the matrix is approximately $5n^2 = 50{,}000$ while $2mn^2 + mn - 2n^2 - n$ is approximately 2,000,000. The LU factorization needs nearly twice as much storage as the Cholesky factorization since L, U^T and $\hat{L}$ have idential patterns of nonzeroes, i.e., ℓ_{ij}, u_{ji} and $\hat{\ell}_{ij}$ are simultaneously nonzero or zero [36].

A third alternative for factorization is a special case of methods known as group- or block-elimination [36, p. 59]. Isaacson and Keller [36] discuss one technique which is a highly efficient direct method but which requires slightly more storage than the Cholesky decomposition. Following their discussion, we seek a factorization of the form

$$M = L\,U = \begin{vmatrix} A_1 & & & & \\ B_2 & A_2 & & & \\ & B_3 & A_3 & & \\ & & \ddots & \ddots & \\ & & & B_m & A_m \end{vmatrix} \begin{vmatrix} I_1 & C_1 & & & \\ & I_2 & C_2 & & \\ & & \ddots & \ddots & \\ & & & & C_{m-1} \\ & & & & I_m \end{vmatrix}$$

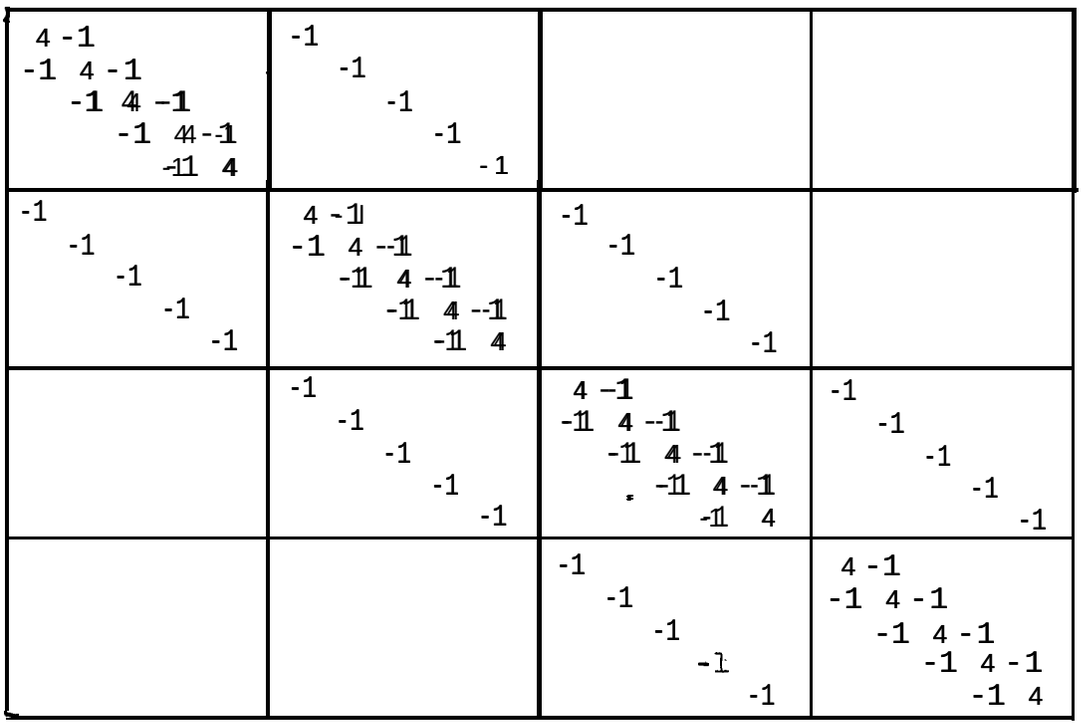


Figure 3 . Example of a "LAP" Matrix with m = 4, n = 5.

where the identity matrices I_j and the matrices A_j, B_j, and C_j are all of order n_j ($j = 1, 2, \ldots, m$). Consequently,

$$A_1 = M_{11} \qquad C_1 = A_1^{-1} M_{12}$$

$$B_i = M_{i,i-1} \qquad i = 2, 3, \ldots, m,$$

$$A_i = M_{ii} - B_i C_{i-1} \qquad i = 2, 3, \ldots, m,$$

and

$$C_i = A_i^{-1} M_i \qquad i = 2, 3, \ldots, m-1.$$

There are $m-1$ matrices A_i and $m-1$ matrices C_i which may each contain n^2 nonzero entries. The B_i matrices need no additional storage. Hence, the block-LU decomposition requires $2mn^2 - n^2$ storage locations versus the $2mn^2 + mn - 2n^2 - n$ required by Cholesky factorization.

In summary, extensions to both principal pivoting methods and various factorization techniques are stymied by storage problems. Similar difficulties in solving large systems of linear equations were recognized by numerical analysts. These difficulties rekindled their interest in iterative (versus direct) methods of solution--that is, in determining techniques to accelerate the convergence of existing methods and in developing new approaches. In Section 2, some results in the former category will provide motivation for the iterative technique we propose for solving the linear complementarity problem (q,M) when M is a block tridiagonal, positive definite Minkowski matrix whose diagonal blocks have symmetric tridiagonal structure.

SECTION 2

ALGORITHMS FOR THE BLOCK TRIDIAGONAL LINEAR COMPLEMENTARITY PROBLEM

2.1. Introduction.

It is ironic that the algorithm we develop in this section arises as a generalization of the least computationally attractive method of Algorithm I, II and III. Yet the computational experience reported in Section 4 demonstrates that this generalization is at least competitive with, if not superior to, techniques currently available [7], [19], [29], [45], [46], [47], [59] for solving the engineering application described in Section 3.

Because of the analogy of the proposed algorithm with relaxation techniques for systems of linear equations, we open this section with some remarks about these methods. For the remainder of the paper, we will observe the following notational conventions. All vectors are column vectors. By a slight abuse of notation, we let $z = (z_1, z_2, \ldots, z_n)$ denote the column vector z in R^n. Similarly, if $z_i \in R^{n_i}$ and $\sum_{i=1}^{m} n_i = N$, then we may let $z = (z_1, z_2, \ldots, z_m)$ denote the column vector z in R^N. Finally, the algorithms to be described will generate a sequence of iterates z^k, $k = 1, 2, \ldots$, converging to a solution. The value of z^k is determined by a specified transformation on z^{k-1}. Therefore the sequence z^k, $k = 1, 2, \ldots$, is totally determined by an initial vector z^0. We denote the sequence z^k, $k = 1, 2, \ldots$, by $\{z^k\}$ and suppress its implicit dependence on z^0.

2.2. Point Successive Overrelaxation (SOR) Algorithm for Linear Systems.

The point successive overrelaxation algorithm for solving the linear system $Mz + q = 0$, where M is an $m \times m$ matrix, is an accelerated version of the earlier Gauss-Seidelmethod [58]. This latter method generates a sequence of iterates $z^k \in R^m$ according to the formula:

$$z_i^{k+1} = -\left(\sum_{j<i} m_{ij} z_j^{k+1} + \sum_{j>i} m_{ij} z_j^k + q_j \right)/m_{ii}$$

$$-i = 1, 2, \ldots, m.$$

Each component of z^{k+1} is recursively determined in terms of the current values of the others. The recursion formula may be rewritten in the following format in which $\omega = 1$: Determine z_i^{k+1} $(i = 1, 2, \ldots, m)$ by

$$z_i^{k+1} = z_i^k + \omega(\bar{z}_i^{k+1} - z_i^k) \qquad (1)$$

where $m_{ii}\bar{z}_i^{k+1} + \left(\sum_{j<i} m_{ij} z_j^{k+1} + \sum_{j>i} m_{ij} z_j^k + q_j \right) \ldots$

We interpret the term $(\bar{z}_i^{k+1} - z_i^k)$ as a direction in which to proceed from the current value of $z_i = z_i^k$. The parameter ω is thus thought of as a weighting factor to indicate how far to move in this direction.

We have noted that in the Gauss-Seidel algorithm $\omega = 1$. In 1950, young [61] and Frankel [25] simultaneously, but independently, recognized the efficacy of using values of ω different from unity to

gain faster convergence rates. The scalar ω is called the relaxation parameter, and $\omega > 1$ ($\omega < 1$) corresponds to overrelaxation (underrelaxation). The method of Young and Frankel (using $\omega > 1$) is called the point successive overrelaxation algorithm.

The word "point" in the name of the method has an interesting geometric origin. Suppose we are solving Laplace's equation, $\nabla^2 u = f$, over a rectangular region by a finite difference method. This first requires forming a grid over the region. We then seek an approximation to the unknown function u at the grid points only. This is achieved by assigning a variable z_i to each grid point (see Figure 4) and obtaining, by well-known techniques (see [24, p. 192]), a linear system $Mz + q = 0$. We define the error at the i^{th} grid point to be the absolute value of the difference between z_i and the function u evaluated at that grid point. If the grid is square and has n points on a side, then the maximum of these errors is $O(1/n^2)$. When the point SOR method is applied to the linear system, the algorithm changes the value of only one variable z_i at a time, i.e., only one grid point is examined at a time. Hence the word "point" in the algorithm name.

2.3. Block Successive Overrelaxation (SOR) Algorithm for Linear Systems.

In certain situations, it is natural to consider simultaneously changing the values of all variables associated with a coordinate line of the grid points. Such methods are known as line- or block-iterative techniques. The word 'block" refers to the fact that the variables

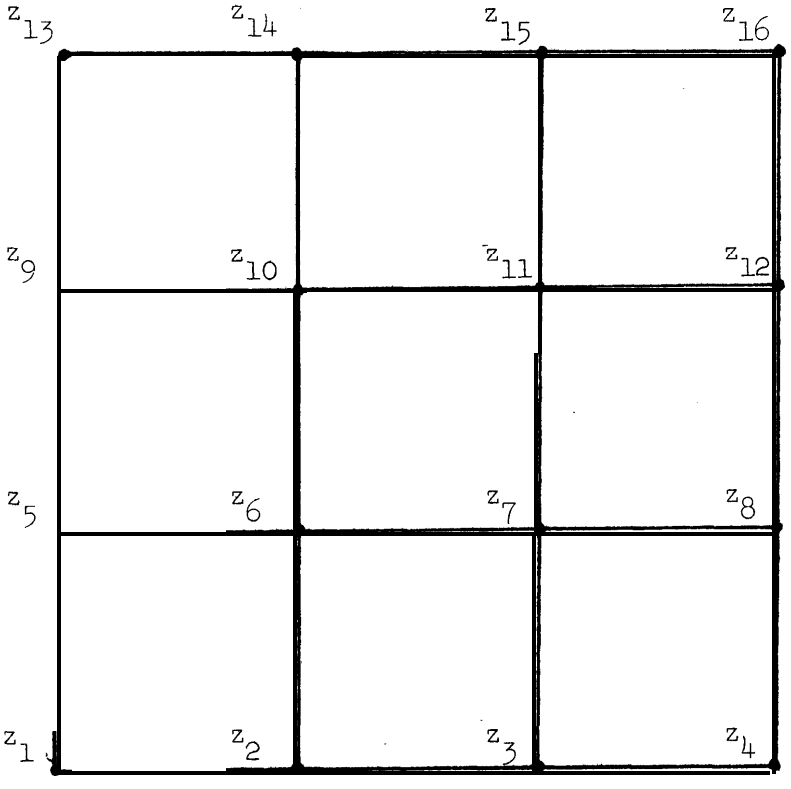


Figure 4. Grid for Finite Difference Equations

which are simultaneously changed correspond to a diagonal block (or principal submatrix) of the matrix M. Varga[58, p. 96] indicates that block methods are not new developments but may be traced back to the work of Gerling [28] in 1843.

We again use the example of the finite difference discretization of a differential equation over a rectangle. In the corresponding linear system, $Mz + q = 0$, we conformably partition the vectors z and q and the matrix M. We will view z as a direct sum of vectors $z = (z_1, z_2, \ldots, z_m)$ where $z_i \in R^{n_1}$. Thus $z = (z_{11}, z_{12}, \ldots, z_{1n_1}; z_{21}, z_{22}, \ldots, z_{2n_2}; \ldots z_{m,n_m})$. A corresponding relabeling of the grid in Figure 4 is shown in Figure 5. Finally, $q = (q_1, q_2, \ldots, q_m)$ is similarly relabled and partitioned.

The corresponding changes in the recursion equation (1) may now be stated. Recursively determine the subvectors z_i^{k+1} $(i = 1, 2, \ldots, m)$ by

$$z_i^{k+1} = z_1^k + \omega(\bar{z}_i^{k+1} - z_i^k) \qquad (2)$$

where $$M_{ii}\bar{z}_i^{k+1} + \left(\sum_{j<i} M_{ij}z_j^{k+1} + \sum_{j>i} M_{ij}z_j^k + q_i\right) = 0 .$$

Varga [58, p. 91] remarks that in the numerical solution of many physical problems, the matrix M is endowed with properties which guarantee that block SOR will converge to a solution faster than will point SOR. In these applications, the matrix M is irreducible and Stieltjes (i.e., symmetric Minkowski). Arms, Gates and Zondek [1] state that if M is merely a Minkowski matrix, then block SOR still has the advantage. For a more extensive treatment of successive

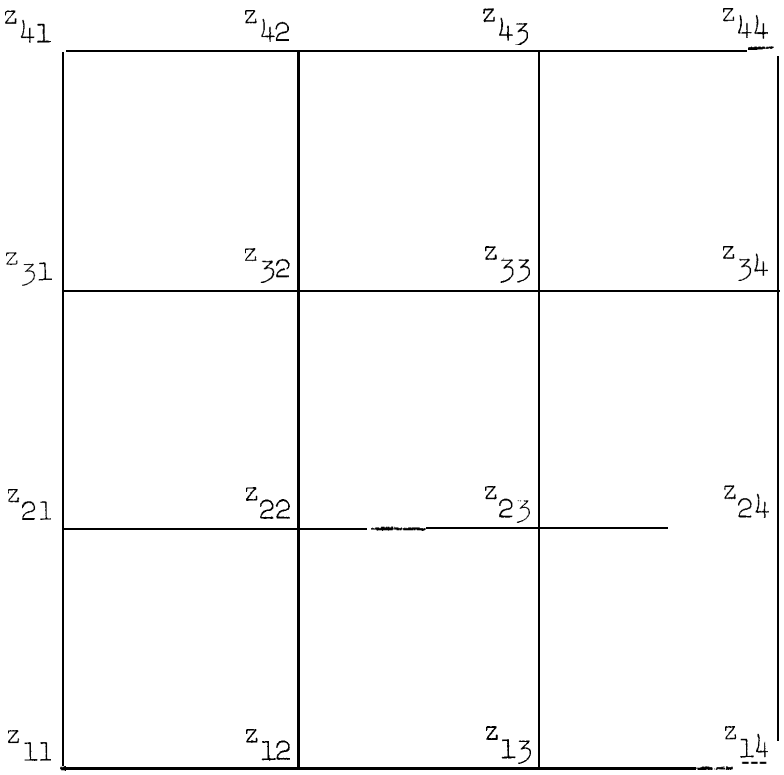


Figure 5. The Grid Variables Relabeled

overrelaxation techniques and their many variants, see [36],[58], [59], [60].

2.4. Algorithm IV: Modified Block Successive Overrelaxation Algorithm for Linear Complementarity Problems.

The philosophy of generalizing point SOR to block SOR in order to solve linear systems finds an analogue in generalizing Algorithm III, the modified point SOR technique for linear complementarity problems. In this section, we show that by restating Algorithm III, a certain generalization suggests itself. In the following sections, an analysis of the latter algorithm will illustrate three points. First, under reasonable hypotheses, the sequence of iterates generated by the algorithm will form a monotonically increasing or decreasing sequence of vectors converging to the solution of the problem. (Under these hypotheses, the results in [52] guarantee that a unique solution exists.) Second, the method may be interpreted as a manifold sub-optimization technique applied to a related quadratic programming problem. Third, values of ω greater than unity may be used to accelerate the convergence of the algorithm.

Recall Cryer's [20] description of the modified point SOR algorithm for the linear complementarity problem (q,M) where $M \in R^{m \times m}$ is positive definite. The parameter $\epsilon > 0$ is chosen small enough to-insure that the errors in the values of the z-variables are sufficiently small. We shall make a slight modification in notation.

Algorithm III (Modified Point SOR)

Step 0. Let $z^0 = (z_1^0, z_2^0, \ldots, z_m^0)$ be an arbitrary nonnegative m-vector and $\omega \in (0,2)$. Set $k = 0$.

Step 1. Let $\hat{z}_i^{k+1} = -\left(\sum_{j<i} m_{ij} z_j^{k+1} + \sum_{j>i} m_{ij} z_j^k + q_i\right)/m_{ii}$

$i = 1, 2, \ldots, m.$

Let $z_i^{k+1} = \max[0, z_i^k + \omega(\hat{z}_i^{k+1} - z_i^k)]$

Step 2. Define $J = \{i: z_i^{k+1} > 0\} \cup \{i: z_i^{k+1} = 0, (Mz^{k+1} + q)_i < 0\}$. If $\max_{i \in J} |(Mz^{k+1} + q)_i| < \epsilon$ stop. An "approximate" solution is at hand. Otherwise, go to Step 1 with k replaced by k+1.

The algorithm is essentially the point SOR algorithm for linear systems with the precaution that if a z-variable ever becomes negative, it is immediately set equal to zero. Cryer [19] gives a convergence proof for Algorithm III under the assumption that M is symmetric and positive definite. Historically, an identical modification of the Gauss-Seidel method has appeared in several varied contexts, see [5], [26], [34].

Algorithm III may be viewed in a slightly different but equivalent way for values of $\omega \geq 1$.

Proposition 1. If $\omega \geq 1$, then Step 1 is equivalent to the following:

step 1': (a) Let $\bar{z}_i^{k+1}$ solve the linear complementarity problem

$$(q', M') \equiv \left(\sum_{j<i} m_{ij} z_j^{k+1} + \sum_{j>i} m_{ij} z_j^k + q_i,\ m_{ii}\right).$$

(b) Let $\omega_i^{k+1} = \max\{\bar{\omega}: \bar{\omega} \leq \omega,\ z_i^k + \bar{\omega}(\bar{z}_i^{k+1} - z_i^k) \geq 0\}$.

(c) Let $z_i^{k+1} = z_i^k + \omega(\bar{z}_i^{k+1} - z_i^k)$; $i = 1, 2, \ldots, m$.

Proof. The analysis is divided into two cases:

Case 1. $\sum_{j<i} m_{ij} z_j^{k+1} + \sum_{j>i} m_{ij} z_j^k + q_i < 0$. Consequently, $\bar{z}_i^{k+1} = \hat{z}_i^{k+1}$. From (b) ω_i^{k+1} equals ω if and only if $z_i^k + \omega(\bar{z}_i^{k+1} - z_i^k) \geq 0$. In this case Steps 1 and 1' give the same value for z_i^{k+1}. Suppose, on the other hand, that $z_i^k + \omega(\bar{z}_i^{k+1} - z_i^k) < 0$, and therefore Step 1 sets z_i^{k+1} equal to zero. But $\max\{0, z_i^k + \omega(\bar{z}_i^{k+1} - z_i^k)\} = 0$ if and only if there is an $\omega' \leq \omega$ for which $z_i^k + \bar{\omega}(\bar{z}_i^{k+1} - z_i^k) \geq 0$ (resp. < 0) when $\bar{\omega} \leq \omega'$ (resp. $\bar{\omega} > \omega'$). In this situation, ω_i^{k+1} is chosen to be ω' in (b) and thus $z_i^{k+1} = 0$ in (c) of Step 1'.

Case 2.

$$\sum_{j<i} m_{ij} z_j^{k+1} + \sum_{j>i} m_{ij} z_j^k + q_i \geq 0 . \qquad (3)$$

Therefore, in Step 1 we have $\hat{z}_i^{k+1} - z_i^k \leq -z_i^k$, $z_i^k + \omega(\hat{z}_i^{k+1} - z_i^k) \leq 0$ for all $\omega \geq 1$ and z_i^{k+1} is set equal to zero. In Step 1', equation (3) implies that $\bar{z}_i^{k+1} = 0$ and that $\omega(\bar{z}_i^{k+1} - z_i^k) = -z_i^k$. Consequently, $\omega \geq 1$ implies that the value of ω_i^{k+1} chosen in (b) of Step 1' is unity and thus z_i^{k+1} is set equal to zero in (c). ∎

In Section 2.5, we show that Algorithm III with the Step 1' substitution will converge for all $\omega \in (0,2)$. Under this new interpretation regarding the choice of the relaxation parameter at

each iteration, a generalization of the preceding algorithm may be proposed. By a slight change in notation, we shall pass from a point-iterative to a block-iterative technique. We will use the notation described in Section 2.3 on block SOR for linear systems. In particular, M is partitioned into submatrices M_{ij} $(i, j = 1, 2, \ldots, m)$ where M_{ii} is of order n_i, $z = (z_1, z_2, \ldots, z_m)$ and $q=(z_1, q_2, \ldots, q_m)$ where z_1 and q_i are n_i-vectors. Algorithm III, with the substitution of Step 1', then forms the basis for the following generalization. We refer to the new algorithm as the Modified Block SOR Algorithm for the linear complementarity problem (q,M) where M is positive definite and M_{ii} $(i = 1, 2, \ldots, m)$ is symmetric.

Algorithm IV (Modified Block SOR)

Step 0. Let $z^0 = (z_1^0, z_2^0, \ldots, z_m^0)$ be an arbitrary nonnegative vector and $\omega \in (0,2)$. Set k = 0 and i = 1

Step 1. Let $\bar{z}_i^{k+1}$ solve the linear complementarity problem

$$\left(\sum_{j<i} M_{ij} z_j^{k+1} + \sum_{j \geq i} M_{ij} z_j^k + q_i, M_{ii} \right)$$

Step 2. Let $\bar{\omega}_i^{k+1} = \max\{\bar{\omega} : \bar{\omega} \leq \omega, \; z_i^k + \bar{\omega}(\bar{z}_i^{k+1} - z_i^k) \geq 0\}$.

Let $z_i^{k+1} = z_i^k + \bar{\omega}_i^{k+1}(\bar{z}_i^{k+1} - z_i^k)$.

Step 3. If i = m, go to Step 4. Otherwise, return to Step 1 with i replaced by i+1.

Step 4. Define $J = \{(i,j):(z_i^{k+1})_j > 0 \text{ or } (\sum_{r=1}^{m} M_{ir}z_r^{k+1} + q_i)_j < 0\}$.

If $\max_{(i,j) \in J} |(\sum_{r=1}^{m} M_{ir}z_r^{k+1} + q_i)_j| < \epsilon$, stop. An approximate solution is at hand. Otherwise, return to Step 1 with k replaced by k+1 and i = 1.

The differencesbetween this algorithm and block SOR for linear systems are evident. In Step 1, we solve a linear complementarity problem $(\bar{q},\bar{M})$ instead of solving a linear system $\bar{M}\bar{z} + \bar{q} = 0$. Also, the nonnegativity constraint of the complementarity problem is handled in Step 2 by requiring that movement in the direction $-(z_i^{k+1} - z_i^k)$ be constrained to remain in the nonnegative orthant. In solving linear systems, the nonnegativity restriction is absent and thus ω_i^{k+1} always equals ω.

The computational bottleneck to the modified block SOR algorithm, if one exists, will occur in Step 1 where linear complementarity problems must be repeatedly solved. In general, if M_{ii} is an arbitrary positive definite matrix, then the standard methods for solving $(\bar{q}_i,\bar{M}_{ii})$, (e.g., the principal pivoting technique of Cottle [13] or Lemke's method [38] may be used--possibly at the expense of large core storage requiremehts and perhaps not particularly rapid convergence.) However, if we apply the modified block SOR algorithm to matrices whose diagonal blocks M_{ii} are tridiagonal Stieltjes matrices, then Algorithms I and II may be profitably applied to yield an algorithm of high overall efficiency. An example of this type will be discussed in Section 3.

2.5. Convergence of Algorithm IV .

It is natural to look at the corresponding problem of the minimization of a quadratic function over the nonnegative orthant to help demonstrate the convergence of the algorithm. We shall use a method of proof similar to that of Cryer [20] and Schechter [53],[54]. If the matrix M is positive definite, then the Kuhn-Tucker conditions for the problem

$$\text{minimize} \quad f(z) = \frac{1}{2} z^T M z + qz$$
$$\text{subject to} \quad z \geq 0$$

are the necessary and sufficient conditions which a global minumum satisfies. If we further assume that M is symmetric, then the Kuhn-Tucker conditions are equivalent to the linear complementarity problem (q,M).

The first result will show that the successive iterates $\{z^k\}$ cause the sequence $\{f(z^k)\}$ to be strictly monotonically decreasing.

Theorem 1. Let

$$f(x,y) = \frac{1}{2} \begin{pmatrix} x \\ y \end{pmatrix}^T \begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} p \\ s \end{pmatrix}^T \begin{pmatrix} x \\ y \end{pmatrix}$$

where $p, x \in R^n$, $s, y \in R^{N-n}$, and x and y are arbitrary vectors. Assume-that A is symmetric and positive definite. Let $\bar{x}$ solve the linear complementarity problem $(Hy + p, A)$ where $H = (B+C^T)/2$. Then $f(x + \omega(\bar{x}-x),y) \leq f(x,y)$ for $\forall\, \omega \in (0,2)$. Furthermore, quality holds if and only if $\bar{x} = x$.

Proof. We examine the minimization of $g(u) \equiv f(u,y)$ over the non-negative orthant. Rearranging terms, $g(u) = \frac{1}{2} u^T A u + (Hy + p)^T u + c$ where c is a constant. Since A is symmetric and positive definite, the minimizing vector $\bar{x}$ is the solution to the related linear complementarity problem $(Hy + p, A)$. For notational convenience, let $r = H\bar{x} + p$.

We will assume that $\bar{x} \neq x$. By a principal rearrangement of A, we may further assume that $\bar{x} = (\bar{x}_K, 0)$, where K is the index set $(1, 2, \ldots, k)$ and $\bar{x}_K = (\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k) > 0$. Letting $\bar{K}$ be the index set $(k+1, k+2, \ldots, n)$, the corresponding partitionings are

$$r = (r_K, r_{\bar{K}}) \quad \text{and} \quad A = \begin{pmatrix} A_{KK} & A_{K\bar{K}} \\ A_{\bar{K}K} & A_{\bar{K}\bar{K}} \end{pmatrix}.$$

Let $d = x - x = (\bar{x}_1 - x_1, \bar{x}_2 - x_2, \ldots, \bar{x}_k - x_k \; x_{k+1}, -x_{k+2}, \ldots, -(d_K, -x_{\bar{K}})$. We want to show that $g(x) > g(v)$ for all v in the open line segment $V = (x, x + 2d)$. Noting that V may be rewritten as $\{v : v = \bar{x} + \lambda d, \forall \lambda \in (-1,1)\}$, we consider two cases $\lambda \leq 0$ and $A > 0$.

Case 1. $A \in (-1,0]$. Since g is strictly convex, then for all $A \in (-1,0]$,

$$\begin{aligned} g(\bar{x} + \lambda d) &< (1 - |\lambda|)\, g(\bar{x}) + |\lambda|\, g(\bar{x} - d) \\ &= (1 - |\lambda|)\, g(\bar{x}) + |\lambda|\, g(x) \\ &\leq (1 - |\lambda|)\, g(x) + |\lambda|\, g(x) . \end{aligned}$$

Thus,

$$g(\bar{x} + Ad) < g(x) \ . \qquad (4)$$

Case 2. $\lambda \in (0,1)$. Consider the Taylor series expansion $g(\bar{x} + Ad) = g(\bar{x}) + \lambda d^T(A\bar{x} + r) + \frac{1}{2}\lambda^2 d^T A d$. By showing that $d^T(A\bar{x} + r) \leq 0$, we may conclude that $g(\bar{x} + \lambda d) < g(\bar{x} - hd)$ for all $\lambda \in (0,1)$. But equation (4) shows $g(\bar{x} - \lambda d) < g(x)$ for all $A \in (0,1)$; thus $g(\bar{x} + hd) < g(x)$ for all $A \in (0,1)$ also.

Using the index set K and the corresponding partitionings described above, we have

$$d^T(A\bar{x} + r) = \begin{pmatrix} d_K \\ -x_{\bar{K}} \end{pmatrix}^T \left[\begin{pmatrix} A_{KK} & A_{K\bar{K}} \\ A_{\bar{K}K} & A_{\bar{K}\bar{K}} \end{pmatrix} \begin{pmatrix} -A_{KK}^{-1} & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} r_K \\ r_{\bar{K}} \end{pmatrix} + \begin{pmatrix} r_K \\ r_{\bar{K}} \end{pmatrix} \right]$$

$$\begin{pmatrix} d_K \\ -x_{\bar{K}} \end{pmatrix}^T \left[\begin{pmatrix} -I & 0 \\ -A_{\bar{K}K} A_{KK}^{-1} & 0 \end{pmatrix} \begin{pmatrix} r_K \\ r_{\bar{K}} \end{pmatrix} + \begin{pmatrix} r_K \\ r_{\bar{K}} \end{pmatrix} \right]$$

$$= -x_{\bar{K}}^T (r_{\bar{K}} - A_{\bar{K}K} A_{KK}^{-1} r_K) \ .$$

Now recall that $\bar{x} = (\bar{x}_K, 0)$ satisfies the system

$$\begin{pmatrix} A_{KK} & A_{K\bar{K}} \\ A_{\bar{K}K} & A_{\bar{K}\bar{K}} \end{pmatrix} \begin{pmatrix} \bar{x}_K \\ 0 \end{pmatrix} + \begin{pmatrix} r_K \\ r_{\bar{K}} \end{pmatrix} \geq 0, \quad \bar{x}^T(A\bar{x} + r) = 0 \ .$$

Therefore $A_{KK}\bar{x}_{KK} + r_K = 0$ and $A_{\bar{K}K}\bar{x}_K + r_{\bar{K}} \geq 0$. Substituting $\bar{x}_K = -A_{KK}^{-1}r_K$ in the latter of these two systems gives $r_{\bar{K}} - A_{\bar{K}K}A_{KK}^{-1}r_K \geq 0$. Finally, we conclude the argument by noting that $x \geq 0$ implies $d^T(A\bar{x} + r) \leq 0$. ∎

Theorem 1 means that we can use the function f to monitor the progress of the algorithm. If we can guarantee that f is bounded from below on the nonnegative orthant, then we will be assured that the sequence of successive iterates $\{z^k\}$ contain a convergent subsequence. Positive definiteness of M is one sufficient condition for the boundedness of f. A necessary condition is that M be copositive [31]. For f is unbounded on the nonnegative orthant if there is a nonnegative vector x for which x^TMx is negative. Hence f is bounded below on the nonnegative orthant only if x^TMx is nonnegative for every nonnegative vector x.

Each iteration of the algorithm updates the m subvectors of the vector $z^k = (z_1^k, z_2^k, \ldots, z_m^k)$. For future notational convenience, let $f:(v) = f(z_1^k, z_2^k, \ldots, z_{i-1}^k, v, z_{i+1}^{k-1}, z_{i+2}^{k-1}, \ldots, z_m^{k-1})$. Theorem 1 thus shows that $f_i^k(z_i^k) \leq f_i^k(z_i^{k-1})$ with equality if and only if $z_i^k = z_i^{k-1}$. Consequently- $f(z^k) \leq f(z^{k-1})$ with equality if and only if $z^k = z^{k-1}$. In the case of equality, we can prove that z^k solves the linear complementarity problem.

Proposition 2. If the algorithm generates iterates z^j, $j = 1,2,\ldots,k$ and $z^{k-1} z^k$, then z^k solves (q,M).

Proof. From Step 2 of the algorithm, we see that $z^k_i = z^{k-1}_i$ if and only if $\bar{z}^k_i = z^{k-1}_i$. Suppose we are solving for $\bar{z}^k_i$ in Step 1. We require

$$w_i \equiv \sum_{j<i} M_{ij} z^k_j + M_{ii}\bar{z}^k_i + \sum_{j>i} M_{ij} z^{k-1}_j + q_i \geq 0 \text{ and } \bar{w}^T_i \bar{z}^k_i = 0 .$$

Since $\bar{z}^k_i = z^{k-1}_i$ and $z^{k-1} = z^k$, then

$$w^k_i \equiv \sum_{j<i} M_{ij} z^k_j + M_{ii} z^k_i + \sum_{j>i} M_{ij} z^k_j + q_i = w_i \geq 0 \text{ and } (w^k_i)^T z^k_i = 0 .$$

This holds for all i = 1, 2, . . . , m; so z^k solves (q,M). ∎

Finally, we prove that the sequence of iterates have a unique limit z and that the limit solves (q,M).

Proposition 3. If M is positive definite, then the sequence of iterates $\{z^k\}$ are contained in a compact set and hence contain a convergent subsequence with limit z.

Proof. (Similar to Cryer). From Theorem 1, $\{f(z^k)\}$ is a monotonically decreasing sequence. Since M is positive definite and f is a quadratic function, f is bounded from below and thus there is some value to which $f(z^k)$ converges as $k \to \infty$.

The set $S = \{x \mid f(x) \leq f(z^0), x \geq 0\}$ is compact. It is closed because f is continuous. Furthermore, S is bounded. Suppose $v_i \in S$, i = 1, 2, . . . , and $\|v_i\| \to \infty$ where $\|\cdot\|$ is the Euclidean

norm. In the quadratic function $f(v) = \frac{1}{2} v^T M v + qv$, we may assume that M is symmetric and has real eigenvalues. Let $\lambda_1 > 0$ be the smallest eigenvalue. Then, by the Fisher Minimax Theorem [4, p. 72] $v_i^T M v_i \geq \lambda_1 v_i^T v_i = A_1 \|v_i\|^2$; thus $v_i^T M v_i \longrightarrow \infty$ as $i \longrightarrow \infty$. Since the quadratic term $\frac{1}{2} v_i^T M v_i$ dominates the linear term $q^T v_i$ as $\|v_i\|$ becomes large, we conclude that $f(v_i) \longrightarrow \infty$ as $i \longrightarrow \infty$. But this contradicts the assumption that $f(v_i) \leq f(z^0) < \infty$ for all i.

Therefore the iterates $\{z^k\}$ are elements of a compact set and have a convergent subsequence with some limit point $z \in S$. ∎

Proposition 4. Using the notation developed in the algorithm and assuming that M is positive definite, then $\lim_{k \to \infty} (\bar{z}_i^k - \bar{z}_i^{k-1}) = 0$ and $\lim_{k \to \infty} (z^k - z^{k-1}) = 0$.

Proof. From the proof of Theorem 1, for any k and each i,

$$f_i^k(\bar{z}_i^k - d_i^k) = f_i^k(\bar{z}_i^k) - d_i^k)^T (M_{ii} \bar{z}_i^k + q_i) + \frac{1}{2} (d_i^k)^T M_{ii} d_i^k$$

and

$$-f_i^k(\bar{z}_i^k + \lambda d_i^k) = -f_i^k(\bar{z}_i^k) - \lambda(d_i^k)^T (M_{ii} \bar{z}_i^k + q_i) + \frac{1}{2} \lambda^2 (d_i^k)^T M_{ii} d_i^k$$

where $d_i^k = \bar{z}_i^k - z_i^{k-1}$. Adding these equations, we have

$$\begin{aligned} f_i^k(z_i^{k-1}) - f_i^k(\bar{z}_i^k + \lambda d_i^k) &= -(\lambda+1)(d_i^k)^T (M_{ii} \bar{z}_i^k + q_i) + \frac{1}{2} (1-\lambda^2)(d_i^k)^T M_{ii} d_i^k \\ &\geq (1-\lambda^2)(d_i^k)^T M_{11} d_1^k \\ &\geq 0 \qquad \text{since } \lambda \in (-1,1) . \end{aligned}$$

Recall that ω_i^k is the scalar chosen in Step 2 of iteration k and thus A has the value $\lambda_i^k \equiv \omega_i^k - 1$. These values satisfy the following bounds:

$$-1 < \min(0, \omega-1) \leq \lambda_i^k \leq \omega - 1 < 1 ,$$

for all k and all i. Therefore there is a scalar α independent of k and i for which $1 - (\lambda_i^k)^2 \geq \alpha > 0$.

Since the $\{f(z^k)\}$ converge, the $\{f_i^k(z_i^k)\}$ also converge. Therefore

$$\lim_{k \to \infty} (f_i^k(z_i^k) - f_i^k(z_i^k + \lambda_i^k d_i^k)) = \lim_{k \to \infty} (f_i^k(z_i^k) - f_i^k(z_i^{k+1})) = 0$$

and also

$$\lim_{k \to \infty} (d_i^k)^T M_{ii} d_i^k = 0 .$$

Since M_{ii} is positive definite, then $\lim_{k \to \infty} d_i^k = 0$, i.e., $\lim_{k \to \infty} (z_i^k - z_i^{k-1}) = 0$. Finally, $\lim_{k \to \infty} (z_i^k - z_i^{k-1}) = \lim_{k \to \infty} (1-\lambda_i^k) d_i^k = 0$.

■

Proposition 5. Let I be the index set of a convergent subsequence of the iterates $\{z^k\}$ generated by the algorithm. Assume the subsequence converges to the vector z. Then $Mz + q \geq 0$.

Proof. If the inequality does not hold, then there are integers i, j and N and some $\delta > 0$ for which $k > N$ and $k \in I$ implies

$$\left(\sum_{t<i} M_{it}z_t^k + M_{ii}z_i^k + \sum_{t>i} M_{it}z_t^k + q_i\right)_j \leq -6$$

However

$$\left(\sum_{t<i} M_{it}z_t^{k+1} + M_{ii}\bar{z}_i^{k+1} + \sum_{t>i} M_{it}z_t^j + q_i\right)_j \geq 0 .$$

Subtracting the second inequality from the first gives

$$\left(\sum_{t<i} M_{it}(z_t^k - z_t^{k+1}) + M_{ii}(z_i^k - \bar{z}_i^{k+1})\right)_j \leq -\delta .$$

Since the terms in parentheses become arbitrarily close to zero, we have a contradiction. ∎

Proposition 6. Let I be the index set for the convergent subsequence above. Then $z^T(Mz + q) = 0$.

Proof. Suppose the contrary. Then there are integers i, j and N and some $\delta > 0$ for which $k > N$ and $k \in I$ implies that $(z_i^k)_j > \delta$ and

$$\left(\sum_{t<i} M_{it}z_t^k + M_{ii}z_i^k + \sum_{t>i} M_{it}z_t^k + q_i\right)_j > \delta . \qquad (5)$$

However

$$(\bar{z}_i^k)_j \left(\sum_{t<i} M_{it}z_t^k + M_{ii}\bar{z}_i^k + \sum_{t>i} M_{it}z_t^{k-1}\right)_j = 0 . \qquad (6)$$

Suppose $(\bar{z}_i^k)_j > 0$. Then dividing equation (6) by $(\bar{z}_i^k)_j$ and subtracting the result from equation (5) gives

$$(M_{ii}(z_i^k - \bar{z}_i^k) + \sum_{t>i} M_{it}(z_t^k - z_t^{k-1}))_j > \delta.$$

As before, Proposition 5 leads to a contradiction. Alternatively, suppose $(\bar{z}_i^k)_j = 0$. Pick $N' > N$ sufficiently large to guarantee that $|(\bar{z}_i^k - z_i^k)| < \frac{\delta}{2}$ when $k > N'$. If $(\bar{z}_i^k)_j = 0$ for all $k > N'$ when $k \in I$, then $|(\bar{z}_i^k - z_i^k)_j| = (z_i^k)_j > \frac{\delta}{2}$, a contradiction. If there is some $k \in I$ which is greater than N' for which $(\bar{z}_i^k)_j > 0$, then the analysis in the preceding paragraph applies and a contradiction follows. ∎

In summary, these results show that the algorithm generates a sequence of vectors $\{z^k\}$ belonging to a compact set S. Given any convergent subsequence of $\{z^k\}$, its limit point z solves the linear complementarity problem (q,M).

Theorem 2. The entire sequence $\{z^k\}$ has a unique limit point z, and z solves (q,M).

Proof. Since M has positive principal minors, the linear complementarity problem (q,M) has a unique solution (see [44], [52]). Propositions 5 and 6 show that the limit of any convergent subsequence of $\{z^k\}$ solves (q,M). Consequently, every convergent subsequence has a-common (and hence unique) limit point z. Finally, the entire sequence $\{z^k\}$ converges to z since every convergent subsequence does [49, p. 371. ∎

2.6. On the Monotonicity of the Iterates z^k, k = 0, 1, 2, ...

If M is merely positive definite, one cannot conclude much more about the sequence $\{z^k\}$ than that it converges to a solution of the linear complementarity problem (q,M). However, if we further assume that M is Minkowski and require that $0 < \omega \leq 1$, then a very interesting result obtains. We shall use the following characterization of Minkowski matrices from [17].

Lemma 1. (Cottle and Veinott [17]). M is a Minkowski matrix if and and only if the solution z^* to the linear complementarity problem (q,M) is the unique vector minimum* of the polyhedral set $Z = \{z : Mz + q \geq 0,\ z \geq 0\}$.

Lemma 2. If M is Minkowski, $q_1 \leq q_2$ and z_i^* solves (q_i, M), then $z_1^* \geq z_2^*$.

Proof. By Lemma 1, z_i^* is the vector minimum of $Z_i = \{z : Mz + q_i \geq 0,\ z \geq 0\}$, i = 1, 2. But $q_1 \leq q_2$ impliesthat $Z_1 \subset Z_2$, so $z_1^* \in Z_2$. Therefore $z_1^* = z_2^*$. ∎

Theorem 3. Let M be a Minkowski matrix having diagonal blocks M_{ii}, i = 1, 2, . . . , m. If $\omega \in (0,1]$ and $z^0 = 0$, then $z^{k+1} \geq z^k$ for all k = 0, 1, 2, ...

*I.e., $z^* \in Z$ and $z^* \leq z$ for all $z \in Z$.

Proof. The proof is by induction. Assume M is partitioned into submatrices M_{ij}, $i, j = 1, 2, \ldots, m$, and that q and z^k are conformably partitioned (e.g., $z^k = (z_1^k, z_2^k, \ldots, z_m^k)$). Since $z^1 \geq 0 = z^0$, we may assume that $z^k > z^{k-1} > \ldots > 0$ and $z_j^{k+1} \geq z_j^k$ for $j < i$. Let $\bar{z}_i^{k+1}$ solve the linear complementarity problem

$$\left(\sum_{j<i} M_{ij} z_j^{k+1} + \sum_{j>i} M_{ij} z_j^k + q_i, M_{ii} \right).$$

Since

$$\sum_{j<i} M_{ij} z_j^{k+1} + \sum_{j>i} M_{ij} z_j^k + q_i \leq \sum_{j<i} M_{ij} \bar{z}_j^k + \sum_{j>i} M_{ij} z_j^{k-1} + q_i,$$

Lemma 2 implies that $\bar{z}_i^{k+1} \geq \bar{z}_i^k$.

Recall that $z_i^{k+1} = z_i^k + \omega(\bar{z}_i^{k+1} - z_i^k)$ where $\omega \in (0,1]$. We next show that $z_i^{k+1} \geq z_i^k$ for all i and all k. Clearly, $\bar{z}_i^1 \geq z_i^1 \geq z_i^0 = 0$, so may assume $\bar{z}_i^r \geq z_i^r$ for $r \leq k$. Therefore,

$$z_i^{k+1} = \omega \bar{z}_i^{k+1} + (1-\omega) z_i^k \geq \omega \bar{z}_i^k + (1-\omega) z_i^k \geq \omega z_i^k + (1-\omega) z_i^k = z_i^k . \quad \blacksquare$$

Theorem 3 may be made more intuitive by examining a simple case in which m = 2 and $n_1 = n_2 = 1$. In Figure 6, we illustrate the sequence $\{z^k\}$ generated by the algorithm when ω is equal to one. The zigzagging which occurs causes slow convergence as we approach the solution z^*. This problem is mitigated when values of ω greater than-one are used. However, in those cases, we lose the monotonicity of the vectors $\{z^k\}$.

The next theorem shows that we can also approach z^* from above.

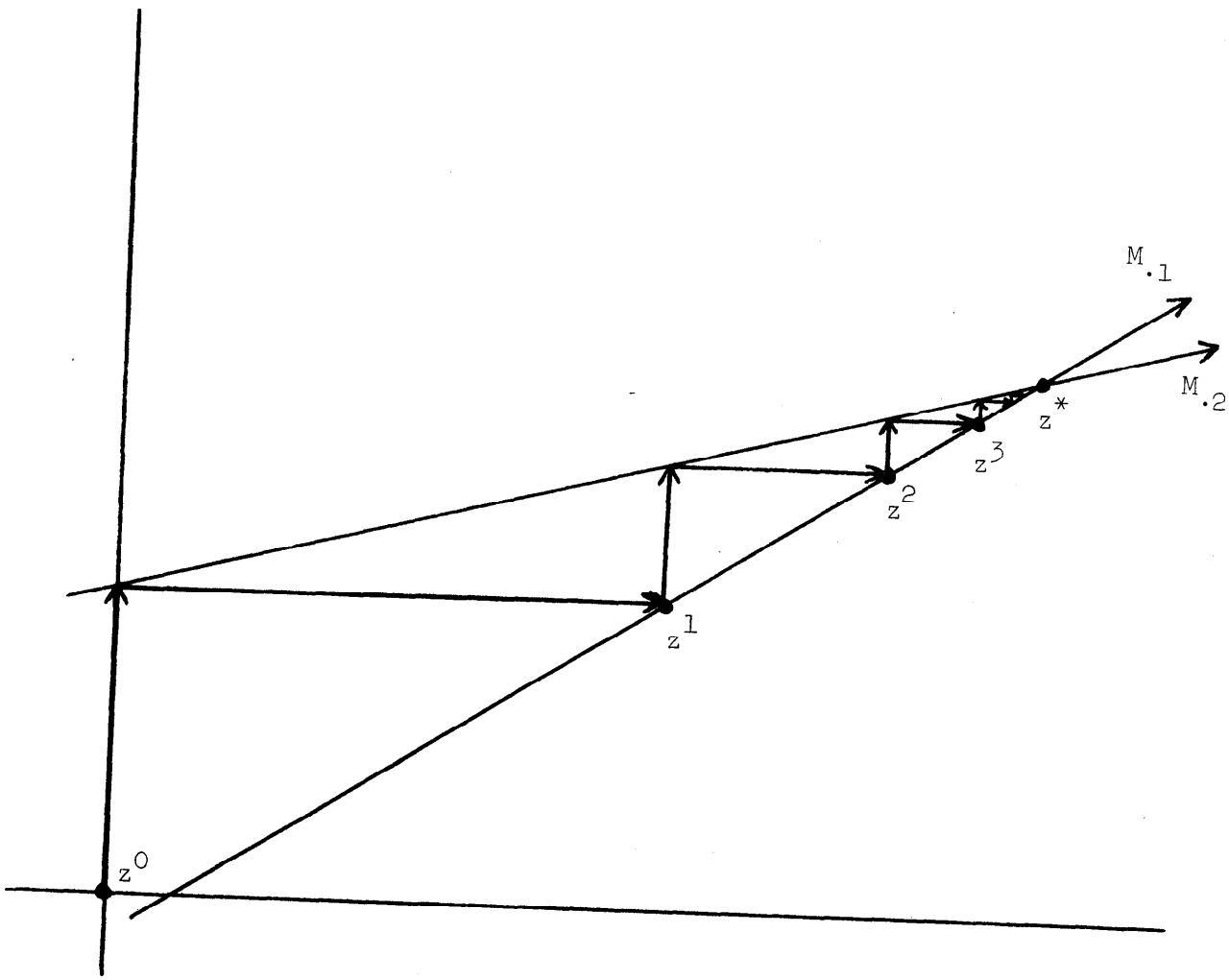


Figure 6. Path of' Algorithm IV with

$\omega = 1$, $m = 2$, $n_i = 1$.

Theorem 4. Let M be a Minkowski matrix having diagonal blocks M_{ii}, $i = 1, 2, \ldots, m$. Furthermore, let z^* solve (q,M). If $\omega \in (0,1]$, $Mz^0 + q \geq 0$ and $z^0 \geq z^*$, then $z^{k+1} \leq z^k$ for all $k = 0, 1, 2, \ldots$

Proof. The proof is by induction. We shall assume that $z^0 \geq z^1 \geq \cdots \geq z^k$ and $Mz^j + q \geq 0$ for $j \leq k$. We may further assume that $z_j^k \geq z_j^{k+1}$ for $j < i$. Let $\bar{z}_i^{k+1}$ solve the linear complementarity problem $\left(\sum_{j<i} M_{ij} z_j^{k+1} + \sum_{j>i} M_{ij} z_j^k + q_i, M_{ii} \right)$. Then, by assumption

$$0 \leq M_{ii} z_i^k + \sum_{j<i} M_{ij} z_j^k + \sum_{j>i} M_{ij} z_j^k + q_i$$

$$\leq M_{ii} z_i^k + \sum_{j<i} M_{ij} z_j^{k+1} + \sum_{j>i} M_{ij} z_j^k + q_i . \qquad (7)$$

Thus, by Lemma 2, $\bar{z}_i^{k+1} \leq z_i^k$ and so $z_i^{k+1} \leq z_i^k$. Furthermore,

$$M_{ii} \bar{z}_i^{k+1} + \sum_{j<i} M_{ij} z_j^{k+1} + \sum_{j>i} M_{ij} z_j^k + q_i \geq 0. \qquad (8)$$

Since z_i^{k+1} is a convex combination of $\bar{z}_i^{k+1}$ and z_i^k, equations (7) and (8) imply that

$$M_{ii} z_i^{k+1} + \sum_{j<i} M_{ij} z_j^{k+1} + \sum_{j>i} M_{ij} z_j^k + q_i \geq 0 .$$

These arguments hold for all $i = 1, 2, \ldots, m$. Since $M_{ij} \leq 0$ (componentwise) for $i \neq j$, $Mz^{k+1} + q \geq 0$. Consequently, $z^k \geq z^{k+1}$ and $Mz^{k+1} + q \geq 0$ for all k. ∎

Intuitively, one might guess that a "dual" version of Theorem 4 exists. For instance, if M and ω are as above, $z^0 \leq z^*$, and z^0

is an element of the same cone but lies on the other side of the apex, i.e., $Mz^0 + q \leq 0$, then $z^k \leq z^{k+1}$ for all k = 0, 1, 2, ... However, $Mz^0 + q \leq 0$ may imply that $z^0 \not\geq 0$, an undesirable situation. A slight modification of these hypotheses will correct this problem.

Theorem 5. Let M be a Minkowski matrix having diagonal blocks M_{ii}, i = 1, 2, . . . , m. Furthermore, let z^* solve (q,M) and assume that $(z_i^*)_t > 0$ implies that $(\sum_{j=1}^{m} M_{ij}z_j^0 + q_i)_t \leq 0$. If $w \in (0,1]$ and $0 \leq z^0 \leq z^*$, then $z^k \leq z^{k+1}$ for all k = 0, 1, 2, ...

Proof. The proof is by induction. We first establish that $(z_i^*)_t = 0$ implies that $(z_i^r)_t = 0$ for all r = 0, 1, 2, ... Since $0 \leq z^0 \leq z^*$, we may assume that $0 \leq z^r \leq z^*$ for r = 0, 1, 2, . . . , k and $0 \leq z_j^{k+1} \leq z_j^*$ for $j < i$. Therefore

$$\sum_{j<i} M_{ij}z_j^{k+1} + \sum_{j>i} M_{ij}z_j^k + q_i \geq \sum_{j<i} M_{ij}z_j^* + \sum_{j>i} M_{ij}z_j^* + q_i .$$

By Lemma 2, $0 \leq z_i^{k+1} \leq z_i^*$ and consequently $0 \leq z_i^{k+1} \leq z_i^*$. Thus, $0 \leq z^r \leq z^*$ for all r = 0, 1, 2, ... and we have resolved our first problem.

Next, suppose $\hat{z} = (z_1^{k+1}, z_2^{k+1}, \ldots, z_{i-1}^{k+1}, z_i^k, z_{i+1}^k, \ldots, z_m^k)$ is known and $\hat{z}$ satisfies the hypotheses of the theorem. We may determine z_i^{k+1} by applying Algorithm III (modified point SOR) to the linear complementarity problem

$$(p,A) \equiv \left(\sum_{j<i} M_{ij}z_j^{k+1} + \sum_{j>i} M_{ij}z_j^k + q_i,\ M_{ii} \right) .$$

We let $x^0 = z_i^k$ be the initial guess, $x^t = (x_1^t, x_2^t, \ldots, x_{n_i}^t)$ be the successive iterates, and let $x^* = \lim_{t \to \infty} x^t$. (Note $x^* = z_i^{k+1}$.) We shall demonstrate that if x^t satisfies the hypotheses, then x^{t+1} will also. Assume $y \equiv (x_1^{t+1}, x_2^{t+1}, \ldots, x_{j-1}^{t+1}, x_j^t, x_{j+1}^t, \ldots, x_{n_i}^t)$ has been generated by the algorithm and satisfies the hypotheses. There are two cases.

Case 1. $(z_i^*)_j = 0$. Since $p \geq \sum_{r<i} M_{ir} z_r^* + q$, then $x^* \leq z_i^*$ and $x_j^* = 0$. Since x_j^0 must also equal zero, we may assume that $x_j^t = 0$ for all $t = 0, 1, 2, \ldots$.

Case 2. $(z_i^*)_j > 0$. Then $(Ay + p)_j \leq 0$. But $x_j^{t+1} = \max\{0, x_j^t - \omega(Ay + p)_j\} \geq x_j^t$. Furthermore, $A(x_1^{t+1}, x_2^{t+1}, \ldots, x_j^{t+1}, x_{j+1}^t, x_{j+2}^t, \ldots, x_{n_i}^t) + p \leq 0$ since $a_{ij} \leq 0$ for $i \neq j$.

We may conclude that $z_k^{k+1} = x^* \leq z^*$, $z_i^k \leq z_i^{k+1}$ and thus $z_i^k \leq z_i^{k+1}$. Since $M_{ij} \leq 0$ (elementwise) for $i \neq j$, then $(z_1^{k+1}, z_2^{k+1}, \ldots, z_i^{k+1}, z_{i+1}^k, z_{i+2}^k, \ldots, z_m^k)$ satisfies the hypotheses. The rest follows by induction. ∎

2.7. The Algorithm Interpreted as a Manifold Suboptimization Technique.

In this section, we shall transfer our attention from the linear complementarity problem to its related quadratic program. In order to facilitate the following discussion, we create a more general setting for the problem. We may view the function to be minimized as one defined on the product space $V = \prod_{i=1}^{m} V_i$. Consequently we have $z = (z_1, z_2, \ldots, z_m) \in V$ where $z_i \in V_i$. Each z_i will be

restricted to the subset $E_i \subset V_i$; thus $z \in E = \prod_{i=1}^{m} E_i$. Let $\langle \cdot, \cdot \rangle_i$ be the scalar product corresponding to V_i, and let M_{ij} be a linear transformation from V_j to V_i. Then the function $f(z) = f(z_1, z_2, \ldots, z_m)$ may be defined as

$$f(z) = \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{n} \langle z_i, M_{ij} z_j \rangle_i + \sum_{i=1}^{m} \langle q_i, z_i \rangle_i$$

where $q_i \in V_1$. In the case of Algorithm IV (Modified Block SOR), $V_i = R^{n_i}$, $E_i = \{x : x \in V_i, x \geq 0\}$, and $M_{ij} \in R^{n_i \times n_j}$. Recall that we assumed that the matrix M (having partitions M_{ij}) is positive definite and that the M_{ii} are, furthermore, symmetric. With this notation, we may state an algorithm for the minimization of f over E.

Algorithm V.

Step. Let $z^0 = (z_1^0, z_2^0, \ldots, z_m^0) \in E$ and let $\omega \in (0,2)$ be given. Set $k = 0$ and $i = 1$.

Step 1. Determine $\bar{z}_i^{k+1} \in E_i$ for which

$$f(z_1^{k+1}, z_2^{k+1}, \ldots, z_{i-1}^{k+1}, \bar{z}_i^{k+1}, z_{i+1}^{k}, \ldots, z_m^{k})$$

$$\leq f(z_1^{k+1}, z_2^{k+1}, \ldots, z_{i-1}^{k+1}, v, z_{i+1}^{k}, \ldots, z_m^{k})$$

for every $v \in E_i$.

Step 2. Let $\omega_i^{k+1} = \max\{\bar{\omega} : \bar{\omega} \leq \omega, \ z_i^k + \bar{\omega}(\bar{z}_i^{k+1} - z_i^k) \in E_i\}$.

Let $z_i^{k+1} = z_i^k + \omega_i^{k+1}(\bar{z}_i^{k+1} - z_i^k)$.

Step 3. If $i = n$, go to Step 4. Otherwise return to Step 1 with i replaced by $i+1$.

Step 4. Is z^{k+1} "reasonably" close to the solution? If so, stop. Otherwise, return to Step 1 with k replaced by $k+1$ and $i = 1$.

Notice that Steps 0, 1, 3 and 4 of Algorithm V are essentially identical to the corresponding steps of Algorithm IV. For the problem described above, the algorithms are, in fact, identical. In Step 1 of Algorithm V, we perform a constrained minimization of f on the manifold of V determined by using fixed values in $E_1, E_2, \ldots, E_{i-1}, E_{i+1}, \ldots, E_m$ and letting the minimization take place in E_i, the constraint set in the space V_i. This is equivalent to solving

$$\text{minimize} \quad f_i(u) = \frac{1}{2} u^T M_{ii} u + \Big(\sum_{j<i} M_{ij} z_j^{k+1} + \sum_{j>i} M_{ij} z_j^k + q_i \Big)^T u \qquad (9)$$

$$\text{subject to} \quad u \in E_1.$$

But M_{ii} is a symmetric positive definite matrix by assumption. Hence $\bar{z}_i^{k+1}$ solves (9) if and only if $\bar{z}_i^{k+1}$ solves the linear complementarity problem

$$\Big(\sum_{j<i} M_{ij} z_j^{k+1} + \sum_{j>i} M_{ij} z_j^k + q_i, \; M_{ii} \Big).$$

However, this is Step 1 of Algorithm IV.

If we let $\omega = 1$, then $\omega_i^{k+1} = 1$ and $z_i^{k+1} = \bar{z}_i^{k+1}$ for all k and all i. In this case, Algorithm V is a typical example of a manifold suboptimization algorithm [63]. When ω is greater than 1, we have an accelerated version of a manifold suboptimization technique. The results of Section 2.5 apply and we have convergence for any value of ω strictly between 0 and 2.

2.6. Related Manifold Suboptimization Techniques.

Methods similar to Algorithm V have appeared in the literature on the minimization of functionals on Hilbert spaces or reflexive Banach spaces. J. Cea [8] treats the case in which the bilinear form (corresponding to our quadratic form $u^T Mv$) is continuous, symmetric and coercive. The sets E_i are closed convex subsets of V_1. Under these hypotheses, Cea proves that if $\omega = 1$, then the z^k, $k = 1,2,\ldots$, converge weakly to the solution.

A. Auslender [2] treats the case in which V_i and E_i are defined as above but where the gradient of f satisfies a uniform Lipschitz condition on the closed, bounded, convex sets of V. If $E_i \not\subseteq V_i$, he requires $\omega \in (0,1]$ for convergence of his algorithm. In the unconstrained ease, i.e., Ei $= V_i$, ω is permitted to assume any value strictly between 0 and 2. If V is finite dimensional, the Lipschitz condition on f is relaxed and replaced by a much weaker condition.

R. Glowinski [30] uses the same hypotheses as Cea. However, Glowinski's algorithm modifies Steps 1 and 2 as follows. He minimizes f over V_i instead of E_i in Step 1. In Step 2, he uses a fixed

value of ω_i for i = 1, 2, . . . , m and guarantees that $z_i^{k+1} \in E_i$ by letting $z_i^{k+1} = P_i(z_i^k + \omega_i(z_i^{k+1} - z_i^k))$ where P_i is the "orthogonal projection operator from V_i to E_i corresponding to the norm* induced by M_{ii}." Glowinski states, without proof, that if the $\omega_i \in (0,2)$, i = 1, 2, . . . , m, then the iterates $\{z^k\}$ converge strongly to the solution.

The research of J.-C. Miellou [43] and of B. Martinet [42] is also of related interest.

* $\|v\| = \langle M_{ii}v, v\rangle$ where $\langle \cdot, \cdot \rangle$ is a scalar product

SECTION 3

APPLICATION--THE JOURNAL BEARING PROBLEM

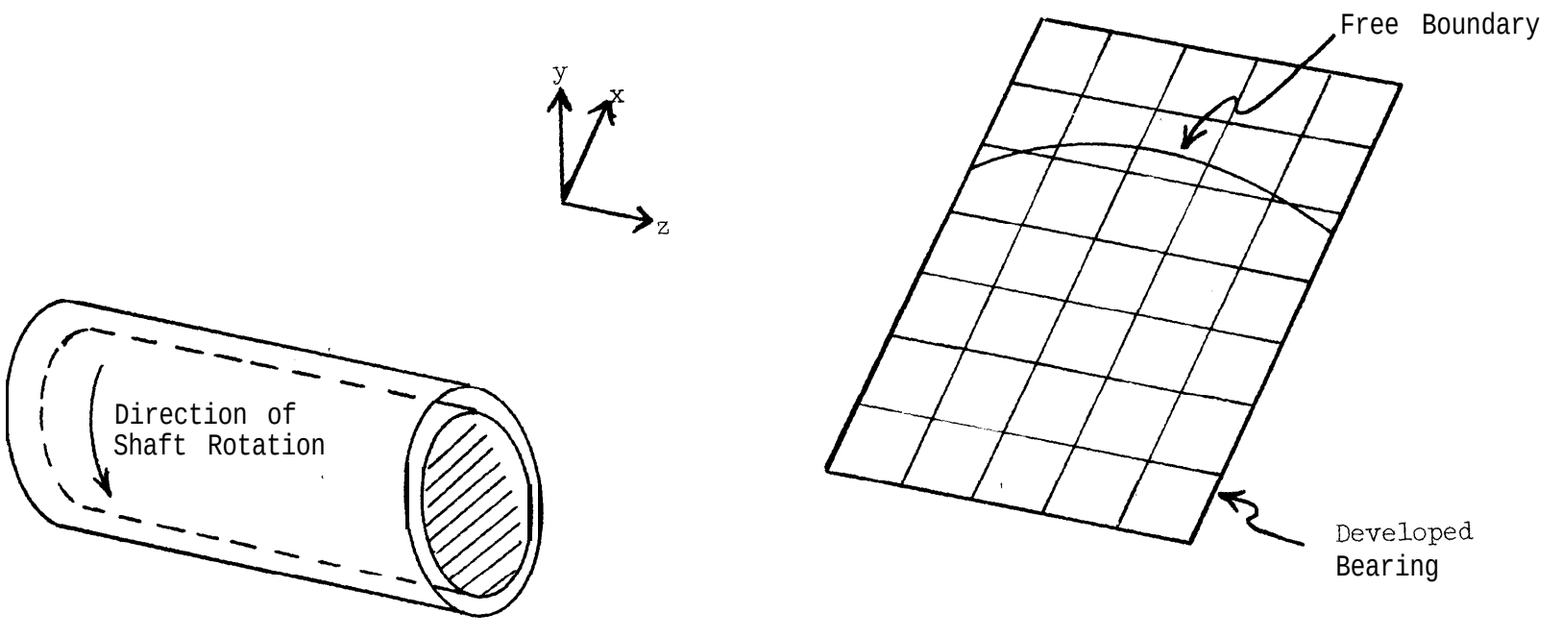
3.1. Statement of the Problem.

A journal bearing consists of a rotating cylindrical shaft (the journal) which is separated from a bearing surface by a film of lubricating fluid. The journal and bearing are of length L and have parallel longitudinal axes (of rotation). A typical journal bearing is shown in Figure 7 as is an unfolding of the bearing surface into the plane. A cross-section perpendicular to the axis of rotation is depicted in Figure 8. The mathematical description of the system will be stated using various coordinate systems as need dictates. A description of the cross-section is most easily couched in polar coordinates whereas a description of the entire journal bearing has a more natural setting in rectangular coordinates.

We wish to know the distribution of pressure on the lubricating film. An important underlying assumption of the model is that the lubricating film is so thin that there is no variation in pressure in the axial direction. Therefore (in Figure 8), the pressure is constant on the "line" from the journal to the bearing for each value of θ. Consequently, one may view the problem as the determination of the pressure distribution on the lubricant of the bearing surface.

An initial understanding of the journal bearing model may be obtained by first examining the cross-section of Figure 8. We shall review Cryer's [19] description. The thickness* of the film

*I.e., depth, not viscosity.



Surface Velocity = u

Rotation Velocity V = $U/\pi D$

Figure 7. Side View and Exposed (i.e., Developed) View of a Finite Length Journal Bearing

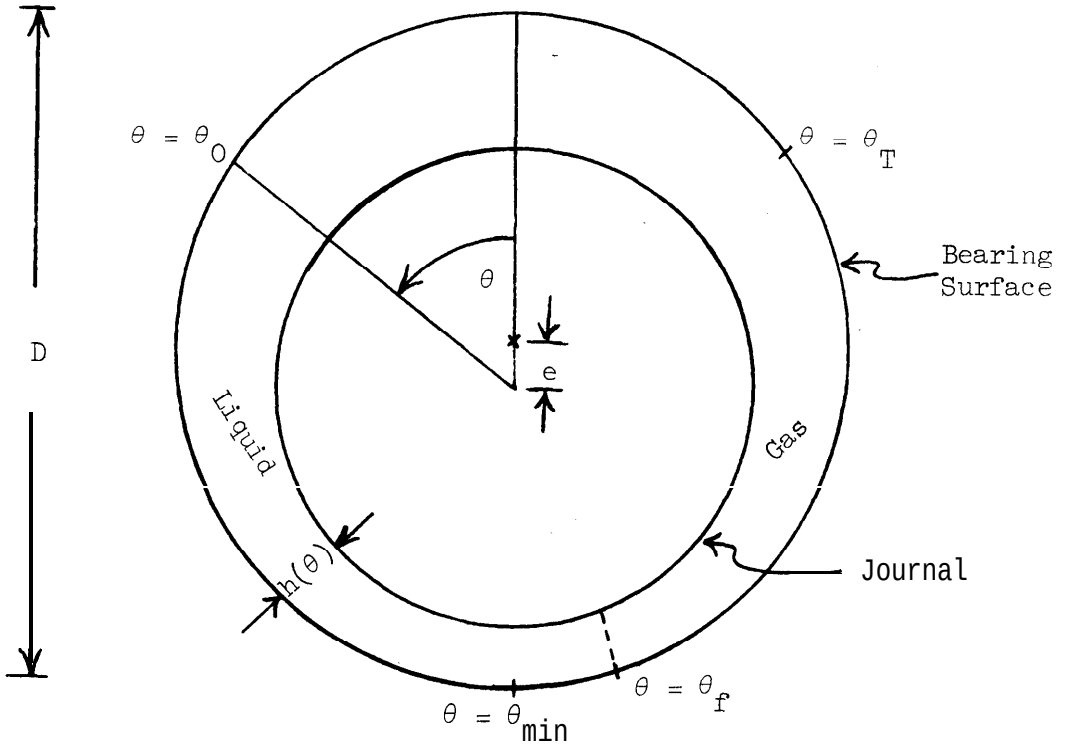


Figure 8. Cross Section of a Journal Bearing

("*" marks the center of the bearing)

is minimum at θ_{min} where the angle θ measures rotation about the z-axis, the axis of rotation. In the case of a partial bearing (one which does not completely encase the journal) the lubricant flows out at θ_T and is replenished at θ_0. In the case of a full bearing, where $\theta_T = \theta_0 + 2\pi$, the liquid which may have vaporized is assumed to condense at θ_T into its previous liquid state. (In the full model of Figure 6, the lubricant can also flow out of both ends of the journal bearing.) The thickness of the film is denoted by $h(\theta,z)$; it satisfies

$$h(\theta,z) > 0 \qquad \theta \in [\theta_0,\theta_T]$$

$$\frac{\partial h}{\partial \theta} < 0 \qquad \theta \in (\theta_0,\theta_{min})$$

$$\frac{\partial h}{\partial \theta} > 0 \qquad \theta \in (\theta_{min},\theta_T) .$$

The pressure on the film can be expected to increase between $\theta = \theta_0$ and $\theta = \theta_{min}$ and to decrease between $\theta = \theta_{min}$ and θ_T. It is assumed that when $\theta = \theta_f$, the pressure becomes so low that the lubricant vaporizes. The interface between the two boundaries of the lubricant is called the free boundary (see [39]). In the finite length bearing of Figure 7, the location of the free boundary depends on the axial coordinate z and is denoted by $\theta_f(\cdot)$. The pressure is zero (i.e., atmosphere) along and beyond the free boundary θ. This is discussed in more detail in Section 3.4. In Figure 9, we illustrate the profile of the pressure distribution on the lubricant at the

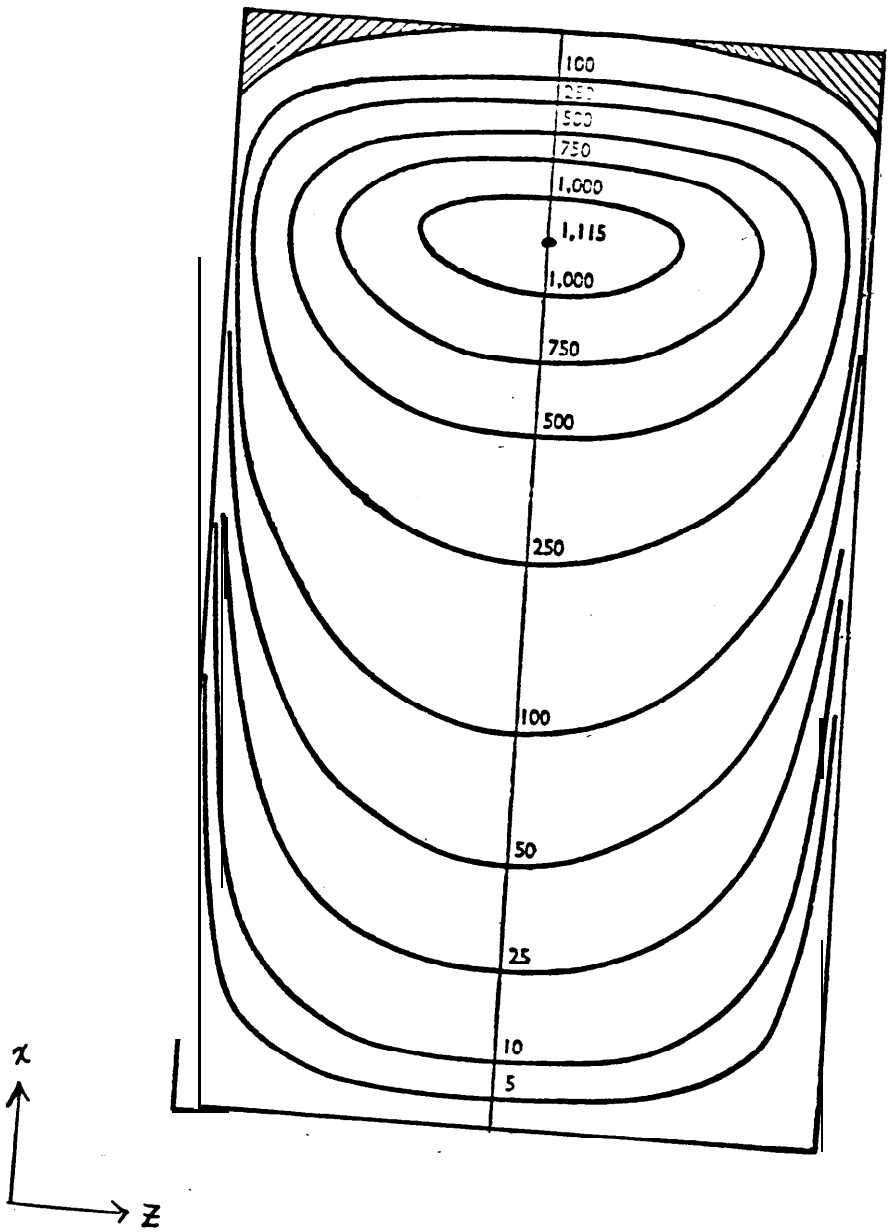


Figure 9. Isobars for Developed Bearing with $\epsilon = 0.8$, $\ell_L = 1$
(Excerpted with Permission of the Publisher Instit. Mech. Eng., from [4].)

(developed) bearing surface. Tcomputations and graph were Cone by Cameron and Wood [4]. This journal bearing has an eccentricity ratio $\epsilon = e/r$ equal to 0.8 and a bearing diameter-to-length ratio of D/L equal to 1. The isobars (constant pressure contours) are given in nondimensional units $(10^3R^3/\mu Ur^2)p$. The variable p is the pressure, R is the bearing radius, μ is the viscosity of the lubricant, U is the surface velocity of the journal, r is the minimum clearance between the bearing and the journal and e is the distance between the two axes (see Figure 8).

3.2. The Reynolds Equation

In 1886, Osborne Reynolds [48] developed the now-classic equation governing the mechanism of hydrodynamic lubrication by incompressible fluids. The equation, a special case of the more general Navier-Stokes equation [45, p. 4], is deduced from seven essential assumptions on the physical properties of the system (see [45, p. 6]).

(i) The dimensions are sufficiently large to justify ignoring the curvature of the journal bearing when studying a small section of it.

(ii) The pressure across the film (from the journal to the bearing) is constant; i.e., $\partial p/\partial y = 0$.

(iii) The flow is laminar, i.e., there is no turbulence in the film.

(iV) There are no external forces acting on the film.

(v) The fluid inertia is small compared to the viscous shear. This means that the rotational forces of the journal acting on the lubricant are much larger than the natural tendency (e.g., fluid gravity) of the fluid to remain at rest.

(vi) There is no slippage of the fluid at the bearing surfaces.

(vii) If u and w are the velocities of the fluid in the x- and z-directions, respectively (see Figure 13), then all velocity gradients are negligible compared to $\partial u/\partial y$ and $\partial w/\partial y$.

Pinkus and Sternlicht [45] note that in most practical cases, the bearing is stationary and only the shaft is moving. In these cases, the most general form of the Reynolds equation is

$$\frac{\partial}{\partial x}\left(\frac{\rho h^3}{\mu}\frac{\partial p}{\partial x}\right)+\frac{\partial}{\partial z}\left(\frac{\rho h^3}{\mu}\frac{\partial p}{\partial z}\right)=6U\frac{\partial(\rho h)}{\partial x}+12\rho V_0 . \qquad (10)$$

The variable ρ represents the density of the lubricant, μ is the absolute viscosity, and V_0 represents a velocity resulting from the motion of the journal center. In the ensuing discussion of equation (10), we will postulate that $V_0 \equiv 0$ and that ρ and μ are constants.

In order to gain a better understanding of the model of a journal bearing of finite length, we first examine a simpler model. By means of this special case, we may motivate the boundary conditions for the problem of more general interest.

3.3. A Limiting Case: The Infinite Length Full Journal Bearing.

If we suppose that the length L of the journal bearing is infinite, certain further simplifications may be made. We may disregard the effect of fluid flow from the ends of the bearing and therefore $\partial p/\partial z$, the pressure gradients in the axial direction, will be zero.

Obviously, an infinitely long journal bearing is a physical impossibility and does not closely approximate the dimensions of those used in practice. However, it does provide some understanding of' the behavior of more realistic bearings. Some notable similarities between the finite and infinite length models are the following. The infinite case provides upper limits on both the pressure exerted on the fluid film and on the loads which the film will support. Moreover, Pinkus and Sternlicht [45, pp. 69-71] show that the solution to equation (10) (which describes the finite length journal bearing) is a perturbation to the solution of the infinite length journal bearing problem. The perturbation involves adding the product of the solutions of two differential equations of a single variable. (To the authors knowledge, this realization has not borne fruit due to the difficulty of solving the latter two differential equations.)

As Pinkus and Sternlicht indicate [45, p. 68], the difficulty in obtaining satisfactory solutions for journal bearing problems lies not only in solving a given formulation but in adequately defining the boundary conditions for the formulation. For the remainder of the paper, we shall assume $\theta_0 = 0$. In order to determine these boundary conditions for the simpler model, we first recall that there is no pressure variation in the axial direction. Consequently, it is sufficient to examine an arbitrary cross-section perpendicular to the axial direction (see Figure 8). Generalization to the finite length case (where, for a given θ, there is pressure variation in the axial direction z) may then be thought of as the examination of a collection of cross-sections along the z-axis, say at $z = z_1, z_2, z_3, \ldots, z_N$, where

the boundary conditions for the i^{th} cross-section are analogous to those for the infinite-length journal bearing model.

Replacing the variable x by $R\theta$, where θ is in radians and R is the bearing radius, and recalling that $\partial p/\partial z = 0$, then the Reynolds equation for an infinite journal bearing is

$$\frac{\partial}{\partial\theta}\left[h^3 \frac{\partial p}{\partial\theta}\right] = 6\mu R \frac{\partial h}{\partial\theta} . \qquad (11)$$

We may use full instead of partial derivatives because both p and h are now functions of θ alone. Furthermore, since $6\mu R$ is a constant, a change of units is sometimes made to allow setting it equal to unity.

In 1904, Sommerfeld obtained* the first solution to equation (11); he addressed the full journal bearing case in which the boundary values were $P(0) = p(2\pi) = p_\alpha$. He also assumed that both journal and bearing were cylindrical and hence (h being a function of θ only),

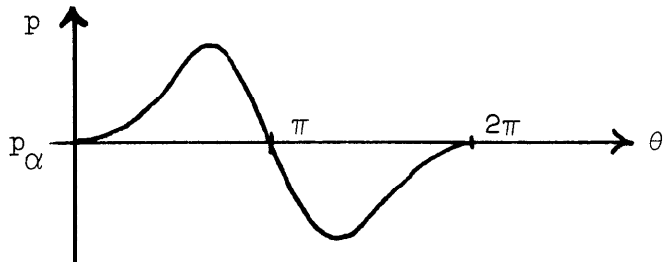
$$h(\theta) = r(1 + \epsilon \cos\theta) .$$

The parameter p_α is the ambient (or atmospheric) pressure and is usually set equal to zero. Sommerfeld's expression for the pressure distribution was

$$p(\theta) = p_\alpha + \frac{6\mu UR\epsilon}{r} \frac{(2 + \epsilon \cos 0) \sin\theta}{(2 + \epsilon^2)(1 + \epsilon \cos\theta)^2} \qquad (12)$$

Setting p_α equal to zero, the graph of $p(\theta)$ becomes

*by a clever transformation of variables



The important thing to notice is that equation (12) yields regions of high negative pressure. This model was unacceptable since a lubricating fluid cannot support such high negative pressures and still remain an incompressible fluid. The underlying problem was that as θ increased beyond $\theta_{min} = \pi$, the width of the film increased and consequently the pressure exerted on the film decreased. Eventually, at $\theta = \theta_f$, the pressure became so low that the tensile strength of the fluid was overcome and the fluid vaporized. Since the Reynolds equation only holds for <u>incompressible</u> lubricating fluids and the region of the journal bearing beyond the free boundary, i.e., $\theta > \theta_f$, contained a <u>compressible</u> gaseous lubricant, it was no longer valid to apply equation (11) over the region (θ_f, θ_T).

Thus, a different set of boundary conditions was needed to provide a more realistic solution to the problem of determining both the region (θ_0, θ_f) in which the lubricant exists as a liquid and the pressure $p(\theta)$ in that region. From the literature, one infers that the boundary conditions commonly used today are due to Swift [55]. They state that when the pressure falls to zero, the

circumferential pressure gradient $dp/d\theta$ also falls to zero. In other words,

$$\text{(i)} \quad P(0) = 0$$

$$\text{(ii)} \quad p(\theta_f) = 0$$

$$\text{(iii)} \quad \frac{dp}{d\theta}(\theta_f) = 0 \ .$$

Clearly, the pressure function p may be continuously extended on the interval $[\theta_f, \theta_T]$ by setting it equal to zero on that interval. From the results of Cryer [19], the free-boundary θ_f occurs at the largest value of $\theta = \bar{\theta}$ for which p is nonnegative on $[\theta_0, \bar{\theta}]$.

Sommerfeld's technique for solving the differential equation with these boundary conditions is still applicable and yields the following complex expression for $p(\theta)$ in terms of another angle, ψ

$$p(\psi) = \frac{6\mu UR}{r^2(1-\epsilon^2)^{3/2}} \left\{ \psi - \epsilon \sin \psi - \frac{(2+\epsilon)\psi - 4\epsilon \sin \epsilon + \epsilon^2 \sin \psi \cos \psi}{2(1 + \epsilon \cos(\psi_f - 7\text{-}r)} \right\} \tag{13}$$

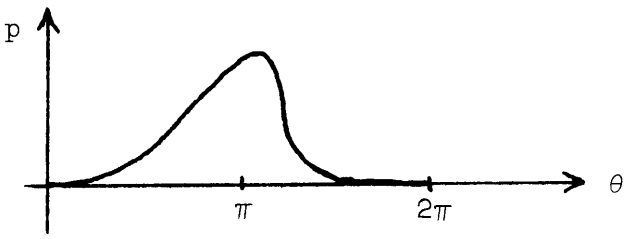
where $$\cos \psi \frac{\epsilon + \cos \theta}{1 + \epsilon \cos \theta}$$

and ψ_f corresponds to θ_f.

The location of the free boundary θ_f is not immediately apparent from the original problem. However, the boundary condition $p(\theta_f) = 0$ yields an implicit formula for ψ_f.

$$\epsilon(\sin(\psi_f - \pi)\cos(\psi_f - \pi) - \psi_f) + 2(\psi_f \cos(\psi_f - \pi) - \sin(\psi_f - \pi)) = 0$$

The solution under these new boundary conditions has the following graph.



Notice that the regions of negative pressure found in the graph of equation (12) do not appear in this graph. This is the principal reason for using the boundary conditions of Swift.

3.4. The Use of Finite Difference Techniques.

Before leaving the case of the infinite length journal bearing, we wish to discuss a situation where Somrnerfeld's technique does not apply and where no other means of obtaining an exact solution is currently known. An example of this might be one where the bearing is not cylindrical and hence the width function h does not have the common form $h(\theta) = r(1 + \epsilon \cos \theta)$. In 1941, Christopherson [10] proposed a technique forsolving free boundary problems for journal bearings by means of approximating the differential equation by finite differences. Later, improvements on Christopherson's method were made by Raimondi and Boyd [46] and by Gnanadoss and Osborne [29]. The former solved the difference equations by modifying the Liebman (or Gauss-Seidel) method, the latter by modifying successive over-relaxation (SOR). In 1971, Cryer [19] analyzed the numerical aspects

of Christopherson's algorithm with the SOR modification when applied to the infinite journal bearing case. He proved that if

(i) Δ is the interval length on the approximation grid,

(ii) $p_j = p(j\Delta)$ is the true value of the pressure at $j\Delta$, i.e., at the j^{th} grid point, $j = 1, 2, \ldots, N$,

(iii) P_j is the discrete approximation value at the j^{th} grid point, $j = 1, 2, \ldots, N$, and

(iv) Δ is sufficiently small, then there is a $K < \infty$ for which

$$\max_j |p_j - P_j| < K\Delta^2$$

Furthermore, he showed that the boundary conditions (in particular, the "free boundary") cause this problem to be equivalent to a linear complementarity problem (q,M). The matrix M corresponds to the finite difference equations which are fully discussed in [19].

3.5* The Finite Length Journal Bearing Model and an Approximation.

A realistic mathematical model of a finite length journal bearing has great potential for becoming very complicated. For instance, the lubricant can be admitted through oil grooves to the bearing at any angle and the larger the angle, the more pronounced is its effect on the resulting pressure distribution. Further, the lubricant is not always admitted at atmospheric (i.e., zero) pressure. These and other factors contribute a significant complexity to the formulation.

In our discussion of the finite length case, we shall treat a fairly simple model, one in which the bearing is a full (as opposed to partial) cylindrical bearing. Ambient pressure is taken to be zero. As in the infinite length case, the lubricant that vaporizes at the free boundary is assumed to condense along the line where $\theta = 0$. The boundary conditions are a natural generalization of (i)-(iii) for the infinite length case (see [29]). As indicated before, it is easier to present the finite length case in rectangular coordinates. Referring to the bearing surface of Figure 7, we shall let $p(x,z)$ represent the pressure on the lubricant along the bearing surface. The boundary conditions are

(i)' $p(0,z) = 0$ for all z,

(ii)' $p(2\pi D, z) = 0$ for all z,

(iii)' $\frac{\partial p}{\partial z}(x, \frac{L}{2}) = 0$ for all x,

(iv)' $p(\theta_f(z),z) = 0$ for all z, and

(v)' $\frac{\partial p}{\partial n}(\theta_f(z), z) = 0$ for all z,

where θ_f is the free boundary, and $\frac{\partial p}{\partial n}(\theta_f(z), z)$ is the normal derivative of p at $(\theta_f(z), z)$, i.e., the derivative of p in the direction normal to the tangent of the free boundary θ_f at $(\theta_f(z),z)$. (In the case of the infinite length journal bearing, the normal derivative at θ_f becomes $\frac{\partial p}{\partial \theta}(\theta_f) = 0$ as in (iii) of Section 3.3.)

Since even this relatively simple model of the finite length journal bearing has eluded attempts to obtain a closed form solution by analytic means, other avenues have been explored and have met with

more success. These alternate methods have included electrolytic tank models, d-c analogues and finite difference models (see [45]). It is the last category to which Christopherson's method belongs.

To develop the discrete model, we shall first follow Pinkus and Sternlicht [45, pp. 79-81] in deriving a five-point finite difference approximation to the Reynolds equation. By a change of variables, we first obtain a dimensionless version of equation (1). Let $\bar{x} = x/D$, $\bar{z} = z/L$, $\bar{h} = h/2r$, and $\bar{p} = (r^2/\mu VR^2)p$ where V is the speed of the journal measured in revolutions per unit time. This yields

$$\frac{\partial}{\partial \bar{x}}\left(\bar{h}^3 \frac{\partial \bar{p}}{\partial \bar{x}}\right) + \left(\frac{D}{L}\right)^2 \frac{\partial}{\partial \bar{z}}\left(\bar{h}^3 \frac{\partial \bar{p}}{\partial \bar{z}}\right) = 6\pi \frac{\partial \bar{h}}{\partial \bar{x}} \quad . \tag{14}$$

Dropping the bars above the variables and referring to Figure 16, we have the following finite difference representations.

$$\frac{\partial}{\partial x}\left(h^3 \frac{\partial p}{\partial x}\right) = \frac{h^3_{i,j+1/2}\left(\frac{p_{i,j+1}-p_{i,j}}{\Delta x}\right) - h^3_{i,j-1/2}\left(\frac{p_{i,j}-p_{i,j-1}}{\Delta x}\right)}{\Delta x},$$

$$\frac{\partial}{\partial z}\left(h^3 \frac{\partial p}{\partial z}\right) = \frac{h^3_{i+1/2,j}\left(\frac{p_{i+1,j}-p_{i,j}}{\Delta z}\right) - h_{i-1/2,j}\left(\frac{p_{i,j}-p_{i-1,j}}{\Delta z}\right)}{\Delta z}$$

$$\frac{\partial h}{\partial x} = \frac{h_{i,j+1/2} - h_{i,j-1/2}}{\Delta x}$$

After rearranging terms, the evaluation of the equation at grid point

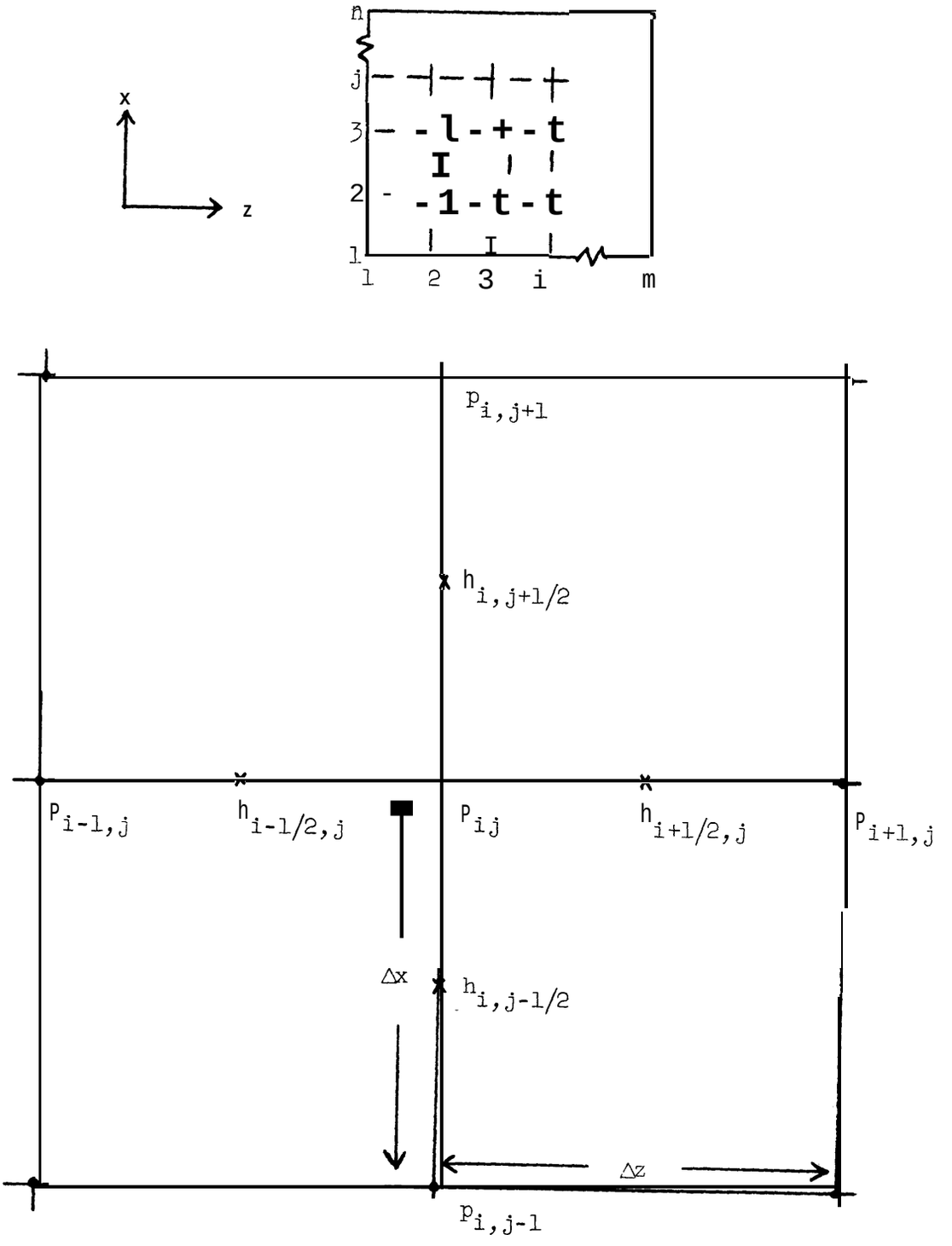


Figure 10. Finite Difference Approximation

(i,j) takes the form

$$W_{ij} = q_{ij} + a_{i,j,1}p_{i,j-1} + a_{i,j,2}p_{i-1,j} + a_{i,j,3}p_{i,j} + a_{i,\,j,4}p_{i+1,j} + a_{i,j,5}p_{i,j+1} \tag{15}$$

where

$$a_{i,j,1} = -h^3_{i,j-1/2}/(\Delta x)^2,$$

$$a_{i,j,2} = -\left(\frac{D}{L}\right)^2 h^3_{i-1/2,j}/(\Delta z)^2 ,$$

$$a_{i,\,j,3} = \left(\frac{D}{L}\right)^2 \frac{(h^3_{i+1/2,j} + h^3_{i-1/2\ j})}{(\Delta z)^2} + \frac{(h^3_{i,j+1/2} + h^3_{i,j-1/2})}{(\Delta x)^2} ,$$

$$a_{i,j,4} = -\left(\frac{D}{L}\right)^2 h^3_{i+1/2,j}/(\Delta z)^2 ,$$

$$a_{i,j,5} = -h^3_{i,j+1/2}/(\Delta x)^2 ,$$

$$q_{i,j} = 6\pi \frac{(h_{i\ j+1/2} - h_{i,j-1/2})}{\Delta x}$$

and $w_{ij} = 0$ if the pressure at (i,j) satisfies the Reynolds equation. If $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, n$, then the discretized version of equation (1) is an $(mn) \times (mn)$ linear system. For each i, we define the entries of the matrix M as

$$
\begin{aligned}
m_{i,j-n} &= a_{k,\ell,1} \\
m_{i,j-1} &= a_{k,\ell,2} \\
m_{i,j} &= a_{k,\ell,3} \\
m_{1,\,j+1} &= a_{k,\ell,4} \\
m_{i,j+n} &= a_{k,\ell,5}
\end{aligned}
\tag{16}
$$

and $m_{i,r} = 0$ for all other r

where $k-1$ is the largest integer not exceeding i/n and where $\ell = i-kn$. In addition we let the subvector $p_i = (p_{i1}, p_{i2}, \ldots, p_{in})$ and the vector $p = (p_1, p_2, \ldots, p_m)$; we define the vector q similarly.

The matrix M and corresponding vector q form the basis for an approximation to the model of a finite length journal bearing having a free boundary θ. As in the infinite length case, there is an associated linear complementarity problem (q,M) whose equivalence is illustrated by a synthesis of Christopherson's original application [10] of his method to the finite length journal bearing problem and Cryer's later discussion [19] of the method and its application. Intuitively, the complementarity problem arises as follows. Denote the region where the lubricant exists in its liquid (vaporized) state as the positive (zero) region. These appellations refer, of course, to the pressure on the lubricant in those regions. In the positive region, the Reynolds equation is required to be satisfied. Hence, if the grid point (i,j) belongs to the positive region, then $w_{ij} = 0$ and the discretized version, equation (15), becomes

$$w_{ij} = q_{ij} + m_{1,\,j-n}p_{j-n} + m_{i,j-1}p_{j-1} + m_{i,j}p_j + m_{1,\,j+1}p_{j+1} + m_{i,j+n}p_n = 0.$$

On the other hand, if the grid point (i,j) belongs to the 'interior" of the zero region, then the pressure variables associated with the adjacent points $(p_{i,j-n}, p_{i,j-1}, p_{i,j+1}, p_{i,j+n})$ have zero value. Consequently, equation (15) becomes $w_{ij} = q_{ij} = 6\pi(h_{i,j+1/2} - h_{i,j-1/2})/\Delta x$. However, the location of the free boundary and the zero region requires $h_{i,j+1/2} - h_{i,j-1/2}$, and hence w_{ij}, to be nonnegative. Summarizing, we have a variable p_{ij} and an algebraic expression w_{ij} associated with the point (i,j) and related by $w = Mp + q$. If p_{ij} is positive, then w_{ij} equals zero and if p_{ij} is zero, then w_{ij} is nonnegative, i.e., p and w satisfy the conditions of the linear complementarity problem (q,M).

If the bearing is cylindrical in the example discussed above, then h(x,z) is independent of z and consequently $h_{i-1/2,j}$ and $h_{i+1/2,j}$ is independent of i. From this observation, we may draw several conclusions about the matrix M.

(i) M is a symmetric block tridiagonal Minkowski matrix where $M_{ij} \in R^{n\times n}$ and i, j = 1, 2, . . . , m.

(ii) $M_{i+1,i} = M_{i,i+1} = \alpha_i I$ where $\alpha_i < 0$ and i = 1, 2, ..= , m.

(iii) M_{ii} is a tridiagonal matrix whose subdiagonal and superdiagonal entries are identical and whose diagonal entries are identical.

With this structure, the Modified Block SOR Algorithm may be brought to bear on the journal bearing problem. The computational experience reported in the next section demonstrates the efficacy of this approach.

SECTION 4

COMPUTATIONAL ASPECTS OF ALGORITHM IV

4.1. Storage Requirements.

We first address the question of storage requirements for the most general form of Algorithm IV. In this case, M is merely assumed to be positive definite with symmetric diagonal blocks, M_{ii}. The matrix M is partitioned so that M_{ii} is a square matrix of order n_i for i = 1, 2, . . . , m. Then, for each i, there are, say, N_i non-zero double precision matrix entries and n_i double precision entries for each of the subvectors q_i and z_i. If one uses sparse matrix techniques to store the entries of M, additional storage demands are made in the form of row and column index vectors. In the algorithm itself, the updating of the solution vector iterate z_i^k requires sufficient space to solve the complementarity problem $(\bar{q},\bar{M}) = (q_i + \sum_{j<i} M_{ij} z_j^{k+1} + \sum_{j>i} M_{ij} z_j^k, M_{ii})$. This means allocating space for a copy of $\bar{q}$ and $\bar{M}$ as well as any additional space required by the complementarity subroutine. Notice that it is not necessary to have all the initial data constantly available in core. For instance, it is sufficient to have the vector z, the subvector q_i, the submatrices M_{ij} for j = 1, 2, . . . , m and appropriate storage for solving $(\bar{q},\bar{M})$.

By restricting our attention to the block tridiagonal case where M_{ii} is symmetric and tridiagonal and both $M_{i,i+1}$ and $M_{i+1,i}$ are diagonal matrices, we find certain economies in storage. Suppose the diagonal blocks M_{ii} are of order $n_i = n$ for i = 1, 2, . . . , m.

Then, one can easily show that the storage required for M, z and q is $6mn + 4n - 2m - 2$ (8-byte) words.

What further requirements does Step 1 impose? If M_{ii} is solved by Lemke's algorithm or the principal pivoting method, we need approximately n^2 more 8-byte words. If we further assume, as above, that M_{ii} is Minkowski, then Algorithms I-III are applicable. Recall that Algorithms I-III preserve the sparsity of the data. Their additional requirements are approximately 40n, 60n and 40n bytes of storage, respectively. Of course, savings (of $8mn-8n$ bytes) are achieved when M is symmetric and more dramatic savings occur when M corresponds to the finite length journal bearing problem described in Section 3.5. In the latter case, M has attributes (i)-(iii) found on page 58.

These storage estimates represent the minimum necessary for the algorithm. Computational refinements (e.g., reduction of multiplications by zero) make further storage demands in the manner of sparse matrix techniques (i.e., in the form of index sets incorporated into the computer program).

4.2. The Computer Codes. .

Three computer codes have been written for Algorithm IV, the Block Modified SOR Algorithm. They differ from each other in the way that each solves the subproblems found in Step 1. The programs are written in IBM 360/370 Fortran IV and use double precision (8-byte) floating-point arithmetic.

The matrix M is required to be symmetric, block tridiagonal, and positive definite. Furthermore, the diagonal blocks M_{ii} are required to be tridiagonal Minkowski matrices and the off-diagonal blocks $M_{i,i+1}$ and $M_{i+1,i}$ must be diagonal matrices. The "tri-diagonal" linear complementarity problems occurring in Step 1 are solved by Algorithms I, II and III, respectively.

4.3. Computational Experience.

A computational study of the problem (q,M) was performed in which we used two types of matrices M. The "JB" matrix corresponds to equations (15)-(16), the 5-point finite difference approximation to the Reynolds equation arising in the free boundary problem for the journal bearing problem. (The eccentricity ϵ equals 0.8 and the ratio D/L equals 1.) The "LAP" matrix corresponds to the five-point difference approximation to Laplace's equation. (See Figure 3 for an example.) In both cases, the diagonal blocks M_{ii} are of order n and m is set equal to n. Thus the matrix M is of order $N = n^2$ When the JB matrix is used, the q-vector comes in two varieties. One type corresponds to the finite difference equations for the journal bearing. The other is a random vector in which the absolute values of the components are chosen from a uniform distribution on [0,2] and their sign is determined by the formula

$$\operatorname{sgn}(q_j) = \begin{cases} +1 & \text{if } j(\text{mod } \alpha) \leq \beta \\ -1 & \text{if } j(\text{mod } \alpha) > \beta \end{cases}$$

where a and β are given constants. For instance, if $\alpha = 20$ and $\beta = 10$, then the q-vector has a repeating pattern of 11 positive and 9 negative entries. The LAP matrix is used only with the random q-vectors described above.

Algorithm IV uses two parameters, a stopping criterion tolerance ϵ and a relaxation parameter ω. We have set ϵ equal to 10^{-7} and for each experiment, have determined (to within 0.02) the value ω_{exp} of the parameter ω which minimizes the number of iterations to achieve the desired level of error in the solution. (In one of the three codes, we solve Step 1 by Algorithm III, the modified point SOR algorithm. Algorithm III uses its own relaxation parameter ω' and for each experiment, we have determined (to within 0.1) the value ω'_{exp} of the parameter ω' which minimizes the total solution time when $\omega = \omega_{exp}$.)

Finally, we shall use the following nomenclature for the algorithms tested. Let BSORF, BSORP and BSORS denote the three versions of the Modified Block SOR Algorithm with the first solving Step 1 by Algorithm I--the factorization method, the second by Algorithm II--the modified principal pivoting method, and the third by Algorithm III--the modified point SOR algorithm. Also, let PSOR denote the Modified Point SOR Algorithm as coded for symmetric block tridiagonal matrices for which M_{ii} is a tridiagonal matrix and both $M_{i,i-1}$ and $M_{i,i+1}$ are diagonal matrices.

The first experiment is a general comparison of the four methods applied to a sample of each type of problem. The results are summarized in Tables 1, 2 and 3. (The number of iterations of BSORF, BSORP and BSORS is the same for each w.)

Table 1. Data: JB matrix, JB q-vector

n	BSORF ω_{exp}	BSORF iter	BSORF sec	BSORP sec	BSORS ω'_{exp}	BSORS sec	PSOR ω_{exp}	PSOR iter	PSOR sec
15	1.30	18	0.133	0.183	1.3	1.797	1.58	43	0.282
31	1.54	37	0.881	2.529	1.3	20.517	1.76	87	2.296
63	1.74	78	7.388	33.862	1.3	182.291	1.88	179	20.616

Table 2. Data: JB matrix, random q-vector, n = 16

α	β	BSORF ω_{exp}	BSORF iter	BSORF sec	BSORP sec	BSORS ω'_{exp}	BSORS sec	PSOR ω_{exp}	PSOR iter	PSOR sec
4	8	1.12	15	0.183	0.216	1.2	2.013	1.36	26	0.183
8	16	1.24	18	0.249	0.266	1.2	2.995	1.52	36	0.249
16	32	1.22	18	0.216	0.299	1.2	2.961	1.56	39	0.266

Table 3: Data: LAP matrix, random q-vector, n=16

α	β	BSORF ω_{exp}	BSORF iter	BSORF sec	BSORP sec	BSORS ω'_{exp}	BSORS sec	PSOR ω_{exp}	PSOR iter	PSOR sec
4	8	1.34	22	0.316	0.316	1.1	2.329	1.46	31	0.216
8	16	1.50	33	0.332	0.482	1.1	4.143	1.62	43	0.299
16	32	1.32	21	0.282	0.332	1.1	2.579	1.46	33	0.232

One notices that BSORF is almost always uniformly faster and BSORS uniformly slower than the others. Further comparison seems to be very dependent on the sign configuration of the q-vector. From the results of Section 3.5, we may deduce that the sign configuration of the q-vector used in Table 1 is that the first $n(n-1)/2$ entries are negative, the next n are zero (or negative if n is even) and the remainder are positive. Here, we see a pronounced ordering of convergence speed (as measured in seconds), especially as n increases. From fastest to slowest, it is BSORF, PSOR, BSORP, and BSORS. In contrast, the q-vectors used in Tables 2 and 3 have a large number of reversals in their sign configurations. Furthermore, a significantly larger fraction of the z-variables are positive in the experiments of Tables 2 and 3 than in Table 1. These two characteristics tend to be levelling effects, i.e., the running times of BSORF, BSORP and PSOR are nearly equal (as well as we can tell in light of the systematic error involved in measuring execution time in the multi-programming environment of the IBM 360/91).

The second experiment dealt specifically with the hypothesis that when the number of positive components of the solution vector was small, then BSORF was considerably faster than PSOR and that as the number of positive components increased, the running times became equal. A LAP matrix was used with m and n equal to 30. A sequence of constant vectors q^t were used in which the first $30t$ components were -3 and the remaining $900-30t$ components were + 1. The results, summarized in Table 4, support the hypothesis. Since the number of positive components of the solution vector is at least as large as the

number of negative entries in the q-vector (see [9]), this experiment may serve as a guideline in the choice of an algorithm for a specific problem.

Table 4. Data: LAP matrix, random q-vector, n = 30

	BSORF			PSOR			Ratio PSOR/BSORF		No. pos.
t	ω_{exp}	iter	sec	ω_{exp}	iter	sec	iter	sec	z-compon.
1	1.08	7	0.099	1.20	19	0.449	2.714	4.535	60
2	1.26	14	0.216	1.40	32	0.732	2.286	3.389	118
3	1.40	20	0.349	1.50	42	0.998	2.100	2.860	174
6	1.58	36	0.765	1.68	60	1.431	1.667	1.871	346
9	1.66	50	1.148	1.76	79	1.880	1.580	1.638	480
12	1.72	60	1.580	1.78	89	2.113	1.483	1.337	610
30	1.74	97	2.995	1.82	124	2.961	1.278	0.989	900

(In this Table, the relaxation parameter ω_{exp} was determined to within 0.02 for both BSORF and PSOR.)

The third experiment attempts to relate the solution time to n. From Table 1, we find that a growth rate of order 3/2 holds between the order of the matrix (i.e., n^2) and the solution time for BSORF (i.e., $t \propto (n^2)^{3/2}$). Doubling n increases the running time of BSORF, BSORP, BSORS and SOR by a factor of about 8, 13.5, 9 and 7.5, respectively. The results of further testing with random q-vectors are summarized in Tables 5 and 6. These approximately support the factors determined from Table 5.

Table 5. Data: JB matrix, random q-vector

n	a	β	ω_{exp}	BSORF iter	sec
16	4	8	1.12	15	0.183
32	4	8	1.14	18	1.181
16	8	16	1.24	18	0.249
32	8	16	1.36	32	1.896
16	16	32	1.22	18	0.216
32	16	32	1.50	39	1.747

Table 6. Data: LAP matrix, random q-vector

n	a	β	ω_{exp}	BSORF iter	sec
16	4	8	1.34	22	0.316
32	4	8	1.36	33	1.880
16	8	16	1.50	33	0.332
32	8	16	1.62	48	2.046
16	16	32	1.32	21	0.282
32	16	32	1.72	67	2.346

The fourth experiment demonstrates the sensitivity of the Modified Block SOR Algorithm to the relaxation parameter ω. The test problems used LAP matrices of order 1024 and random q-vectors. Since the number of iterations is identical for BSORF, BSORP and BSORS, we present the results only for BSORF. Summarized in Tables 7, 8 and 9, this experiment indicates that the convergence is fairly robust, e.g., if ω_{exp} is the optimal value, then we still achieve good convergence rates for $\omega \in [\omega_{exp} - .2, \omega_{exp} + .2]$.

Table 7 . Data: LAP matrix, random q-vector, n = 32, α = 4, β = 8

BSORF					
ω	iter	sec	ω	iter	sec
1.10	59	3.011	1.40	34	1.713
1.20	47	2.396	1.50	39	1.980
1.30	37	1.836	1.60	47	2.396
1.32	35	1.880	1.70	59	3.011
1.34	33	1.730	1.80	85	4.309
1.36	33	1.880	1.90	153	7.870
1.38	33	1.697			

Table 8. Data: LAP matrix, random q-vector, n = 32, a = 8, β = 16

BSORF

ω	iter	sec	ω	iter	sec
1.10	>200	--	1.62	48	2.046
1.20	175	7.288	1.64	51	2.063
1.30	140	5.807	1.70	61	2.529
1.40	109	4.459	1.80	87	3.577
1.50	81	3.377	1.90	163	6.739
1.60	53	2.163			

Table 9. Data: LAP matrix, random q-vector, n = 32, a = 16, β = 32

BSORF

ω	iter	sec	ω	iter	sec
1.40	>200	--	1.70	72	2.612
1.50	174	5.973	1.72	67	2.346
1.60	124	4.176	1.74	71	2.612
1.64	105	3.560	1.76	77	2.728
1.66	95	3.178	1.80	95	3.394
1.68	85	2.928	1.90	175	5.923

The fifth experiment measures how much of the total solution time is used by Step 1 alone. The results, reported in Table 10, indicate that the subproblems use nearly one-third to one-half of the total time. The times reported are somewhat inaccurate due to the resolution of the timer (16 milliseconds). Despite this, the results emphasize the importance of having a very efficient linear complementarity algorithm for use in Step 1. Further investigation along these lines

might study the dependence of the solution time on the partitioning of the matrix, i.e., on the values of $n_1, n_2, \ldots, n_m$.

Table 10. Subproblem Solution Time vs. Total Solution Time

Matrix	n	a	β	Subproblem Time (sec)	Total Time (sec)
JB	31	nonrandom		0.688	1.999
J-B	63	nonrandom		4.304	13.369
JB	16	4	8	0.208	0.448
JB	16	8	16	0.176	0.416
JB	16	16	32	0.304	0.644
LAP	32	4	8	1.409	3.178
LAP	32	8	16	1.664	3.807
LAP	32	16	32	1.792	4.808

The sixth experiment studies the possibility of accelerating the convergence by varying the value of the relaxation parameter during the progress of the algorithm. It is sometimes profitable when solving systems of linear equations by overrelaxation methods to let $z^0 = 0$, $\omega^1 = 1$ and $\omega^k = \bar{\omega}$ for some fixed $\bar{\omega}$ and all $k \geq 2$. The intended effect of this procedure is to reduce the variation in the components of z^1 which would result if ω^1 were given a value greater than unity. The overrelaxation technique then proceeds with some appropriately chosen value* of the relaxation parameter. In applying this scheme to Algorithm IV, we repeat the experiments reported in Tables 1-3 and 5-9 and set

*theoretically or empirically based

$\bar{\omega}$ equal to the ω_{exp} determined in those experiments. As a point of interest, we also determine the value $\bar{\omega} = \hat{\omega}_{exp}$ that minimizes the number of iterations necessary to satisfy the convergence criterion when using $\omega^1 = 1$. The results are summarized in Table 11. They indicate that the scheme has a minor effect, if any. When there is a change, it is usually a variation of one more iteration than in the preceding experiments. (However, one test showed a decrease of one iteration.)

The eighth, and last, experiment studies another approach to solving the linear complementarity problem (q,M). In Section 2.6, we indicated that when M is a Minkowski matrix, then the solution to (q,M) is the unique vector minimum of the polyhedral set $\{z: Mz + q \geq 0,\ z \geq 0\}$. It is thus a simple exercise to show that the problem (q,M) is equivalent to the linear programming problem

$$\begin{aligned} &\text{Minimize } c^T z \\ &\text{subject to } Mz \geq -q \\ &\qquad\qquad z \geq 0 \ . \end{aligned}$$

for any strictly positive vector c. Letting c be a vector of ones, we solved the linear program with a production code LPM1 [41] written at the Systems Optimization Laboratory at Stanford University. The data was a JB matrix of order 225 and the q-vector corresponded to the journal bearing problem. The LPM1 code took 4.93 seconds with most of the time spent in the Phase I procedure. (Recall that BSORF took .133 seconds to solve (q,M).)

Table 11. Varying the Relaxation Parameter Using BSORF

Matrix	α	β	n	$\omega^1 = \omega^k = \omega_{exp}$ ω_{exp}	iter	$\omega = 1$ $\omega^k = \omega_{exp}$, $k \geq 2$ iter	$\omega = 1$ $\omega^k = \hat{\omega}_{exp}$, $k \geq 2$ $\hat{\omega}_{exp}$	iter
JB	*	*	15	1.30	18	19	1.30	19
	*	*	31	1.54	37	38	1.54	38
	*	*	63	1.74	78	79	1.74	79
	4	8	16	1.12	15	15	1.12	15
	8	16	16	1.24	18	19	1.24	19
	16	32	16	1.22	18	21	1.26	17
	4	8	32	1.14	18	18	1.14	18
	8	16	32	1.36	32	33	1.38	32
	16	32	32	1.50	39	40	1.50	40
LAP	4	8	16	1.34	22	23	1.34	23
	8	16	16	1.50	33	34	1.50	34
	16	32	16	1.32	21	21	1.32	21
	4	8	32	1.36	33	32	1.36	32
	8	16	32	1.62	48	48	1.62	48
	16	32	32	1.72	67	67	1.72	67

"*" indicates that the q-vector corresponds to journal bearing data.

Next, we solved the dual problem using the same data. In this case, the zero vector was a initially feasible point, thus no Phase I was necessary. The solution time for LPM1 solving the dual problem was 4.09 seconds.

Since the matrix is block tridiagonal, it can be partitioned so that the nonzero entries exhibit a "staircase" structure. Under this partitioning, the corresponding linear programming problem was solved by the Ho-Manne nested decomposition algorithm [35], an algorithm especially developed for problems with this structure. The running time was 11.46 seconds. In all cases, the numerical accuracy was comparable.

Further experimentation might investigate whether a reordering of variables might reduce solution time. One possibility is the so-called 'checkerboard" ordering.* Forsythe and Wasow [24, p.259] have reported, however, that the (unpublished) work of M. R. Powers has indicated the convergence of the SOR method for linear equations may not be very sensitive to various orderings.

9.4. Choice of the Relaxation Parameter ω.

The problem of determining a 'theoretically optimal' value of ω for the PSOR algorithm applied to tridiagonal Minkowski matrices is discussed in Section 1.3. The setting was the application of over-relaxation to systems of linear equations. We now review and extend the key notation and results.

*Also known as the "black-white" or "odd-even parity" ordering. See [24, p. 245].

Most generally, an algorithm may be expressed in the operator form $z^{k+1} = \mathscr{L} z^k$ where $z^{k+1}, z^k \in R^{n \times n}$. If we let z^* represent the solution and $e^k = z^* - z^k$ be the error vector, then $z^* = \mathscr{L} z^*$ and $e^{k+1} = \mathscr{L} e^k$. We will let $\|\cdot\|$ be any vector norm or its induced matrix norm with usage dictating its meaning.

Since $\lim_{k \to \infty} (\|e^k\|/\|e^0\|)^{1/k} < \rho(\mathscr{L})$,[58] where $\rho(\mathscr{L})$ is the spectral radius of $\mathscr{L}$ and $e^0 = z^* - z^0$ is the initial error vector, we want to minimize $\rho(\mathscr{L})$. In the specific case where we are applying successive overrelation to the linear system $Mz + q = 0$, the operator $\mathscr{L}$ is formed as follows. We write $M = D - E - F$ where D and $(E + F)$ is a regular splitting of M (see [58, p. 88]) and let $L = D^{-1}E$, $U = D^{-1}F$. (The splitting used depends on whether we are doing point or block SOR. Since the operator $\mathscr{L}$ is dependent on ω and M, we express it as

$$\mathscr{L}_\omega(M) = (I - \omega L)^{-1} [\omega U + (1-\omega)L] .$$

When M is a tridiagonal or block tridiagonal Minkowski matrix, it belongs to the class of consistently ordered 2-cyclic matrices [58, pp. 99-101]. Consequently, the relaxation parameter ω_b that minimizes $\rho(\mathscr{L}_\omega)$ can be uniquely specified in terms of $B = L + U$, the Jacobi matrix associated with M. From a formula of Young [62, p. 169], the optimal parameter value is

$$\omega_b(M) = 2/(1 + \sqrt{1 - \rho^2(B)}\ ).$$

As discussed in Section 1.3, one approach to theoretically determining an "optimal" value of ω for Algorithm IV is to imitate the procedure for systems of linear equations outlined above. This is a plausible idea since if z^* solves (q,M), then for large enough values of k, $(Mz^k + q)_i$ will equal zero when z_i^* is positive. In other words, after a number of iterations, Algorithm IV will appear to act as a block SOR algorithm solving a subsystem of linear equations extracted from the original problem. More specifically, let $N = \sum_{i=1}^{m} n_i$, $\mathcal{I} = \{1, 2, \ldots, N\}$, T be an index set from $\mathcal{I}$ and T' be its complement. Also let M_{TT} be the principal submatrix of M corresponding to rows and columns $j \in T$ and let $\mathcal{I}(z) = \{i \in \mathcal{I} : z_i > 0\}$. The results of Cryer [20] are easily generalized to form the basis of the conjecture that the optimal ω for Algorithm IV is $\omega_{opt} = \omega_b(M_{TT})$ where z^* solves (q,M), $w^* = Mz^* + q$ and $T = \mathcal{I} - \mathcal{I}(w^*)$.

In the case where M is a tridiagonal Minkowski matrix, the theoretical estimate of ω_{opt} is not supported very well by experimental evidence [20]. However, when M is a block tridiagonal matrix, the correlation between theory and practice improves considerably. In order to demonstrate this, we first need to develop some technical machinery.

Recall that the expression for ω_b requires the evaluation of $\rho(B)$. In general, this is difficult to do theoretically. If the matrix is symmetric, an approximation may be obtained by setting

$$\rho(B) = \min\{\|B\|_\infty, \|B^T\|_\infty\} = \min\{\max_i \sum_j |b_{ij}|, \max_j \sum_i |b_{ij}|\}$$

or by a variety of iterative methods (e.g., the power method [36, p. 147]). In the special case of the LAP matrix, we can state o(B) explicitly. We deal first with the Jacobi matrix B arising in the PSOR algorithm, Algorithm III. We decompose M = D - E - F into a diagonal matrix D and strictly lower and upper triangular matrices E and F. The matrix M is partitioned the usual way into submatrices M_{ij} for i, j = 1, 2, ... , m so that M_{ii} is n × n. We will next determine $\rho(B_{KK})$ where K = (1, 2, . . . , kn} for any k = 1, 2, . . . , m. (Note that B_{KK} is the Jacobi matrix associated with the LAP matrix M_{KK}.)

Theorem 6. Let B_{KK} be the matrix described above. Then $\rho(B_{KK}) = \frac{1}{2}(\cos \pi/(k+1) + \cos \pi/(n+1))$.

Proof. Define the s X s matrix TS = (tij) by $t_{12} = 1$, $t_{i,i+1} = t_{i+1,i} = 1$ for i = 2, 3, . . . , s-l, $t_{s,s-1} = 1$, and $t_{ij} = 0$ otherwise. Let I_S be an s × s identity matrix. Recall that if G and H are $s_1 \times s_2$ and $s_3 \times s_4$ matrices, then their tensor product (or Kronecker product [32, pp. 97-98]) P = G ⊗ H is an $s_1 s_3 \times s_2 s_4$ matrix of the form:

$$P = \begin{vmatrix} g_{11}^H & g_{12}^H & \cdots & g_{1s_2}^H \\ \cdot & \cdot & & \\ g_{s_1 1}^H & g_{s_1 2}^H & \cdots & g_{s_1 s_2}^H \end{vmatrix}$$

It is easy to show that $\hat{B}_K \equiv 4B_K = T_K \otimes I_N + I_K \otimes T_N$. Let Q_S be the orthogonal s x s matrix whose column vectors are the eigenvectors of T_S and let L_S be the diagonal matrix of eigenvalues; thus $T_S Q = Q_S L_S$. The matrix $Q = Q_K \otimes Q_N$ is orthogonal since ?K and Q_N are, hence $Q^T \hat{B}_K Q$ has the same eigenvalues as $\hat{B}_K$. (Note that we have suppressed the explicit dependence of Q on K and N.) Using the fact that $(G_1 \otimes G_2)(G_3 \otimes G_4) = (G_1 G_3) \otimes (G_2 G_4)$ for any matrices, $G_1, \ldots, G_4$ one can show that $Q^T \hat{B}_K Q = L_K \otimes I_N + I_K \otimes L_N$. But this is a KN x KN diagonal matrix with entries $\lambda_{ki} + \lambda_{nj}$ where λ_{k1}, i = 1, 2, ... , k and A_{nj}, j = 1, 2, ... , n are the diagonal entries of L_K and L_N, respectively. From [33, p. 154], we know that $\lambda_{rj} = 2 \cos \pi j/(r+1)$ for j = 1, 2, . . . , r. Thus the spectral radius of $\hat{B}_K$ is

$$\rho(\hat{B}_K) = \max_{\substack{1 \le i \le k \\ 1 \le j \le n}} |\lambda_{ki} + \lambda_{nj}| = 2 \left(\cos \frac{\pi}{k+1} + \cos \frac{\pi}{n+1} \right)$$

and since $\hat{B}_K = 4B_K$, $\rho(B_K) = \frac{1}{2} (\cos \pi/(k+1) + \cos \pi/(n+1))$. ∎

We now study the block Jacobi matrix associated with Algorithm IV. Let M = D - E - F where, again, M is a LAP matrix and

$$-L = \begin{vmatrix} 0 & & & & \\ M_{21} & 0 & & & \\ & M_{32} & 0 & & \\ & & \ddots & \ddots & \\ & & & M_{m,m-1} & 0 \end{vmatrix}, \quad D = \begin{bmatrix} M_{11} & & & \\ & M_{22} & & \\ & & \ddots & \\ & & & M_{mm} \end{bmatrix}$$

and

$$-U = \begin{vmatrix} 0 & M_{21} & & & \\ & 0 & M_{32} & & \\ & & 0 & \ddots & \\ & & & \ddots & M_{m-1,m} \\ & & & & 0 \end{vmatrix}$$

Let $B = -D^{-1}(L + U)$ and B_{KK} be the leading principal $kn \times kn$ submatrix. The application of Young's formula for ω_b to block SOR algorithms refer to B_{KK}, the Jacobi matrix associated with M_{KK}. The next result gives the spectral radius of these submatrices of B.

Theorem 7. Let B_{KK} be the matrix described above. Then $\rho(B_{KK}) = \cos(\pi/(k+1))/(2 - \cos(\pi/(n+1))$.

Proof. Define the $s \times s$ matrix $U_S = (u_{ij})$ by $u_{ii} = 4$ for $i = 1, 2, \ldots, s$, $u_{12} = -1$, $u_{i,i+1} = u_{i,i-1} = -1$, $u_{s,s-1} = -1$ and $u_{ij} = 0$ otherwise. Let $V_S = U_S^{-1}$ and I_S be an $s \times s$ identity matrix. Also let $T_S = 4I_S - U_S$. Finally, let P_S (resp., Q_S) be the orthogonal $s \times s$ matrix whose column vectors are the eigenvectors

of T_S (resp., V_S) and let L_S (resp., C_S) be the diagonal matrix of eigenvalues. Thus, the matrix $Q = Q_K \otimes P_N$ is orthogonal. (Again, we have suppressed the dependence of Q on K and N.)

Notice that $B_K = T_K \otimes V_N$. Since B_K and $Q^T B_K Q$ have the same eigenvalues, we can instead determine those of the latter matrix. But $Q^T B_K Q = L_K \otimes C_N$, a diagonal matrix. Hence, the eigenvalues of B_K are all possible products of the diagonal entries of L_K and C_N, say $\{\lambda_{ki} c_{nj}\}$ where i = 1, 2, . . . , k and j = 1, 2, . . . , n. As in Theorem 6, $\lambda_{kj} = 2 \cos \pi j/(k+1)$ for j = 1, 2, . . . , k. Furthermore, $\{c_{nj}\}$ are the reciprocals of the eigenvalues of U_N, therefore [33, p. 154], $c_{nj} = 1/(4 - 2 \cos \pi j/(n+1))$ for j = 1, 2, . . . , n. It then follows that

$$\rho(B_K) = \max_{\substack{1 \le i \le k \\ 1 \le j \le n}} \cos(\pi i/(k+1))/(2 - \cos(\pi j/(n+1)))$$

$$= \cos(\pi/(k+1))/(2 - \cos(\pi/(n+1))). \quad \blacksquare$$

There are two problems in applying Theorem 6 or 7 to determine ω_{opt}. The theorems both presuppose that one knows, a priori, the index set $T = \mathcal{I}(z^*)$ since T determines the linear subsystem $M_{TT} z_T + q_T = 0$ which is eventually solved. Furthermore, they both assume that T = {1, 2, . . . , kn) for some $1 \le k \le m$. (The theorems remain true if K = {t+1, t+2, . . . , t+kn} for t = 0, n, 2n,...,(m-1)n and k = 1, 2, . . . , m.) From the Perron-Frobenius theory of nonnegative matrices, if T = (1, 2, . . . , kn, kn+1, . . . , kn+c} where $0 \le k \le n$, then we can bound $\omega_b(M_{TT})$ between $\omega_b(M_{T_1T_1})$ and $\omega_m(M_{T_2T_2})$

where $T_1 = \{1, 2, \ldots, kn\}$ and $T_2 = \{1, 2, \ldots, (k+1)n\}$. This fact mitigates the second problem somewhat although it still leaves the problem of determining T_1 and T_2. One might be able to determine T_1 and T_2 during the progress of Algorithm III or IV by monitoring $\mathcal{I}(z^k)$ until it appears to stabilize at some index set. From [20], we know that gross bounds for ω_{opt} can be obtained by setting $T_1 = \mathcal{I}(-q)$ and $T_2 = \mathcal{I}$. Our computational experience has shown these latter bounds are not very useful unless T_1 is a fairly large index set in which case $\omega_b(M_{T_1T_1})$ and $\omega_b(M_{T_2T_2})$ are fairly close together.

In order to illustrate the use of Theorems 6 and 7, we use the data of the second experiment reported in Section 9.3. We let $T_1 = \{1, 2, \ldots, 30t\}$ and $T_2 = \{1, 2, \ldots, 30(k+1)\}$ where $\{1, 2, \ldots 30k\} \not\subseteq \mathcal{I}(z^*) \subseteq T_2$. The results are summarized in Table 12. For the PSOR algorithm, $\omega_1 = \omega_b(M_{T_1T_1})$ and $\omega_2 = \omega_b(M_{T_2T_2})$ is determined via Theorem 6. For the BSORF algorithm, these quantities are calculated using the results of Theorem 7. In both cases, ω_{exp} was determined (within 0.02) to be the empirically optimal value and can be seen to be remarkably close to ω_2 in most cases. (We would expect $\omega_1 \leq \omega_{exp} \leq \omega_2$.) These results suggest that an adaptive mechanism which sets $\omega^1 = \omega_1$ and changes ω during the operation of the algorithms could prove very worthwhile.

Table 12. Theoretical Rounds for the Relaxation Parameters

		FSOR			BSORF		
t	k+1	ω_1	ω_2	ω_{exp}	ω_1	ω_2	ω_{exp}
1	2	1.07	1.20	1.20	1.00	1.07	1.08
2	4	1.20	1.40	1.40	1.07	1.26	1.26
3	6	1.31	1.52	1.50	1.17	1.39	1.40
6	12	1.52	1.69	1.68	1.39	1.59	1.58
9	17	1.63	1.75	1.76	1.51	1.67	1.66
12	22	1.69	1.79	1.78	1.59	1.71	1.72
30	30	1.82	1.82	1.82	1.75	1.75	1.74

An alternate approach for estimating the optimal relaxation parameter is suggested by some research of Garabedian [27]. In a study of the point SOR method applied to linear systems derived from finite difference approximations to partial differential equations, he proposed an asymptotically good estimate for ω_b (i.e., the estimate became better as the mesh size on the region R of interest tended to zero). He assumed that the mesh size was uniform and of width h and the area of the closure of R was a. Garabedian then suggested using a relaxation parameter $\omega = 2/(1 + 3.015(h^2/a)^{1/2})$. For many shapes of regions, he noted that in several numerical tests carried out by Young, this choice of ω resulted in approximately a 20 percent decrease in convergence rate from the optimal convergence rate. The remarkable success of this estimate lies in the simplicity of its application in comparison with the application of Young's formula. This suggests that a generalization to the block SOR method (and thence to

Algorithm IV) could be worthwhile as future research. The authors have not yet derived similar results for either the block SOR or Modified Block SOR techniques.

REFERENCES

[1] R. J. Arms, L. D. Gates and B. Zondek, "A Method of Block Iteration," J. SIAM 4 (1956),220-229.

[2] A. Auslender, "Une Méthode Générale Pour la Decomposition et la Minimisation de Fonctions Non Différentiables," C.R. Acad. Sci. Paris 271 (1970), A1078-A1081.

[3] A. Auslender and B. Martinet, "Méthodes de Decomposition pour la Minimisation d'une Fonctionelle sur un Espace Produit," C.R. Acad. Sci. Paris 274 (1972), A632-A635.

[4] E. F. Beckenbach and R. Bellman, Inequalities, Springer-Verlag, New York, 1965.

[5] G. V. Berg, Private Communication (Letter to R. W. Cottle, February 5,1968.)

[6] A. Cameron and W. L. Wood, "The Full Journal Bearing," Inst. Mech. Engrs. J. Proc.161 (1949),59-64.

[7] V. Castelli and V. Shapiro,"Improved Method for Numerical Solutions of the General Incompressible Fluid Film Problem," J. of Lubr. Technology, Trans. of the ASME 89 (1967),211-218.

[8] J. Cea, "Recherche Numerique d'un Optimum dans un Espace Produit," Colloquium on Methods of Optimization, Lecture Notes in Mathematics, vol. 112, Springer-Verlag, New York, 1970.

[9] R. Chandrasekaran, "A Special Case of the Complementary Pivot Problem," Opsearch 7 (1970),263-268.

[10] D. G. Christopherson, 'A New Mathematical Method for the Solution of Film Lubrication Problems," Inst. Mech. Engrs. J. Proc. 146 (1941), 126-135.

[11] R. W. Cottle, "Nonlinear Programs with Positively Bounded Jacobians," Doctoral Dissertation, University of California, Berkeley, 1964.

[12] R. W. Cottle, 'Nonlinear Programs with Positively Bounded Jacobians," J. SIAM Appl. Math. 14 (1966), 147-158.

[13] R. W. Cottle, "Principal Pivot Method of Quadratic Programming," Mathematics of the Decision Sciences, Part I (G. B. Dantzig and A. F. Veinott, eds.), American Mathematical Society, Providence, R.I., 1968.

[14] R. W. Cottle, "Monotone Solutions of the Linear Complementarity Problem," Math. Prog. 3 (1972), 210-224.

[15] R. W. Cottle, "Complementarity and Variational Problems," to appear in Proceedings of the Conference on Mathematical Programming and its Applications, Rome, 1974.

[16] R. W. Cottle and G. B. Dantzig, "Complementary Pivot Theory of Mathematical Programming," Linear Algebra and Appl. 1 (1968), 103-125.

[17] R. W. Cottle and A. F. Veinott, Jr., "Polyhedral Sets Having a Least Element," Mathematical Programming 3 (1972), 238-249.

[18] c. W. Cryer, "On the Approximate Solution of Free Boundary Problems Using Finite Differences," J. Assoc. Comput. Mech. 17 (1970), 397-410.

[19] C. W. Cryer, "The Method of Christopherson for Solving Free Boundary Problems for Infinite Journal Bearings by Means of Finite Differences," Math. of Computation 25 (1971), 435-443.

[20] C. W. Cryer, "The Solution of a Quadratic Programming Problem Using Systematic Overrelaxation," J. SIAM Control 9 (1971), 385-392.

[21] G. B. Dantzig, Linear Programming and Extensions, Princeton University Press, Princeton, 1963.

[22] G. B. Dantzig and R. W. Cottle, "Positive Semi-Definite Programming," Nonlinear Programming (J. Abadie, ed.), North Holland Publishing Company, Amsterdam, 1967.

[23] M. Fiedler and V. Ptak, "On Matrices with Nonpositive Off-Diagonal Elements and Positive Principal Minors," Czech. J. Math. 12 (1962), 382-400.

[24] G. E. Forsythe and W. R. Wasow, Finite-Difference Methods for Partial Differential Equations, J. Wiley and Sons, New York, 1960.

[25] S. P. Frankel, "Convergence Rates of Iterative Treatments of Partial Differential Equations," Math. Tables Aids Comput. 4 (1950), 65-75.

[26] V. M. Friedman and V. S. Chernina, "An Iteration Process for the Solution of the Finite-Dimensional Contact Problem," U.S.S.R. Comp. Math. and Math. Phys. 8 (1967), 210-214.

[27] P. R. Garabedian, "Estimation of the Relaxation Factor for Small Mesh Size," Math. Tables Aids Comput. 10 (1956), 183-185.

[28] C. L. Gerling, Die Ansgleichs-Rechnungen der Practischen Geometrie oder die Methode der Kleinsten Quadrate mit ihren Anwendungen für Geodätiche Aufgaben, Gothe. Hamburg, 1843.

[29] A. A. Gnanadoss and M. R. Osborne, "The Numerical Solution of Reynolds' Equation for a Journal Bearing," Quart. J. Mech. Appl. Math. 17 (1964), 241-246.

[30] R. Glowinski, "Sur la Minimisation, par Surrelaxation avec Projection, 'de Fonctionnelles Quadratiquesdans les Espaces de Hilbert," C.R. Acad. Sci. Paris 276 (1973), A1421-A1423.

[31] M. Hall and M. Newman, "Copositive and Completely Copositive Quadratic Forms," Proc. Cambridge Phil. Soc. 59 (1963), 329-339.

[32] P. Halmos, Finite-Dimensional Vector Spaces, D. Van Nostrand co., Inc., Princeton, 1958.

[33] S. J. Hammarling, Latent Roots and Latent Vectors, Univ. Toronto Press, Toronto, 1970.

[34] C. Hildreth, "A Quadratic Programming Procedure," Nav. Res. Logis. Quart. 4 (1957), 79-85.

[35] J. K. Ho and A. S. Manne, "Nested Decomposition for Dynamic Models," to appear in Mathematical Programming, 1974.

[36] E. Isaacson and H. B. Keller, Analysis of Numerical Methods, J. Wiley and Sons, New York, 1966.

[37] E. L. Keller, "The General Quadratic Optimization Problem," Mathematical Programming 5 (1973), 331-337.

[38] c. E. Lemke, "On Complementary Pivot Theory," Mathematics of the Decision Sciences, Part I (G. B. Dantzig and A. F. Veinott, Jr., eds.), American Mathematical Society, Providence, R.I., 1968.

[39] C. E. Lemke, "Recent Results on Complementarity Problems," Nonlinear Programming (J. B. Rosen, 0. L. Mangasarian, and K. Ritter, eds.), Academic Press, New York, 1970.

[40] C. E. Lemke and J. T. Howson, "Equilibrium Points of Bimatrix Games," J. Soc. Indust. Appl. Math. 12 (1964), 413-442.

[41] LPM1 User's Manual, System Optimization Laboratory, Department of Operations Research, Stanford University, Stanford, 1973.

[42] B. Martinet, "Convergence de Certaines Méthodes de Relaxation en Programmation Convexe," C.R. Acad. Sci. Paris 265 (1967), A210-A212.

[43] J.-C. Miellou, "Méthodes de Jacobi, Gauss-Seidel, Sur (Sous)-relaxation par Blocs, Appliquées a une Classe de Problkmes Non Lineaires," C.R. Acad. Sci. Paris 273 (1971), A1257-A1260.

[44] K. G. Murty, "On the Number of Solutions to the Complementarity Problems and Spanning Properties of Complementary Cones," Lin. Alg. and Appl. 5 (1972), 65-108.

[45] O. Pinkus and B. Sternlicht, Theory of Hydrodynamic Lubrication, McGraw-Hill, New York, 1961.

[46] A. A. Raimondi and J. Boyd, "A Solution for the Finite Journal Bearing and Its Application to Analysis and Design: III," Trans. Amer. Soc. Lubrication Engrs. 1 (1958), 194-209.

[47] A. A. Raimondi, Private Communication, August 13, 1973.

[48] O. Reynolds, "On the Theory of-lubrication and its Application to Mr. Beauchamp Tower's Experiments: Phil. Trans. Roy. Soc., London 177 (1886).

[49] H. L. Royden, Real Analysis, 2nd edition, Macmillan Co., London, 1968.

[50] R. Saigal, "A Note on a Special Linear Complementarity Problem," Opsearch 7 (1970), 175-183.

[51] R. Saigal, "Lemke's Algorithm and A Special Linear Complementarity Problem," Opsearch 8 (1971), 201-208.

[52] H. Samelson, R. M. Thrall, and O. Wesler, "A Partition Theorem for Euclidean n-Space," Proc. Amer. Math. Soc. 9 (1958), 805-807.

[53] S. Schechter, "Relaxation Methods for Convex Problems," SIAM J. Numer. Anal. 5 (1968), 601-612.

[54] S. Schechter, "Iteration Methods for Nonlinear Problems," Trans. Amer. Math. Soc. 104 (1962), 179-189.

[55] H. W. Swift, "The Stability of Lubricating Films in Journal Bearings," Proc. Inst. of Civ. Eng. 233 (1932), 267-323.

[56] A. W. Tucker, "A Combinatorial Equivalence of Matrices," Proceedings of Symposia in Applied Mathematics, Vol. 10 (R. Bellman and M. Hall, eds.), American Mathematical Society, Providence, R.I., 1960.

[57] A. W. Tucker, "Principal Pivotal Transforms of Square Matrices," SIAM Review 5 (1963), p. 305.

[58] R. S. Varga, Matrix Iterative Analysis, Prentice Hall, Englewood Cliffs, N.J., 1962.

[59] J. R. Westlake, A Handbook of Numerical Matrix Inversion and Solution of Linear Equations, J. Wiley and Sons, New York, 1968.

[60] D. F. Wilcock and E. R. Booser, Hearing Design and Application, McGraw-Hill, New York, 1957.

[61] J. H. Wilkinson, "Error Analysis of Direct Methods of Matrix Inversion," J. Assoc. Comput. Mach. 8 (1961),281-330.

[62] D. M. Young, Iterative Solution of Large Linear Systems, Academic Press, New York, 1971.

[63] W. I. Zangwill, Nonlinear Programming, A Unified Approach, Prentice-Hall, Inc. Englewood Cliffs, N.J., 1969.