LAKEHEAD

UNIVERSITY

FAST APPROXIMATE INVERSION OF LARGE SPARSE LINEAR SYSTEMS

Paul O. Frederickson

Mathematics Report #7-75

DEPARTMENT OF MATHEMATICAL SCIENCES

FAST APPROXIMATE INVERSION OF LARGE SPARSE LINEAR SYSTEMS

Paul O. Frederickson*

SUMMARY: *The algorithm FAPIN solves large sparse linear systems of a special but important class in* $O(n)$ *operations. In particular, it solves all finite-element or Raleigh-Ritz-Galerkin approximations, over a sufficiently regular mesh, to second order elliptic boundary value problems. Like most iterative algorithms, FAPIN has a highly parallel structure which allows efficient implementation on parallel or vector processors. The main advantage of FAPIN is a rate of convergence independent of* n *for equations in the class considered; this is proved in Theorem 2. Another advantage, stability and insensitivity to perturbations of the system, has been demonstrated by several numerical examples.*

1. APPROXIMATE INVERSION

We suppose we have been given a large sparse linear operator $A : X \rightarrow Y$ and an element $y \in Y$ and have been asked to construct a solution $x \in X$ to the linear system

$$(1) \qquad Ax = y.$$

We are interested in constructing a numerical solution, rather than an

* Research supported by the National Research Council of Canada, through grant number A7334, and by Leibniz Rechenzentrum and the Math. Instutite of the Tech. Univ. München, during the author's sabbatical year.

exact, or rational arithmetic, solution. Thus we will assume we have also been given a tolerance $\varepsilon < 1$ and have as goal the construction of an element $x \in X$ satisfying

(2) $$||y-Ax|| < \varepsilon||y||$$

Any such x will be called an *ε-approximate solution* to (1).

Linear operators $B : Y \to X$ with the property that the operator $I-AB : Y \to Y$ has a spectral radius $\rho < 1$ will play an important role in the theory which follows; we will refer to such an operat~~a~~or B as a *ρ-approximate inverse* to A. In case $\rho \le \varepsilon$ the element $By \in X$ is already an ε-approximate solution to equation (1). Usually, though, ε is smaller than ρ and we must iterate, improving the approximation x^m by computing

(3) $$\begin{aligned} r &\leftarrow y - Ax \\ x &\leftarrow x + Br \end{aligned}$$

$$\begin{aligned} r^m &= y - Ax^m = \\ &= y - A(x^{m-1} + Br^{m-1}) = \\ &= r^{m-1} - ABr^{m-1} = (I - AB)r^{m-1} \end{aligned}$$

It follows directly from (3) that the mth iterate r^m satisfies

(4) $$r^m = (I-AB)^m r^0,$$

and thus a solution to (2) has been constructed after at most M iterations if $\rho^M \le \varepsilon$, or equivalently if $M \ge (-\log \varepsilon)/(-\log \rho)$. (See chapter 3 of either Varga [1] or Young [2] for more detail.)

We will describe one pass of the algorithm FAPIN as one iteration of equation (3) for a suitable operator B. Although the description of B is somewhat complex the cost of one iteration, at 89n arithmetic operations for a general nine-point operator A and 54n for the constant coefficient nine point operator is only about five times that of SOR, for example. This higher cost is justified, for n greater than 2000, by the increased rate of convergence alone.

2. LOCAL OPERATORS

The classical five-point discrete Laplacian is an example of the type of linear operator A that FAPIN is able to invert. A more significant example is a nine-point discrete Laplacian; for the primary restriction on A, which we make precise in this section, amounts to the requirement that the graph of A be contained in the graph of a nine-point Laplacian, or that the matrix representation of A have zeros where the nine point operator does. We will not restrict A to be a constant coefficient operator; in fact, FAPIN is surprisingly insensitive to variations in the coefficients of A. We will, however, require A to be a differential operator. (For the sake of clarity we have restricted our discussion, throughout the paper, to the planar case.)

Denote by Ω a finite subset of the integer lattice in the plane, denote by X the space of real functions on Ω, and let Y be a subspace of X. We say that the linear operator $A : X \to Y$ is a *q-local operator* for some integer q if the value of Ax at a point $i = (i_1, i_2) \in \Omega$ depends only on the values of x in a q-neighborhood of i; more precisely, if

(5) $$[(Ax)_i \neq 0] \Rightarrow [\exists\; j \in \Omega \quad |i-j| \leq q, \text{ and } x_j \neq 0].$$

The norm in equation (5) is the ℓ_∞ or sup-norm, defined for any lattice point $i = (i_1,i_2)$ by $|i| = \max\{|i_1|,|i_2|\}$. Thus a nine-point Laplacian is, in particular, 1-local.

We consider, from now on, only 1-local operators A. We note that the nested disection algorithm of Alan George [3], the best of the exact algorithms, requires the same restriction on A. Fortunately, the usual discretizations of linear second order elliptic boundary value problems in the plane, including Raleigh-Ritz-Galerkin discretizations over either bilinear or linear splines, are included.

Denote by n the number of points in Ω, so that X is isomorphic to R^n. Then any 1-local operator A is represented, through this isomorphism, by a sparse matrix having at most nine nonzero elements in each row. We don't find this usual sort of matrix representation as useful for our computational purposes, however, as a more compressed representation which follows directly from equation (5). Corresponding to every 1-local operator $A : X \rightarrow Y$ there is an array $A_{i,j}$ such that for any point i

(6) $$(Ax)_i = \sum_{|j|\leq 1} A_{i,j} x_{i+j}.$$

The sum extends, of course, over only those j for which $i+j \in \Omega$. Implementation of (6) allows storage of A in $9n$ locations and evaluation of Ax in $9n$ multiplications, both with low overhead. *Moreover, parallel or vector operation is facilitated.*

An inner loop of FAPIN requires an approximate inverse $B : Y \to X$ to A which is itself 1-local; thus we will use representation (6) for B as well as A. Among all 1-local approximate inverses there is an optimum one for our purposes, as one can show using compactness arguments. Their construction is feasible only in the constant coefficient case, however. In general we recommend a sub-optimal approximate inverse referred to in Benson [4] and in Benson and Frederickson [5] as the *DB_q-approximate inverse*. It is constructible, for each $i \in \Omega$, by the linear system of 9 equations

$$\sum_{|j|\leq 1} A_{i+j,k-j}B_{i,j} = \delta_{k,0} \qquad |k| \leq 1, \tag{7}$$

where δ denotes the Kronecker delta, or the identity operator in notation (6).

3. COLLECTION

Local approximate inverses alone are not adequate, however, when n becomes large. The reason is not hard to understand if one observes that a discrete Dirichlet problem of second order in the plane has an exact inverse with the logarithmic character of the plane potential function. No local approximate inverse can come close to this when n is large, which explains the poor convergence rate of the usual iterative methods on such problems. This global character of the exact inverse means that construction of an approximate solution requires the passing of a considerable amount of information from every point of Ω to every other point, a task which can be carried out in $O(n)$ operations *only* if this information is collected, in some way, and distributed in bulk. We now describe how FAPIN does this collection and distribution.

There is a least integer ℓ such that $|i| < 2^{\ell-1}$ for every i ~~X~~ $\in \Omega$; we write Ω^ℓ for Ω and define, using the recurrence

$$(8) \qquad \Omega^{k-1} = \{i \,|\, \exists\; j,\ |j| \le 1,\ 2i+j \in \Omega^k\},$$

the sets Ω^k for $1 \le k \le \ell$. We note that $|i| \le 2^{k-1}$ if $i \in \Omega^k$, and in particular, Ω^1 has at most 9 points. From the programming standpoint the easiest implementation of FAPIN occurs when Ω^ℓ is a square region with exactly $(1+2^\ell)^2$ points. The algorithms of George [3], Schroeder and Trottenburg [6], or Buzbee, Golub and Nielsen [7] are also most easily implemented in this case.

Denote by X^k the linear space of real functions on Ω^k, and define the *collection operator* $p^k : X^k \to X^{k-1}$ by

$$(9) \qquad r_i^{k-1} \leftarrow \sum_j t_j r_{2i+j}^k .$$

Here the coefficients t_j are normalized binomial coefficients

$$(10) \qquad t_j = \frac{1}{4} \binom{2}{j_1+1} \binom{2}{j_2+1}$$

$\binom{n}{j} =$ [illegible] $n!$

$t_{0,-1} = \frac{1}{4} \cdot 2 \cdot 1 = \frac{1}{2}$ $\qquad t_{1,1} = \frac{1}{4} \; \frac{2!}{2!} \cdot \frac{2!}{2!} = \frac{1}{4}$

$t_{0,0} = 1$

$t_{0,1} = \frac{1}{2}$

$t_{-1,0} = \frac{1}{2}$

$t_{1,0} = \frac{1}{2}$

and thus vanish for $|j| > 1$. We then use the same coefficients to define the sequence of *interpolation operators* $Q^k : X^{k-1} \to X^k$ through ?

(11) $$x_i^k \leftarrow \sum_j t_{i-2j} x_j^{k-1} .$$

We define the subspace Y^k of X^k by $Y^k = P^{k+1}(Y^{k+1})$, beginning with $Y^\ell = Y$, and we define the sequence of operators $A^k : X^k \to Y^k$ by

(12) $$A_{i,j}^{k-1} \leftarrow \sum_u \sum_v t_u \, A_{2i+u,\underbrace{2j}_{?}\overset{-}{+}u\overset{+}{-}v}^k \, t_v .$$?

Implementation of P^k requires only $3\times(n/3)$ multiplications and $8\times(n/3)$ additions for all k in the range $1 < k \le \ell$ if (10) is coded as

(13) $$r_{i_1,i_2} \leftarrow r_{2i_1,2i_2} + .5\times(r_{2i_1-1,2i_2} + \ldots)$$

and Q_k requires only $6n/3$ operations if it is well coded. Similarly, construction of all of the operators A^k from the given $A^\ell = A$ requires about 56n operations.

4. BEST APPROXIMATION

Theorem 1. *The operator* A^{k-1} *defined by equation (12) satisfies the identity*

(14) $$A^{k-1} = P^k \, A^k \, Q^k .$$

and is the Raleigh-Ritz-Galerkin best approximation to A^k *in the* *subspace* $U^k = Q^k(x^{k-1})$ *of* X^k.

The proof of equation (14) involves comparing (12) with the expression which results when the right hand side is expanded, using (11), (6), and then (9). The rest is familiar. Theorem 1 can be restated by saying that the ladder diagram

$$\begin{tikzcd}
X^\ell \arrow[d, "A^\ell"] & X^{\ell-1} \arrow[l, "Q^\ell"'] \arrow[d, "A^{\ell-1}"] & \cdots \arrow[l] & X^2 \arrow[d, "A^2"] & X^1 \arrow[l, "Q^2"'] \arrow[d, "A^1"] \\
Y^\ell \arrow[r, "P^\ell"'] & Y^{\ell-1} \arrow[r] & \cdots & Y^2 \arrow[r, "P^2"'] & Y^1
\end{tikzcd}$$

commutes in any square. It follows immediately that the whole diagram commutes, or

<u>Corollary 1.1</u>: *The operator* A^k *is the Raleigh-Ritz-Galerkin best approximation to* A^ℓ *in the subspace* $Q^\ell \ldots Q^{k+1}(X^k)$ *of* X^ℓ.

We make use of Theorem 1 primarily in the form of

<u>Corollary 1.2</u>: *If* $A^{k-1}x^{k-1} = y^{k-1}$ *and* $y^{k-1} = P^k y^k$, *then* $Q^k x^{k-1}$ *is the best approximate solution to* $A^k x^k = y^k$ *in* U^k, *in the sense that* $P^k(y^k - A^k(Q^k x^{k-1})) = 0$.

5. THE ALGORITHM

FAPIN should be viewed as an iterative algorithm. At the beginning of each pass one has an approximation x to the solution to equation (1), which may or may not be zero during the first pass, and one has evaluated the residual vector $r \leftarrow y - Ax$. The pass really begins when one applies (9) repeatedly, creating $r^{\ell-1},\dots,r^1$ from $r^\ell = r$. Next $x^1 = B^1 r^1$ is computed, and then one works back up from $k = 2$ to $k = \ell-1$, first interpolating and then refining this approximation:

$$
\begin{aligned}
&x^k \leftarrow Q^k x^{k-1} \\
&x^k \leftarrow x^k + B^k(r^k - A^k x^k)
\end{aligned}
\tag{15}
$$

At the top level, $k = \ell$, these assignments are replaced by

$$
\begin{aligned}
&x^\ell \leftarrow x^\ell + Q^\ell x^{\ell-1} \\
&x^\ell \leftarrow x^\ell + B^\ell(y - A^\ell x^\ell)
\end{aligned}
\tag{16}
$$

To be more precise, we state this in an ALGOL like notation, the semantics of which should be obvious. Note that the array x^k has dimension different from that of the array x^{k-1}, for example, which means that the actual ALGOL or FORTRAN programs are considerably more complex.

```
FOR it ← 1 STEP 1 UNTIL nit DO
BEGIN
    $r^\ell \leftarrow y^\ell - A^\ell(x^\ell)$;
    FOR k ← $\ell$ STEP -1 UNTIL 2 DO
        $r^{k-1} \leftarrow P^k(r^k)$;
    $x^1 \leftarrow B^1(r^1)$;
    FOR k ← 2 STEP 1 UNTIL $\ell$-1 DO
    BEGIN
        $x^k \leftarrow Q^k(x^{k-1})$;
        $r^k \leftarrow r^k - A^k(x^k)$;
        $x^k \leftarrow x^k + B^k(r^k)$
    END;
```

```
        $x^{\ell} \leftarrow x^{\ell} + Q^{\ell}(x^{\ell-1})$;
        $r^{\ell} \leftarrow y^{\ell} - A^{\ell}(x^{\ell})$;
        $x^{\ell} \leftarrow x^{\ell} + B^{\ell}(r^{\ell})$
    END
```

The actual programs evaluate the norm of r^{ℓ} while evaluating r^{ℓ} in the third line, and this is used to allow an early exit when the tolerance ε has been achieved. The above statement of the algorithm, with this WHILE clause omitted, defines the linear approximate inverse operator B advertised in section 2.

The above program segment must be preceded by two more FOR loops which first collect (or project) the A^k using equation (12) and then use equation (7), or a more nearly optimum algorithm, to define the approximate inverse operators B^k. These steps are considerably simplified in the constant coefficient case, of course. Moreover, they need not be re-executed if only the right hand side y of equation (1) has been modified.

At this point the similarities between FAPIN and the algorithm of Brandt [8] or Fedorenko [9] is apparent. In particular, the idea of interpolating a solution to a coarser problem as a first approximation is common to all three.

7. NUMERICAL EXPERIENCE

The following data, taken from runs made in April 1974 using an ALGOL implementation on the Telefunken TR 400 computer of Leibniz

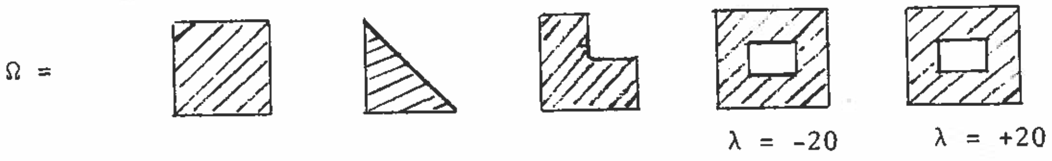
Rechenzentrum in München, are presented as primary evidence that FAPIN has a spectral radius $\rho < 1/2$ when applied to a discrete Poisson problem on a square; Theorem 1 and its proof are included only to make this evidence believable. Here T denotes CPU time per iteration on the TR 440, in seconds, and $n = (2^{\ell}+1)^2$ is the total number of equations.

ℓ	2	3	4	5	6
n	25	81	289	1089	4225
T	.10	.45	2.0	8.3	32
ρ	.35	.44	.47	.43	.48

fig. 1

More detail is given by Broy [10], who has also found modifications of FAPIN with even faster convergence.

We have claimed that FAPIN is not sensitive to perturbations of the operator A or, more generally, to the region Ω. The following data, taken from runs made in the summer of 1975 using FORTRAN implementations on the IBM 360/50 at Lakehead University and the CDC 7600 at Los Alamos demonstrate this independence.



ρ =	.450	.496	.494	.444	.467

fig. 2

In each case Ω was the indicated part of a 33×33, or $\ell = 5$, grid. In the last two cases the equation being solved was $\nabla^2 u + \lambda q(x)u = f(x)$, with $q(x) = 1$ within the inner square and zero elsewhere. In the first three cases $q(x)$ was identically zero within Ω, and zero boundary conditions were used in each case.

An even more striking demonstration is provided by an extension of the last two examples of fig. 2 above to show the behaviour as λ passes the first eigenvalue λ_0 of the discrete $(n = 4225)$ problem. There is no need for A to be positive definite for FAPIN to work, and A can be very nearly singular.

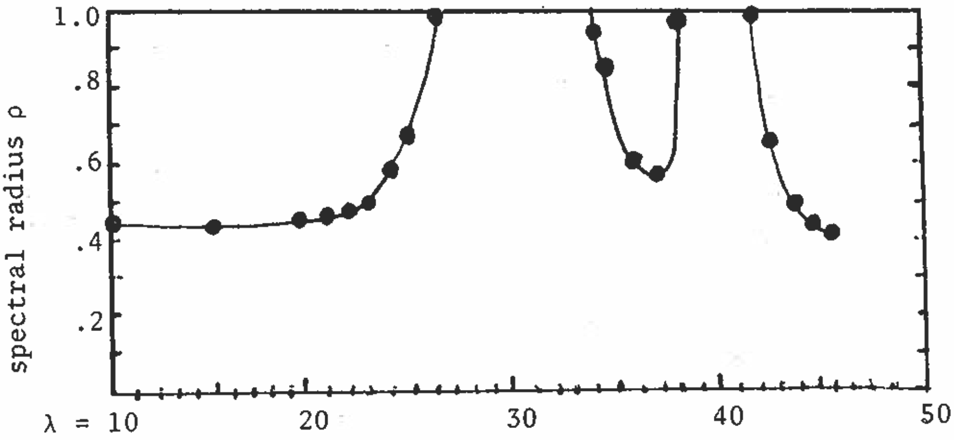


fig. 3

8. ORTHOGONAL DECOMPOSITION

Denote by V^k the nullspace of P^k if $k > 1$, with $V^1 = Y^1$, and denote by U^k the orthogonal complement of V^k in Y^k; we will use the orthogonal decomposition

(17) $$Y^k = U^k \oplus V^k$$

in the error analysis. Because of the close relationship between P^k and Q^k the subspace U^k has an alternative representation

(18) $$U^k = Y^k \cap Q^k(X^{k-1}).$$ $p.8: \mathcal{U}^k = Q^k(X^{k-1})$

(In fact, $2P^k$ is an approximate inverse to Q^k, being an example of a quasi-interpolation as defined by de Boor and Fix [11] or Frederickson [2].) Denote by $S^k : Y^{k-1} \to U^k$ the exact inverse to P^k on U^k, so that $S^kP^k : Y^k \to U^k$ and $I-S^kP^k : Y^k \to V^k$ are orthogonal projections.

We now extend (17) to the decomposition

(19) $$Y^k = V^{k,k} \oplus V^{k,k-1} \oplus \dots \oplus V^{k,1}$$

in which the subspaces $V^{k,j}$ are defined inductively by $V^{k,k} = V^k$ and

$V^{k,k-\alpha} = S^k S^{k-1} \dots S^{k-\alpha+1} V^{k-\alpha,k-\alpha}$

$V^{11} = Y^1$

$V^{2,1} = S^2 V^{11}$

(20) $$V^{k,j} = S^kV^{k-1,j}.$$

We also re-norm Y^k at this time, leaving it unchanged on $V^{k,k}$ and by induction on $V^{k,j}$ through (20) and the requirement that S^k be an isometry. This we extend to Y^k in such a way that the decomposition (19) is orthogonal. Finally, we define the orthogonal projection $M^{k,j} : Y^k \to V^{k,j}$ by induction, using

(21) $$M^{k,k} = I-S^kP^k; \quad M^{k,j} = S^kM^{k-1,j}P^k, \quad 1 \le j < k.$$

9. CONVERGENCE

We want to develop a bound on the maximum error or residual e which can remain after a single pass of FAPIN which begins with a residual r. If we represent this single pass as a linear operator $B : Y \to X$ we are interested in the spectral radius of $I-AB : Y \to Y$, since $e = (I-AB)r$. Since this global approximate inverse B does not have a simple representation, we find that we must trace the development of the error

$$e^k = r^k - A^k x^k \tag{22}$$

as it moves upward from $k = 1$. In our computations we also use this equation in the form

$$x^{k-1} = (A^{k-1})^{-1}(r^{k-1} - e^{k-1}) \tag{23}$$

To measure the accuracy of the approximate inverse B^k on $V^{k,k}$ we define, for $1 \leq k \leq \ell$, the constants

$$m_i = ||M^{k,k-i}(I-A^kB^k)M^{k,k}|| \tag{24}$$

We also need to use a measure of how well S^kA^{k-1} approximates A^kQ^k; this is done by defining, $1 < k \leq \ell$, the constants

$$\begin{aligned} n_i &= ||M^{k,k-i}(I-A^kB^k)(S^k-A^kQ^k(A^{k-1})^{-1})P^k|| && 0 \leq i \leq k-1 \\ c_{i,j} &= ||M^{k,k-i}(I-A^kB^k)A^kQ^k(A^{k-1})^{-1}P^kM^{k,k-j}|| && 0 \leq i,\ j \leq k-1 \end{aligned} \tag{25}$$

Finally, in order to put our convergence bound in a simple form, we define

$$\bar{m} = \sup 4^i m_i \qquad 1 \le i,\ 1 \le k \le \ell$$

(26) $$\bar{n} = \sup 4^i m_i \qquad 1 \le i,\ 1 < k \le \ell$$

$$\bar{c} = \sup 4^i 2^j (c_{i.j} - \delta_{i-j}) \qquad 1 \le i,j,\ 1 < k \le \ell$$

where δ_i denotes the Kronecker delta.

Theorem 2. *The spectral radius* ρ *of FAPIN satisfies*

(27) $$\rho < m_0 + \bar{m} + n_0 + \bar{n} + \frac{2\bar{c}}{3/2 - 8\bar{c}/7} \left(2(n_0 + \bar{n}/7) + (5/2 + 8\bar{c}/7)\ (m_0 + \bar{m}/7)\right)$$

provided that $\bar{c} < 7/16$.

Proof: We will consider a real positive matrix E^ℓ with the property that the error e^ℓ after any pass of FAPIN satisfies

(28) $$||M^{\ell,i} e^\ell|| \le \sum_{j=1}^{\ell} E^\ell_{i,j}\ ||M^{\ell,j} r^\ell||,$$

where r^ℓ is the residual before that pass. It follows from the orthogonality of the projections $M^{\ell,j}$ that $||e^\ell|| \le ||E^\ell|| \cdot ||r^\ell||$, or

(29) i.e. $||M|| = 1$ $$\rho \le ||E^\ell||$$

In order to bound the norm of E^ℓ we will construct a sequence $F = \{F_S\}_{S=-\infty}^{+\infty}$, independent of ℓ, with the property that

(30) $$E^{\ell}_{i,j} \le F_{j-i}$$

from which it follows that, for the corresponding normalization of Y,

(31) $$||E^{\ell}|| \le F(x) = \sum_{s=-\infty}^{\infty} F_s x^s \qquad x > 0$$

Since it is very near to the infinum of $F(x)$ we use $F(2)$ as the bound in Theorem 2. Using the notation of inequality (28) but with k replacing ℓ, we will denote by E^k a matrix bounding the error e^k at the kth level during any pass of FAPIN. We will also consider a sequence $F^k = \{F^k_s\}$ with the property that $E^k_{i,j} \le F^k_{j-1}$, as in (30), and the associated function $F^k(x)$ for $x = 2$. The linearity of FAPIN allows us to assume, with no loss of generality, that $||r^{\ell}|| = 1$ and that, for some m, $||M^{\ell,j}r^{\ell}|| = 0$ if $j \ne m$. Thus

(32) $$||M^{k,j}r^{k}|| = \begin{cases} 1 & \text{if } j = 1,\ k > m \\ 0 & \text{otherwise} \end{cases}$$

This means that we are interested only in the mth column of the matrix E (but for artibrary m). In particular, it is sufficient to construct E^k and F^k to satisfy

(33) $$||M^{k,i}e^{k}|| \le E^k_{i,m} \le F^k_{m-i}$$

Now a particular consequence of (32) is that for $k < m$, $r^k = x^k = e^k = 0$. Thus $x^m = B^m r^m$, and

$$e^m = (I-A^m B^m) r^m. \tag{34}$$

It follows immediately from (24) and (34) that E^m satisfies (33) if we define it by

$$E^m_{m-i,m} = m_i \qquad i = 0, \ldots, m-1 \tag{35}$$

and F^m satisfies (33), because of (26), if we define it by

$$F^m_i = \begin{cases} m_0 & i = 0 \\ \bar{m}\, 4^{-i} & i > 0 \\ 0 & i < 0 \end{cases} \tag{36}$$

For general k we apply (22) in (16) to get

$$e^k = (I-A^k B^k)(r^k - A^k Q^k x^{k-1}) \tag{37}$$

We substitute (23) in (37), and use the fact that $r^{k,k} = r^k - S^k P^k r^k = 0$ for $k > n$, which puts it in the form

$$\begin{aligned} e^k &= (I-A^k B^k)(S^k - A^k Q^k (A^{k-1})^{-1}) P^k r^k \\ &\quad + (I-A^k B^k)(A^k Q^k (A^{k-1})^{-1} P^k) S^k e^{k-1} \end{aligned}$$

applying (25) and (32) in (38) we have

$$||M^{k,k-i} e^k|| \le n_i + \sum_{j=0}^{k-1} C_{ij}\, ||M^{k,k-j} S^k e^{k-1}|| \tag{39}$$

By (21) and the re-normalization of Y^k, $||M^{k,k-j} S^k e^{k-1}|| = ||M^{k-1,k-j} e^{k-1}||$; using this and equation (26) we can show by induction that (33) is satisfied if we define

$$(40) \qquad E^{k}_{k-i,m} = n_i + \sum_{j=0}^{k-1} C_{i,j}\, E^{k-1}_{k-j,m}$$

Thus F^k_s satisfies (33) if we define it by induction, beginning with (36) and extending with

$$(41) \qquad F^{k}_{i-(k-m)} = \begin{cases} n_0 + \overline{c}\,\overline{F}^{k-1} & i = 0 \\ F^{k-1}_{i-(k-m)} + 4^{-i}(\overline{n}+\overline{c}\,\overline{F}^{k-1}) & i > 0 \\ 0 & i < 0 \end{cases}$$

in which we use

$$(42) \qquad \overline{F}^{k} = \sum_{j=0}^{k-1} 2^{-j}\, F^{k}_{j-(k-m)}$$

The combination of (36) and (41) with (42) yields the recurrence.

$$(43) \qquad \overline{F}^{k} = \begin{cases} m_0 + \overline{m}/7 & k = m \\ (n_0+\overline{n}/7) + (1/2+8\overline{c}/7)\overline{F}^{k-1} & k > m \end{cases}$$

which has the solution

$$(44) \qquad F^{k} = \frac{m_0+\overline{m}/7+n_0+\overline{n}/7}{1-\gamma} - \frac{n_0+\overline{n}/7+\gamma(m_0+\overline{m}/7)}{1-\gamma}\, \gamma^{k-m}$$

in which $\gamma = 1/2 + 8\overline{c}/7$. We may also use (41) to construct a recurrence on $F^k(2)$; simplifying this with (44) gives

(45)
$$\begin{aligned} F^{k}(2) &= F^{k-1}(2) + 2^{-(k-m)}(n_0+\overline{n}+2\overline{c}\ \overline{F}^{k-1}) \\ &= F^{k-1}(2) + 2^{-(k-m)}(n_0+\overline{n}+2\overline{c}\ \frac{m_0+\overline{m}/7+n_0+\overline{n}/7}{1-\gamma}) \\ &- (\gamma/2)^{k-m}\ 2\overline{c}\ \frac{n_0+\overline{n}/7+\gamma(m_0+\overline{m}/7)}{1-\gamma} \\ &\le m_0+\overline{m}+n_0+\overline{n} + \frac{4\overline{c}}{2-\gamma}(n_0+\overline{n}/7) + \frac{2\overline{c}(2+\gamma)(m_0+\overline{m}/7)}{2-\gamma} \end{aligned}$$

which we wish to bound, independent of k. The least such is

(46)
$$\begin{aligned} F^{k}(2) \le m_0+\overline{m}+n_0+\overline{n} &+ \frac{4\overline{c}}{2-\gamma}(n_0+\overline{n}/7) \\ &+ \frac{2\overline{c}}{2-\gamma}(2+\gamma)(m_0+\overline{m}/7), \end{aligned}$$

from which the conclusion of Theorem 1 follows, if $\gamma < 1$, completing the proof.

The content of Theorem 2 is best demonstrated by applying it to the constant coefficient discrete Laplacian operator A with representation

(47)
$$A = \begin{matrix} 1 & 1 & 1 \\ 1 & -8 & 1 \\ 1 & 1 & 1 \end{matrix}$$

This is the unique discrete Laplacian which is invariant under collection, or $A^{k} = A$ for all k, which follows from the fact that it is the Raleigh-Ritz-Galerkin discretization over bilinear splines of the Laplacian. For a constant coefficient operator we are able to use Fourier transforms, and thus evaluate

(48)
$$m_0 = 0.045 \qquad \bar{m} = 0.432$$
$$\bar{c} = 0.118$$
$$n_0 = 0.044 \qquad \bar{n} = 0.342$$

To get these constants this small we used a somewhat better approximate inverse than the DB_q one, namely

(49)
$$-400\ B = \begin{matrix} 5 & 6 & 5 \\ 6 & 52 & 6 \\ 5 & 6 & 5 \end{matrix}$$

Applying these constants in Theorem 2 we have

Corollary 2.1: *The algorithm FAPIN has a spectral radius*

(50)
$$\rho < 0.855$$

when solving the discrete Dirichlet problem with operator A *of equation (47) provided the approximate inverse* B *with respresentation (49) is used.*

Experiments indicate that ρ is actually considerably less than 0.20, but the above estimate is sharp enough for theoretical purposes.

Economizing in operations, as we did in (13), we find 54n additions and multiplications per iteration are required for the constant coefficient case and 89n operations for the general, variable coefficient case. Thus a bit of accuracy costs us, in actuality, fewer than 23n arithmetic operations, in this case. The best we can state theoretically, however, is

Corollary 2.2: *FAPIN solves the discrete Dirichlet problem to an accuracy* ε *in at most* $(239/(1-\log_2\varepsilon))n$ *operations, and using fewer than* $4n$ *words of storage.*

10. ACKNOWLEDGMENTS

The support and encouragement of Professors F. L. Bauer and R. Bulirsch of the Technische Universität and C. H. Reinsch of Leibniz Rechenzentrum in München is acknowledged with pleasure.

11. REFERENCES

[1] Varga, Richard S., Matrix Iterative Analysis, Prentice-Hall, Englewood Cliffs, N.J. (1962).

[2] Young, David M., Iterative Solution of Large Linear Systems, Academic Press, New York (1971).

[3] George, Alan, Nested Dissection of a Regular Finite Element Mesh, SIAM J. Numer. Anal. 10 (1973), 345-363.

[4] Benson, Maurice W., Iterative Solution of Large Scale Linear Systems, Thesis, Lakehead University, July 1973.

[5] ________________, and Frederickson, Paul O., Iterative Solution of Large Sparse Linear Systems, to appear.

[6] Schröder, Johann and Ulrich Trottenberg, Reductionsverfahren für Differenzengleichungen bei Randwertaufgaben I, Numer. Math. 22 (1973), 37-68.

[7] Buzbee, B.L., G. H. Golub, and C. W. Nielsen, On direct methods for solving Poisson's equations, SIAM J. Numer. Anal. 7 (1970), 627-656.

[8] Brandt, Achi,Multi-level adaptive technique (MLAT) for fast numerical solution to boundary value problems, Proc. of the Third International Conf. on Num. Meth. in Fluid Mech., Lecture Notes in Physics 18, Springer Verlag.(1972).

R.P.

[9] Fedorenko, P.P., The speed of convergence of one iterative method, Zh. Vychisl. Mat. Fiz. 4,3 (1964), 559-564. (USSR Computational Math. and Math. Phys. ~~19, 227-235.~~)

[10] Broy, Manfred, Numerische *Lösung hochdimensionaler linearer Gleichungssysteme endliecher Netze*, Thesis, Tech. Univ. München, April 1975.

[11] de Boor, Carl and George Fix, Spline Approximation by Quasi-interpolants, *J. of Approx. Theory* 8 (1973), 19-45.

[12] Frederickson, Paul O., Quasi-interpolation, extrapolation, and approximation on the plane, *Proc. First Man. Conf. on Num. Math.*, 159-167, Utilitas Mathematicae, Winnipeg (1971).