[1] Department of Computer Science, Yale University, P. O. Box 2158 Yale Station, New Haven, CT, USA 06520.

[2] Department of Computer Science, University of Toronto, Toronto, Ontario, Canada M5S 1A7.

[3] Department of Mathematics, Shandong University, Jinan, Shandong, PRC.

This work was supported in part by ONR Grant N0014-76-C-0277, ONR Grant N0008-80-0076 under subcontract from Florida State, and Air Force Grant AFOSR-81-0193 while the authors were at Yale University.

Alternating-Direction Incomplete Factorizations

By

Tony F. Chan[1], Kenneth R. Jackson[2], and Benren Zhu[3]

Technical Report #208

August 19, 1981

Abstract

To solve the system of linear equations $Aw = r$ that arises from the discretization of a two-dimensional self-adjoint elliptic differential equation, iterative methods employing easily computed incomplete factorizations, $LU = A+B$, are frequently used. Dupont, Kendall, and Rachford [5] showed that, for the DKR factorization, the number of iterations {arithmetic operations} required to reduce the A-norm of the error by a factor of ε is $O(h^{-1/2}\log\frac{1}{\varepsilon})$ $\{O(h^{-2\frac{1}{2}}\log\frac{1}{\varepsilon})\}$, where h is the stepsize used in the discretization. We present some error estimates which suggest that, if a pair of Alternating-Direction DKR Factorizations are used, then the number of iterations {arithmetic operations} may be decreased to $O(h^{-1/3}\log\frac{1}{\varepsilon})$ $\{O(h^{-2\frac{1}{3}}\log\frac{1}{\varepsilon})\}$. Numerical results supporting this estimate are included.

1. Introduction.

Iterative methods are frequently used to solve the system of linear equations

$$Aw = r \tag{1.1}$$

that arises from the usual five-point discretization of the Dirichlet problem for the two-dimensional self-adjoint elliptic differential equation

$$\frac{\partial}{\partial x}\left[a_1 \frac{\partial u}{\partial x} \right] + \frac{\partial}{\partial y}\left[a_2 \frac{\partial u}{\partial y} \right] + qu = f \quad \text{in} \quad \Omega, \tag{1.2}$$

where, throughout this paper, we assume

1. Ω is an open bounded region in R^2,

2. a_1, a_2 are Lipschitz continuous in $\bar{\Omega}$,

3. a_1, $a_2 \geq \eta > 0$ in $\bar{\Omega}$ for some constant η, and

4. $q \leq 0$ is bounded in $\bar{\Omega}$.

The efficiency of many iterative methods depends upon the selection of an easily-inverted approximation $\tilde{A}$ to A. Several authors [2, 4, 5, 7, 9, 10, 11, 12, 13] have suggested taking $\tilde{A}$ to be an incomplete factorization of A,

$$\tilde{A} = LU = A+B, \tag{1.3}$$

where B is chosen so that L and U are sparse.

For several of these factorizations, there are two directionally dependent forms of B: B_1 and B_2. Stone [13] found that, for his method, experimental results indicated that using the pair of incomplete factorizations alternately,

$$(A+B_1)w_{n+\frac{1}{2}} = (A+B_1)w_n - \omega(Aw_n - r) \tag{1.4}$$

$$(A+B_2)w_{n+1} = (A+B_2)w_{n+\frac{1}{2}} - \omega(Aw_{n+\frac{1}{2}} - r),$$

gave a faster rate of convergence than using either $\tilde{A} = A+B_1$ or $\tilde{A} = A+B_2$ alone in the stationary iteration

$$\tilde{A}w_{n+1} = \tilde{A}w_n - \omega(Aw_n - r). \tag{1.5}$$

Of course, eliminating $w_{n+\frac{1}{2}}$, we can rewrite the pair of equations (1.4) in the form (1.5) using

$$\tilde{A} = M_\omega = [A+B_1][(2-\omega)A+B_1+B_2]^{-1}[A+B_2] \tag{1.6}$$

provided that $[(2-\omega)A+B_1+B_2]$ is nonsingular.[1] We refer to the right side of (1.6) as an Alternating-Direction Incomplete Factorization. Although M_ω itself may be costly to compute, it is relatively inexpensive to solve $M_\omega x = b$, and it is the solution of such systems that is required in the

[1] Note that the formal inverse of M_ω, $[A+B_2]^{-1}[(2-\omega)A+B_1+B_2][A+B_{12}]^{-1}$, is always well-defined.

iteration (1.5) and its Chebyshev or conjugate gradient accelerations.[2]

In general, M_ω is nonsymmetric. Since, in many applications, it is advantageous for $\tilde{A}$ to be symmetric, we also consider

$$S_\omega^{-1} = \frac{1}{2}[M_\omega^{-1}+M_\omega^{-t}], \tag{1.7}$$

the symmetric part of M_ω^{-1}. Again, although S_ω itself may be costly to compute, it is relatively inexpensive to solve $S_\omega x = b$.

For the DKR factorization (an incomplete factorization similar to Stone's), Dupont, Kendall, and Rachford [5] showed that the number of iterations of (1.5) required to reduce the A-norm of the error by a factor of ε is $O(h^{-1}\log\frac{1}{\varepsilon})$ and the associated number of arithmetic operations is $O(h^{-3}\log\frac{1}{\varepsilon})$. Moreover, the iteration can be accelerated by Chebyshev or conjugate gradient methods, decreasing the the number of iterations required to $O(h^{-1/2}\log\frac{1}{\varepsilon})$ and the associated number of arithmetic operations to $O(h^{-2\frac{1}{2}}\log\frac{1}{\varepsilon})$. In this paper, we investigate whether these work estimates can be improved by using either the Alternating-Direction form (1.6) of the DKR factorization (AD-DKR) or the Symmetric Alternating-Direction form (1.7) of the DKR factorization (SAD-DKR).

---

[2] M_ω can be viewed as a one parameter family of preconditionings for A. From this point of view, it follows that, when (1.5) is accelerated by the Chebyshev or conjugate gradient technique, the parameter ω internal to M_ω should be held fixed, while the external parameter ω in (1.5) is varied.

In Section 2, we review the DKR factorization and present a modification. In Section 3, we review some general results concerning the rate of convergence of the stationary iteration (1.5) and its Chebyshev or conjugate gradient acceleration. Since these results are dependent upon the spectrum of $\tilde{A}^{-1}A$, we are led to an investigation of the eigenvalues of $M_\omega^{-1}A$ and $S_\omega^{-1}A$ in the following two sections. More specifically, in Section 4, using the additional restriction that $a_1 = a_2$, we develop eigenvalue estimates for a pair of factors of the iteration matrix $I-\omega M_\omega^{-1}A$ associated with the modified AD-DKR factorization. In Section 5, we explain why we believe that, for a large class of problems, these estimates suggest that the number of iterations of (1.5) required to reduce the A-norm of the error by a factor of ε may be $O(h^{-2/3}\log\frac{1}{\varepsilon})$ with the associated number of arithmetic operations being $O(h^{-2\frac{2}{3}}\log\frac{1}{\varepsilon})$, and, moreover, if (1.5) is accelerated by the Chebyshev or conjugate gradient methods, then the number of iterations may be decreased to $O(h^{-1/3}\log\frac{1}{\varepsilon})$ with the associated number of arithmetic operations being $O(h^{-2\frac{1}{3}}\log\frac{1}{\varepsilon})$. Although these work estimates are not rigorous, numerical results presented in Section 6 strongly support our conjecture that the estimates are accurate. In addition, the numerical results indicate that the estimates are valid for the unmodified as well as the modified forms of the DKR factorization.

2. The DKR Factorization.

In this section, following the notation of [5], we review the DKR factorization and present a modification.

Let $\bar{\Omega}_h$ be the set of points $(jh,kh) \in \bar{\Omega}$, where h is the stepsize associated with the discretization and j, k are integers, and let Ω_h be the set of points $(jh,kh) \in \bar{\Omega}_h$ such that $((j+1)h,kh)$, $((j-1)h,kh)$, $(jh,(k+1)h)$, $(jh,(k-1)h) \in \bar{\Omega}_h$ also. Then $\partial\Omega_h = \bar{\Omega}_h \backslash \Omega_h$. Let $w_{j,k}$ denote the value of the grid-function w at $(jh,kh) \in \bar{\Omega}_h$.

For each point $(jh,kh) \in \Omega_h$, we approximate the right side of (1.2) by the usual five-point self-adjoint difference operator

$$(Aw)_{j,k} = b_{j,k}w_{j,k} + c_{j,k}w_{j+1,k} + f_{j,k}w_{j,k+1} + c_{j-1,k}w_{j-1,k} + f_{j,k-1}w_{j,k-1}. \tag{2.1}$$

For definiteness, we take

$$c_{j,k} = -h^{-2}a_1((j+\tfrac{1}{2})h,kh),$$

$$f_{j,k} = -h^{-2}a_2(jh,(k+\tfrac{1}{2})h),$$

$$b_{j,k} = h^{-2}[a_1((j+\tfrac{1}{2})h,kh) + a_1((j-\tfrac{1}{2})h,kh) + a_2(jh,(k+\tfrac{1}{2})h) + a_2(jh,(k-\tfrac{1}{2})h)] - q(jh,kh),$$

although our results hold for other similar sets of coefficients.

When the linear difference operator A is written in matrix form, the terms in (2.1) that involve $w_{j,k}$ for $(jh,kh) \in \partial\Omega_h$ are incorporated into the right side of (1.1). Therefore, we adopt the convention that $w_{j,k} = 0$ if $(jh,kh) \notin \Omega_h$. For consistency of notation, we also adopt the alternative convention used in [5] that

$c_{j,k} = 0$ if $(jh,kh) \notin \Omega_h$ or $((j+1)h,kh) \notin \Omega_h$, and

$f_{j,k} = 0$ if $(jh,kh) \notin \Omega_h$ or $(jh,(k+1)h) \notin \Omega_h$.

With the latter convention, it is useful to define

$$\tilde{c}_{j,k} = -h^{-2} a_1((j+\tfrac{1}{2})h,kh),$$

$$\tilde{f}_{j,k} = -h^{-2} a_2(jh,(k+\tfrac{1}{2})h),$$

$$q_{j,k} = q(jh,kh),$$

for $(jh,kh) \in \bar{\Omega}$.

In [5], Dupont, Kendall, and Rachford introduced the DKR factorization

$$L_1 L_1^t = A + B_1 \quad \text{with} \quad B_1 = \tilde{B}_1 + D_1, \tag{2.2}$$

where

$$(L_1 w)_{j,k} = v^{(1)}_{j,k} w_{j,k} + t^{(1)}_{j-1,k} w_{j-1,k} + g^{(1)}_{j,k-1} w_{j,k-1}, \tag{2.3}$$

$$(\tilde{B}_1 w)_{j,k} = h^{(1)}_{j,k} w_{j-1,k+1} + h^{(1)}_{j+1,k-1} w_{j+1,k-1} - (h^{(1)}_{j,k} + h^{(1)}_{j+1,k-1}) w_{j,k}, \tag{2.4}$$

$$(D_1 w)_{j,k} = \alpha^{(1)}_{j,k} b_{j,k} w_{j,k}, \tag{2.5}$$

with coefficients given by

$$v^{(1)}_{j,k} = [b_{j,k}(1+\alpha^{(1)}_{j,k}) - h^{(1)}_{j,k} - h^{(1)}_{j+1,k-1} - (t^{(1)}_{j-1,k})^2 - (g^{(1)}_{j,k-1})^2]^{1/2}, \tag{2.6}$$

$$g^{(1)}_{j,k} = f_{j,k}/v^{(1)}_{j,k}, \tag{2.7}$$

$$t^{(1)}_{j,k} = c_{j,k}/v^{(1)}_{j,k}, \tag{2.8}$$

$$h^{(1)}_{j+1,k} = t^{(1)}_{j,k}g^{(1)}_{j,k} = c_{j,k}f_{j,k}/(v^{(1)}_{j,k})^2. \tag{2.9}$$

Since $c_{j,k}$ and $f_{j,k}$ are zero for $(jh,kh) \in \partial\Omega_h$, the coefficients of the factorization can be computed recursively for j and k increasing. We modify[3] this formulation by taking

$$h^{(1)}_{j+1,k} = \tilde{c}_{j,k}\tilde{f}_{j,k}/(v^{(1)}_{j,k})^2, \tag{2.10}$$

and initializing

$$v^{(1)}_{j,k} = [-\gamma^{(1)}_{j,k}(\tilde{c}_{j,k} + \tilde{f}_{j,k})]^{1/2} \tag{2.11}$$

---

[3] Note that the recurrence (2.10) differs from (2.9) only at the points adjacent to $\partial\Omega_h$, where $c_{j,k}$ or $f_{j,k}$ may be zero but $\tilde{c}_{j,k}$ and $\tilde{f}_{j,k}$ are not.

if $(jh,kh) \in \partial\Omega_h$ and either $((j+1)h,kh) \in \Omega_h$ or $(jh,(k+1)h) \in \Omega_h$.[4] Dupont, Kendall, and Rachford [5] showed that, for the unmodified factorization, the quantity under the square root on the right side of (2.6) is positive, whence $L_1L_1^t$ is symmetric and positive-definite. Lemma 4.1 proves that the modified DKR factorization possesses these properties also. However, for $\gamma_{j,k}^{(1)} < \infty$, the modification has the effect of making $\tilde{B}_1$ negative-definite rather that simply negative-semidefinite, as is the case for the unmodified factorization.[5] This difference is critical to the eigenvalue estimates developed in Section 4.

If the grid-points are renumbered with j decreasing and k increasing, then an alternative form of the DKR factorization is given by

$$L_2L_2^t = A+B_2 \quad \text{with} \quad B_2 = \tilde{B}_2+D_2, \tag{2.12}$$

where

$$(L_2w)_{j,k} = v_{j,k}^{(2)}w_{j,k} + t_{j+1,k}^{(2)}w_{j+1,k} + g_{j,k-1}^{(2)}w_{j,k-1}, \tag{2.13}$$

---

[4] The coefficients $\tilde{c}_{j,k}$ and $\tilde{f}_{j,k}$ used in (2.11) do not occur in the matrix A. Moreover, for some domains Ω and their discretizations, the computation of these coefficients may require the evaluation of a_1 and a_2, respectively, outside of $\bar{\Omega}$. If this presents a problem, $\tilde{c}_{j,k}$ and $\tilde{f}_{j,k}$ may be replaced by nearby nonzero values. In most instances, this alteration does not affect the factorization significantly.

[5] A vector w with all components equal is a null-vector for the matrix $\tilde{B}_1$ associated with the unmodified DKR factorization.

$$(\tilde{B}_2 w)_{j,k} = h^{(2)}_{j,k} w_{j+1,k+1} + h^{(2)}_{j-1,k-1} w_{j-1,k-1} - (h^{(2)}_{j,k} + h^{(2)}_{j-1,k-1}) w_{j,k}, \tag{2.14}$$

$$(D_2 w)_{j,k} = \alpha^{(2)}_{j,k} b_{j,k} w_{j,k}, \tag{2.15}$$

with coefficients for the unmodified factorization given by

$$v^{(2)}_{j,k} = [b_{j,k}(1+\alpha^{(2)}_{j,k}) - h^{(2)}_{j,k} - h^{(2)}_{j-1,k-1} - (t^{(2)}_{j+1,k})^2 - (g^{(2)}_{j,k-1})^2]^{1/2}, \tag{2.16}$$

$$g^{(2)}_{j,k} = f_{j,k}/v^{(2)}_{j,k}, \tag{2.17}$$

$$t^{(2)}_{j,k} = c_{j-1,k}/v^{(2)}_{j,k}, \tag{2.18}$$

$$h^{(2)}_{j-1,k} = t^{(2)}_{j,k} g^{(2)}_{j,k} = c_{j-1,k} f_{j,k}/(v^{(2)}_{j,k})^2. \tag{2.19}$$

For the modified factorization, we replace (2.19) by

$$h^{(2)}_{j-1,k} = \tilde{c}_{j-1,k} \tilde{f}_{j,k}/(v^{(2)}_{j,k})^2. \tag{2.20}$$

and initialize

$$v^{(2)}_{j,k} = [-\gamma^{(2)}_{j,k}(\tilde{c}_{j-1,k} + \tilde{f}_{j,k})]^{1/2} \tag{2.21}$$

if $(jh,kh) \varepsilon \partial\Omega_h$ and either $((j-1)h,kh) \varepsilon \Omega_h$ or $(jh,(k+1)h) \varepsilon \Omega_h$. For either the modified or unmodified factorizations, the coefficients can be computed recursively with j decreasing and k increasing.

Again, $L_2L_2^t$ is symmetric and positive-definite for both the modified and unmodified factorizations. Furthermore, for $\gamma_{j,k}^{(2)} < \infty$, the modification has the effect of making $\tilde{B}_2$ negative-definite rather that simply negative-semidefinite, as is the case for the unmodified factorization.

We end this section with a remark about the directional dependence of the factorizations (2.2) and (2.12). Not only are the coefficients computed in a different order, but, also, $\tilde{B}_1$ resembles a second-order difference operator with differences taken along lines $x+y=c$, while $\tilde{B}_2$ resembles a similar operator with differences taken along lines $x-y=c$.

3. Error Estimates.

In this section, we review some general results concerning the rate of convergence of the stationary iteration (1.5) and its Chebyshev or conjugate gradient acceleration.

To begin, we introduce some additional notation. If $x = (x_1,\ldots,x_n)$ and $y = (y_1,\ldots,y_n)$ are two n-vectors, let the inner-product of x and y be $(x,y) = x_1\bar{y}_1+\ldots+x_n\bar{y}_n$, where $\bar{y}_i$ is the complex conjugate of y_i. Let the norm of x be $\|x\| = (x,x)^{1/2}$ and, for any n by n matrix C, let the norm of C be $\|C\| = \max\{\ \|Cx\|\ :\ \|x\| = 1\ \}$. For any symmetric positive-definite matrix P, let the P-norm of x be $\|x\|_P = (Px,x)^{1/2}$ and the P-norm of C be $\|C\|_P = \max\{\ \|Cx\|_P\ :\ \|x\|_P = 1\ \}$. Also, for any matrix C, let the spectral radius of C be $\rho(C) = \max\{\ |\lambda|\ :\ \lambda$ and eigenvalue of C $\}$.

If w is the solution of (1.1), w_n is the n^{th} iterate generated by the stationary iteration (1.5), and $e_n = w_n - w$ is the error in the n^{th} iterate, then

$$e_n = [I-\omega\tilde{A}^{-1}A]e_{n-1} = [I-\omega\tilde{A}^{-1}A]^n e_0, \tag{3.1}$$

where e_0 is the error in the initial guess w_0. Since A is symmetric and positive-definite, it is valid to multiply (3.1) by $A^{1/2}$ to get

$$A^{1/2}e_n = [I-\omega A^{1/2}\tilde{A}^{-1}A^{1/2}]A^{1/2}e_{n-1} = [I-\omega A^{1/2}\tilde{A}^{-1}A^{1/2}]^n A^{1/2}e_0,$$

whence

$$\|e_n\|_A \le \|[I-\omega A^{1/2}\tilde{A}^{-1}A^{1/2}]^n\| \cdot \|e_0\|_A. \tag{3.2}$$

The last inequality is the basis for the following lemma.

Lemma 3.1: If $\rho = \rho(I-\omega A^{1/2}\tilde{A}^{-1}A^{1/2}) < 1$, then the number of iterations of (1.5) required to reduce the A-norm of the initial error by a factor of ε is at most $n+1$, where

$$(n-q) \log \frac{1}{\rho} - \log \binom{n}{q} = \log \frac{1}{\varepsilon} + \log c, \tag{3.3}$$

c is a positive constant,[6] and $q+1$ is the size of the largest Jordan block

---

[6] The constant c is dependent upon the similarity transformation that reduces $A^{1/2}\tilde{A}^{-1}A^{1/2}$ to Jordan normal form (see Theorem 3.1 of [14]).

of $A^{1/2}\tilde{A}^{-1}A^{1/2}$ with an eigenvalue of magnitude ρ. If $A^{1/2}\tilde{A}^{-1}A^{1/2}$ is normal, then

$$n = \log \frac{1}{\varepsilon} \;/\; \log \frac{1}{\rho}. \tag{3.4}$$

Proof: By Theorem 3.1 of [14],

$$\|[I-\omega A^{1/2}\tilde{A}^{-1}A^{1/2}]^n\| \leq c\binom{n}{q}\rho^{n-q},$$

for constants c, q and ρ specified above. This inequality, together with (3.2), proves the validity of (3.3). If $A^{1/2}\tilde{A}^{-1}A^{1/2}$ is normal, then $I-\omega A^{1/2}\tilde{A}^{-1}A^{1/2}$ can be diagonalized by a Hermitian similarity transformation, whence

$$\|[I-\omega A^{1/2}\tilde{A}^{-1}A^{1/2}]^n\| = \rho^n$$

and the well-known result (3.4) follows. **Q.E.D.**

If (1.5) is accelerated by the Chebyshev technique, then the error at the n^{th} step satisfies

$$e_n = P_n(\tilde{A}^{-1}A)e_0, \tag{3.5}$$

where $P_n(z)$ is the translated and normalized Chebyshev polynomial of degree n. (See, for example, [1].) Multiplying (3.5) by $A^{1/2}$ and taking norms, we get that

$$\|e_n\|_A \leq \|P_n(A^{1/2}\tilde{A}^{-1}A^{1/2})\|\cdot\|e_0\|_A. \tag{3.6}$$

The last inequality is the basis of the following lemma.

Lemma 3.2: If the eigenvalues of $A^{1/2}\tilde{A}^{-1}A^{1/2}$ lie in the ellipse

$$E = \{ z \in C : z = 1 - a \cos \theta + i b \sin \theta, 0 \leq \theta \leq 2\pi \}, \tag{3.7}$$

where $0 \leq b < a < 1$, then the number of iterations of the Chebyshev acceleration of (1.5) required to reduce the A-norm of the initial error by a factor of ε is at most n+1, where

$$n \log \frac{1}{r} - q \log n = \log \frac{1}{\varepsilon} + \log c, \tag{3.8}$$

c is a positive constant,[7] q+1 is the size of the largest Jordan block of $\tilde{A}^{-1}A$ with an eigenvalue on the ellipse E, and

$$r = (a + b)/(1 + \sqrt{1 - a^2 + b^2}). \tag{3.9}$$

If $\tilde{A}$ is symmetric, then

$$n = \log \frac{2}{\varepsilon} / \log \frac{1}{r}, \tag{3.10}$$

and, moreover, b = 0 in the expression for r.

Proof: By inequality (2.22) of [8], Section 6.2 of [1], and an

---

[7] The constant c is dependent upon both the similarity transformation that reduces $A^{1/2}\tilde{A}^{-1}A^{1/2}$ to Jordan normal form (see Theorem 3.1 of [14]) and the bound (2.22) of [8] on the Chebyshev polynomials.

argument similar to the one leading to Theorem 3.1 of [14],

$$\|P_n(A^{1/2}\tilde{A}^{-1}A^{1/2})\| \le cn^q r^n \tag{3.11}$$

for the constants c, q, and r specified above. This inequality, together with (3.6), proves the validity of (3.8). If $\tilde{A}$ is symmetric, then $A^{1/2}\tilde{A}^{-1}A^{1/2}$ has real eigenvalues and, moreover, it can be diagonalized by an orthogonal similarity transformation. Hence, it follows from a simplification of the argument used to prove (3.11) that

$$\|P_n(A^{1/2}\tilde{A}^{-1}A^{1/2})\| \le 2r^n,$$

where b = 0 in the expression for r. This together with inequality (3.6) proves the validity of (3.10). **Q.E.D.**

Although variants of the conjugate gradient algorithm have been developed for nonsymmetric problems, the analysis of these methods is not well-developed. Consequently, we limit our discussion of the conjugate gradient acceleration of (1.5) to the case that $\tilde{A}$ is symmetric and positive-definite. In this case, it is well-known that the approximate solution w_n generated by the conjugate gradient acceleration of (1.5) minimizes the A-norm of the associated error, e_n, over all possible errors of the form

$$\tilde{e}_n = p_n(A^{1/2}\tilde{A}^{-1}A^{1/2})e_0,$$

where $p_n(z)$ is a polynomial of degree n satisfying $p_n(0) = 1$. (See, for example, [1].) Since the translated and normalized Chebyshev polynomial,

$P_n(z)$, satisfies these conditions, the following lemma is an immediate consequence of Lemma 3.2.

Lemma 3.3: If $\tilde{A}$ is symmetric and the eigenvalues of $A^{1/2}\tilde{A}^{-1}A^{1/2}$ lie in the interval $[1-a,1+a]$, $0 \leq a < 1$, then the number of iterations of the conjugate gradient acceleration of (1.5) required to reduce the A-norm of the initial error by a factor of ε is at most $n+1$, where

$$n = \log \frac{2}{\varepsilon} \bigg/ \log \frac{1}{r} \tag{3.12}$$

and

$$r = a/(1 + \sqrt{1 - a^2}).$$

To use the results developed in this section to bound the number of iterations of (1.5) or its acceleration, we require estimates of the spectrum of $\tilde{A}^{-1}A$. We turn to this question next.

4. Eigenvalue Estimates.

For the AD-DKR factorization, the iteration matrix associated with (1.5) is

$$I-\omega M_\omega^{-1}A = [A+B_1]^{-1}[(1-\omega)A+B_1][A+B_2]^{-1}[(1-\omega)A+B_2],$$

which is similar to

$$[(1-\omega)A+B_1][A+B_2]^{-1}[(1-\omega)A+B_2][A+B_1]^{-1}.$$

In this section, we develop some eigenvalue estimates for the pair of factors $[(1-\omega)A+B_1][A+B_2]^{-1}$ and $[(1-\omega)A+B_2][A+B_1]^{-1}$. These estimates provide some guidance (which has proven to be very effective in practice) for choosing the parameters $\{\alpha^{(i)}_{j,k}\}$ and ω required by the AD-DKR and SAD-DKR factorizations. Moreover, these estimates are the basis for the conjectures developed in the next section concerning the work required to solve (1.1) to a specified tolerance.

A number of preliminary lemmas are required before we can state and prove the main result of this section.

Lemma 4.1: For either form of the modified DKR factorization (2.2) or (2.12), if $\alpha^{(i)}_{j,k} \geq 0$ and $\beta_i \leq \gamma^{(i)}_{j,k} \leq \infty$, i=1,2, where

$$\beta_i = \min \{ \tfrac{1}{2}[(1+\alpha^{(i)}_{j,k})(1+\rho^{(i)}_{j,k}) + [\{(1+\alpha^{(i)}_{j,k})(1+\rho^{(i)}_{j,k})\}^2-4\rho^{(i)}_{j,k}]^{1/2}] \}$$

$$\rho^{(1)}_{j,k} = \frac{\tilde{c}_{j-1,k}+\tilde{f}_{j,k-1}}{\tilde{c}_{j,k}+\tilde{f}_{j,k}},$$

$$\rho^{(2)}_{j,k} = \frac{\tilde{c}_{j,k}+\tilde{f}_{j,k-1}}{\tilde{c}_{j-1,k}+\tilde{f}_{j,k}},$$

then

$$(v^{(1)}_{j,k})^2 \geq -\beta_1(\tilde{c}_{j,k} + \tilde{f}_{j,k}), \tag{4.1}$$

$$(v^{(2)}_{j,k})^2 \geq -\beta_2(\tilde{c}_{j-1,k} + \tilde{f}_{j,k}), \tag{4.2}$$

$$0 \leq h^{(1)}_{j+1,k} \leq -\frac{1}{\beta_1}\frac{\tilde{c}_{j,k}\cdot\tilde{f}_{j,k}}{\tilde{c}_{j,k}+\tilde{f}_{j,k}}, \tag{4.3}$$

$$0 \leq h^{(2)}_{j-1,k} \leq - \frac{1}{\beta_2} \frac{\tilde{c}_{j-1,k} \cdot \tilde{f}_{j,k}}{\tilde{c}_{j-1,k} + \tilde{f}_{j,k}} \tag{4.4}$$

at all points $(jh,kh) \in \bar{\Omega}_h$ at which $v^{(i)}_{j,k}$ and $h^{(i)}_{j,k}$ are required. Moreover, if the user selected parameters $\{\alpha^{(i)}_{j,k}\}$ and $\{\gamma^{(i)}_{j,k}\}$ are uniformly bounded above independently of the stepsize, h, then

$$0 < h^{-2}H \leq h^{(i)}_{j,k}, \tag{4.5}$$

where H, although problem dependent, is independent of the stepsize, h.[8]

Proof: We prove this lemma for the factorization (2.2) only, as the proof for (2.12) is similar.

Since the initial values of $v^{(1)}_{j,k}$ and $h^{(1)}_{j+1,k}$ for the modified DKR factorization satisfy (4.1) and (4.3) and the basic recurrence relations used to calculate the coefficients for both the original and modified DKR factorizations are essentially the same, the induction argument used in Lemma 1 of [5] also proves (4.1) and (4.3).

To prove (4.5), note that, if $v^{(1)}_{j,k}$ is computed by the recurrence

---

[8] Inequality (4.5) does not hold near $\partial\Omega_h$ for the unmodified DKR factorization.

(2.6), then

$$(v^{(1)}_{j,k})^2 \leq (1+\alpha^{(1)}_{j,k})b_{j,k} = -(1+\alpha^{(1)}_{j,k})(1+\rho^{(1)}_{j,k}+\mu_{j,k})(\tilde{c}_{j,k}+\tilde{f}_{j,k}), \tag{4.6}$$

where

$$\mu_{j,k} = \frac{q_{j,k}}{\tilde{c}_{j,k}+\tilde{f}_{j,k}},$$

whence, by (2.10), (2.11), and (4.6),

$$h^{(1)}_{j+1,k} \geq \frac{-1}{\xi_{j,k}} \frac{\tilde{c}_{j,k}\cdot\tilde{f}_{j,k}}{\tilde{c}_{j,k}+\tilde{f}_{j,k}} \geq \frac{-1}{2\xi_{j,k}} \min(\tilde{c}_{j,k},\tilde{f}_{j,k}) \geq \frac{h^{-2}\eta}{2\xi_{j,k}},$$

where $\xi_{j,k}$ is either $(1+\alpha^{(1)}_{j,k})(1+\rho^{(1)}_{j,k}+\mu_{j,k})$ or $\gamma^{(1)}_{j,k}$ depending upon whether $v^{(1)}_{j,k}$ is computed from (2.6) or (2.11), respectively. The proof is completed by observing that the assumptions on $\alpha^{(i)}_{j,k}$, $\gamma^{(i)}_{j,k}$, a_1, a_2 and q ensure that $\xi_{j,k}$ is bounded above independently of the stepsize, h. **Q.E.D.**

Lemma 4.2: For either form of the modified DKR factorization (2.2) or (2.12), if $\alpha^{(i)}_{j,k} \geq 0$, then $\beta_i \geq 1$. Moreover, if $\alpha^{(i)}_{j,k} = c_0 h^p$ for constants $c_0 > 0$ and $0 < p \leq 2$, then $\beta_i \geq 1+c_1 h^{p/2}$ for some constant $c_1 > 0$.

Proof: The bound $\beta_1 \geq 1$ follows directly from the definition of β_1 in Lemma 4.1. If, in addition, $\alpha^{(i)}_{j,k} = c_0 h^p$ and a_1, $a_2 \geq \eta > 0$ are Lipschitz continuous in $\bar{\Omega}$, then $\beta_1 \geq 1+c_1 h^{p/2}$ by an argument similar to the one used to prove (4.15) in [5]. The corresponding inequalities for β_2 are proved in a similar way. **Q.E.D.**

Lemma 4.3: For A, $\tilde{B}_1$, and $\tilde{B}_2$ defined by (2.1), (2.4), and (2.14), respectively,

$$(Ax,x) = -\sum \{\tilde{c}_{j,k}|x_{j+1,k}-x_{j,k}|^2 + \tilde{f}_{j,k}|x_{j,k+1}-x_{j,k}|^2 + q_{j,k}|x_{j,k}|^2\}, \tag{4.7}$$

$$(\tilde{B}_1x,x) = -\sum h^{(1)}_{j+1,k}|x_{j+1,k}-x_{j,k+1}|^2, \tag{4.8}$$

$$(\tilde{B}_2x,x) = -\sum h^{(2)}_{j,k}|x_{j+1,k+1}-x_{j,k}|^2, \tag{4.9}$$

where we have used the convention that $x_{j,k} = 0$ for $x_{j,k} \notin \Omega_h$ and the sums are taken over all nonzero terms.

Proof: The validity of equations (4.7)-(4.9) can be demonstrated easily by summation by parts, as is the validity of the similar set of equations (4.7)-(4.8) in [5]. **Q.E.D.**

Lemma 4.4: For either form of the modified DKR factorization (2.2) or (2.12), if $\alpha^{(i)}_{j,k} \geq 0$ and $\beta_i \leq \gamma^{(i)}_{j,k} \leq \infty$, then

$$0 \leq -(\tilde{B}_ix,x) \leq \frac{1}{\beta_i}(Ax,x), \tag{4.10}$$

If, in addition, the user-selected parameters $\{\alpha^{(i)}_{j,k}\}$ and $\{\gamma^{(i)}_{j,k}\}$ are uniformly bounded above independently of the stepsize, h, then

$$c_2(x,x) \leq -(\tilde{B}_ix,x) \tag{4.11}$$

for some constant $c_2 > 0$.[9]

Proof: We prove this lemma for the factorization (2.2) only, as the proof for (2.12) is similar.

To prove (4.10), we use an argument similar to the one used to prove (4.11) in [5]. First, observe that $0 \leq -(\tilde{B}_1 x,x)$ follows directly from (4.8) of Lemma 4.3, since $h^{(1)}_{j+1,k} \geq 0$ by Lemma 4.1. To verify the upper bound on $-(\tilde{B}_1 x,x)$, note that, by Lemma 3 of [5],

$$\frac{c \cdot f}{c+f}|a-b|^2 \leq c|a-e|^2 + f|b-e|^2,$$

for any positive c, f and any complex a, b, e. This inequality, together with Lemmas 4.1 and 4.3, shows that

$$-(\tilde{B}_1 x,x) = \sum h^{(1)}_{j+1,k}|x_{j+1,k}-x_{j,k+1}|^2$$

$$\leq -\frac{1}{\beta_1}\sum \frac{\tilde{c}_{j,k}\cdot\tilde{f}_{j,k}}{\tilde{c}_{j,k}+\tilde{f}_{j,k}}|x_{j+1,k}-x_{j,k+1}|^2$$

$$\leq -\frac{1}{\beta_1}\sum \{\tilde{c}_{j,k}|x_{j+1,k}-x_{j,k}|^2 + \tilde{f}_{j,k}|x_{j,k+1}-x_{j,k}|^2\}$$

$$\leq \frac{1}{\beta_1}(Ax,x).$$

---

[9] For the unmodified DKR factorization, inequality (4.11) does not hold, whence $\tilde{B}_i$ is negative-semidefinite rather than negative-definite.

To prove (4.11), observe that, by Lemmas 4.1 and 4.3,

$$-(\tilde{B}_1 x,x) = \sum h^{(1)}_{j+1,k} |x_{j+1,k} - x_{j,k+1}|^2$$

$$\geq h^{-2} H \sum |x_{j+1,k} - x_{j,k+1}|^2$$

$$\geq h^{-2} H \sum_{\text{all } L} \sum_{L} |y_L^{(\ell+1)} - y_L^{(\ell)}|^2,$$

where each L is a diagonal in Ω_h satisfying x+y=c, for some constant c, and $\{y_L^{(\ell)}\}$ is the subset of $\{x_{j,k}\}$ on L. For each L, let y_L be the n-vector with components $\{y_L^{(\ell)}\}$ on the diagonal L, and let C_L be the n by n matrix $h^{-2}\mathrm{diag}(-1,2,-1)$. Then

$$h^{-2} \sum |y_L^{(\ell+1)} - y_L^{(\ell)}|^2 = (C_L y_L, y_L) \geq \lambda_L (y_L, y_L),$$

where λ_L is the minimum eigenvalue of C_L. Since the length of any diagonal L in Ω_h is bounded, there exists a constant $\lambda_* > 0$, independent of both h and L, such that $\lambda_L \geq \lambda_* > 0$. Consequently, (4.11) holds for $c_2 = H\lambda_* > 0$.

Q.E.D.

Lemma 4.5: If

1. $\alpha^{(i)}_{j,k} = c_0 h^p$ for constants $c_0 > 0$ and $0 < p < 2$, and

2. $a_1 = a_2$,

then

$$0 \leq -(\tilde{B}_1 x,x) - (\tilde{B}_2 x,x) \leq (1 - c_3 h^{p/2})(Ax,x).$$

for some constant $c_3 > 0$.

Proof: By Lemmas 4.1 and 4.3,

$$0 \leq - (\tilde{B}_1 x, x) - (\tilde{B}_2 x, x)$$

$$\leq - \frac{1}{\beta_*} \sum \frac{\tilde{c}_{j,k} \cdot \tilde{f}_{j,k}}{\tilde{c}_{j,k} + \tilde{f}_{j,k}} |x_{j+1,k} - x_{j,k+1}|^2 + \frac{\tilde{c}_{j,k} \cdot \tilde{f}_{j+1,k}}{\tilde{c}_{j,k} + \tilde{f}_{j+1,k}} |x_{j+1,k+1} - x_{j,k}|^2,$$

where $\beta_* = \min\{\beta_1, \beta_2\}$. For some constant L,

$$- \frac{\tilde{c}_{j,k} \cdot \tilde{f}_{j+1,k}}{\tilde{c}_{j,k} + \tilde{f}_{j+1,k}} \leq - \frac{\tilde{c}_{j,k} \cdot \tilde{f}_{j,k}}{\tilde{c}_{j,k} + \tilde{f}_{j,k}} (1+Lh),$$

since $a_2 \geq \eta > 0$ is Lipschitz continuous in $\bar{\Omega}$. Also, for any complex values a, b, c, d,

$$|a-b|^2 + |c-d|^2 \leq |a-c|^2 + |a-d|^2 + |b-c|^2 + |b-d|^2.$$

Therefore,

$$0 \leq - (\tilde{B}_1 x, x) - (\tilde{B}_2 x, x)$$

$$\leq - \frac{1+Lh}{\beta_*} \sum \frac{\tilde{c}_{j,k} \cdot \tilde{f}_{j,k}}{\tilde{c}_{j,k} + \tilde{f}_{j,k}} (|x_{j+1,k} - x_{j,k+1}|^2 + |x_{j+1,k+1} - x_{j,k}|^2)$$

$$\leq - \frac{1+Lh}{\beta_*} \sum \frac{\tilde{c}_{j,k} \cdot \tilde{f}_{j,k}}{\tilde{c}_{j,k} + \tilde{f}_{j,k}} (|x_{j+1,k} - x_{j,k}|^2 + |x_{j,k+1} - x_{j,k}|^2 + |x_{j+1,k+1} - x_{j,k+1}|^2 + |x_{j+1,k+1} - x_{j+1,k}|^2).$$

Since $a_1 = a_2 \geq \eta > 0$ are Lipschitz continuous in $\bar{\Omega}$,

$$-\frac{\tilde{c}_{j,k}\cdot\tilde{f}_{j,k}}{\tilde{c}_{j,k}+\tilde{f}_{j,k}} \leq -\frac{1+Lh}{2}\min\{\tilde{c}_{j,k},\tilde{f}_{j,k},\tilde{c}_{j,k+1},\tilde{f}_{j+1,k}\}$$

for L sufficiently large, whence

$$0 \leq -(\tilde{B}_1 x,x) - (\tilde{B}_2 x,x)$$

$$\leq -\frac{(1+Lh)^2}{2\beta_*}\sum\{\tilde{c}_{j,k}|x_{j+1,k}-x_{j,k}|^2 + \tilde{c}_{j,k+1}|x_{j+1,k+1}-x_{j,k+1}|^2$$

$$+ \tilde{f}_{j,k}|x_{j,k+1}-x_{j,k}|^2 + \tilde{f}_{j+1,k}|x_{j+1,k+1}-x_{j+1,k}|^2\}$$

$$\leq -\frac{(1+Lh)^2}{\beta_*}\sum\{\tilde{c}_{j,k}|x_{j+1,k}-x_{j,k}|^2 + \tilde{f}_{j,k}|x_{j,k+1}-x_{j,k}|^2\}$$

$$\leq (1-c_3h^{p/2})(Ax,x),$$

where the last inequality follows from Lemmas 4.2 and 4.4. Q.E.D.

Corollary 4.6: If

1. $\alpha^{(i)}_{j,k} = c_0h^p$ for constants $c_0 > 0$ and $0 < p < 2$,

2. $a_1 = a_2$, and

3. $0 < \omega \leq 1$,

then the AD-DKR iteration matrix M_ω is well-defined.

Proof: From Lemma 4.5, if $0 < \omega \leq 1$, then $[(2-\omega)A+\tilde{B}_1+\tilde{B}_2]$ is positive-definite, whence so is $[(2-\omega)A+B_1+B_2]$. Therefore, $[(2-\omega)A+B_1+B_2]$ is nonsingular and M_ω is well-defined by (1.6). **Q.E.D.**

Theorem 4.7: Assume that

1. $\alpha_{j,k}^{(i)} = c_0h^p$ for constants $c_0 > 0$ and $0 < p \leq 2$,
2. $\beta_i \leq \gamma_{j,k}^{(i)} \leq \gamma_*$ for some constant $\gamma_* < \infty$ independent of the stepsize, h, and
3. $a_1 = a_2$.

Then any eigenvalue λ of either $[(1-\omega)A+B_1][A+B_2]^{-1}$ or $[(1-\omega)A+B_2][A+B_1]^{-1}$ is real and satisfies

$$-1 + c_4h^{p/2} - (\omega-1)c_6h^{-p} \leq \lambda \leq 1 - c_5h^{2-p}, \tag{4.12}$$

if $\omega \geq 1$, and

$$-1 + c_4h^{p/2} \leq \lambda \leq 1 - c_5h^{2-p} + (1-\omega)c_6h^{-p}, \tag{4.13}$$

if $\omega \leq 1$, where c_4, c_5, c_6 are positive constants.

Proof: We prove this result for $[(1-\omega)A+B_1][A+B_2]^{-1}$ only, as the proof for $[(1-\omega)A+B_2][A+B_1]^{-1}$ is similar.

Since A, B_1, and B_2 are symmetric and $L_2L_2^t = A+B_2$ is positive-definite, $[(1-\omega)A+B_1][A+B_2]^{-1}$ is similar to the symmetric matrix $[A+B_2]^{-1/2}[(1-\omega)A+B_1][A+B_2]^{-1/2}$. Consequently, the eigenvalues of $[(1-\omega)A+B_1][A+B_2]^{-1}$ are real. Moreover, for $x = [A+B_2]^{-1/2}y$, $y \neq 0$,

$$\frac{([A+B_2]^{-1/2}[(1-\omega)A+B_1][A+B_2]^{-1/2}y,y)}{(y,y)} = \frac{([(1-\omega)A+B_1]x,x)}{([A+B_2]x,x)},$$

whence any eigenvalue λ of $[(1-\omega)A+B_1][A+B_2]^{-1}$ satisfies

$$\min_{x\neq 0} \frac{([(1-\omega)A+B_1]x,x)}{([A+B_2]x,x)} \leq \lambda \leq \max_{x\neq 0} \frac{([(1-\omega)A+B_1]x,x)}{([A+B_2]x,x)}.$$

In addition, since $B_1 = \tilde{B}_1+D_1$, $B_2 = \tilde{B}_2+D_2$, and, by Assumption 1, $D_1 = D_2 = D$,

$$\frac{([(1-\omega)A+B_1]x,x)}{([A+B_2]x,x)} = \frac{(1-\omega)(Ax,x)+(\tilde{B}_1x,x)+(Dx,x)}{(Ax,x)+(\tilde{B}_2x,x)+(Dx,x)}. \tag{4.14}$$

Thus, to verify that inequalities (4.12) and (4.13) hold, it is sufficient to develop upper and lower bounds for the right side of (4.14), where, throughout this proof, we assume $x \neq 0$.

Since $L_2L_2^t = A+B_2$ is positive-definite, $(Ax,x)+(\tilde{B}_2x,x)+(Dx,x) > 0$. Therefore, if $(1-\omega)(Ax,x)+(\tilde{B}_1x,x)+(Dx,x) \leq 0$, then

$$\frac{(1-\omega)(Ax,x)+(\tilde{B}_1x,x)+(Dx,x)}{(Ax,x)+(\tilde{B}_2x,x)+(Dx,x)} \leq 0 \leq 1 - c_5h^{2-p}$$

for c_5 sufficiently small, as h is bounded above in any discretization of Ω. On the other hand, if $(1-\omega)(Ax,x)+(\tilde{B}_1x,x)+(Dx,x) \geq 0$, then

$$\frac{(1-\omega)(Ax,x)+(\tilde{B}_1x,x)+(Dx,x)}{(Ax,x)+(\tilde{B}_2x,x)+(Dx,x)} \leq 1 + \frac{(\tilde{B}_1x,x)}{(Dx,x)} + (1-\omega)\frac{(Ax,x)}{(Dx,x)}, \tag{4.15}$$

since, by Lemmas 4.2 and 4.4, $(Ax,x)+(\tilde{B}_2x,x) > 0$. By Assumptions 1-2 and the assumptions on a_1, a_2, and q, there exist positive constants m and M such that

$$mh^{-2+p}(x,x) \leq (Dx,x) \leq Mh^{-2+p}(x,x), \tag{4.16}$$

whence, by Lemma 4.4,

$$\frac{(\tilde{B}_1x,x)}{(Dx,x)} \leq -c_5h^{2-p}$$

for $c_5 \leq c_2/M$. Furthermore, from Lemma 4.3 and the definition of D,

$$\frac{(Ax,x)}{(Dx,x)} \leq c_6h^{-p}$$

for $c_6 \geq 2/c_0$. Hence, if $\omega \geq 1$, then

$$\frac{(1-\omega)(Ax,x)+(\tilde{B}_1x,x)+(Dx,x)}{(Ax,x)+(\tilde{B}_2x,x)+(Dx,x)} \leq 1 - c_5h^{2-p},$$

and, if $\omega \leq 1$, then

$$\frac{(1-\omega)(Ax,x)+(\tilde{B}_1x,x)+(Dx,x)}{(Ax,x)+(\tilde{B}_2x,x)+(Dx,x)} \leq 1 - c_5h^{2-p} + (1-\omega)c_6h^{-p},$$

showing that the upper bounds for inequalities (4.12) and (4.13) are valid.

To verify the lower bounds, consider two cases depending upon whether $(Dx,x) > (Ax,x)$. If $(Dx,x) > (Ax,x)$, then

$$(B_ix,x) + (Dx,x) > 0, \quad i = 1,2,$$

by Lemmas 4.2 and 4.4, whence

$$\frac{(1-\omega)(Ax,x)+(\tilde{B}_1x,x)+(Dx,x)}{(Ax,x)+(\tilde{B}_2x,x)+(Dx,x)} \geq \frac{(1-\omega)(Ax,x)}{(Ax,x)+(\tilde{B}_2x,x)+(Dx,x)}.$$

Therefore, if $\omega \leq 1$, then

$$\frac{(1-\omega)(Ax,x)+(\tilde{B}_1x,x)+(Dx,x)}{(Ax,x)+(\tilde{B}_2x,x)+(Dx,x)} \geq 0,$$

and, if $\omega \geq 1$, then

$$\frac{(1-\omega)(Ax,x)+(\tilde{B}_1x,x)+(Dx,x)}{(Ax,x)+(\tilde{B}_2x,x)+(Dx,x)} \geq -(\omega-1).$$

Thus, the lower bounds in the theorem are satisfied in this case provided that c_6 is sufficiently large, since h is bounded above in any discretization of Ω.

On the other hand, if $(Dx,x) \leq (Ax,x)$, then

$$\frac{(Ax,x)}{(Ax,x)+(\tilde{B}_2x,x)+(Dx,x)} \geq \frac{1}{2}. \qquad (4.17)$$

Furthermore, by Lemma 4.5,

$$(\tilde{B}_1x,x) \geq -(1-c_3h^{p/2})(Ax,x) - (\tilde{B}_2x,x),$$

whence

$$\frac{(1-\omega)(Ax,x)+(\tilde{B}_1x,x)+(Dx,x)}{(Ax,x)+(\tilde{B}_2x,x)+(Dx,x)}$$

$$\geq \frac{-(Ax,x)-(\tilde{B}_2x,x)+(Dx,x)+(1-\omega+c_3h^{p/2})(Ax,x)}{(Ax,x)+(\tilde{B}_2x,x)+(Dx,x)}$$

$$\geq -1 + (1-\omega+c_3h^{p/2})\frac{(Ax,x)}{(Ax,x)+(\tilde{B}_2x,x)+(Dx,x)}.$$

Consequently, by (4.17), if $\omega \leq 1$, then

$$\frac{(1-\omega)(Ax,x)+(\tilde{B}_1x,x)+(Dx,x)}{(Ax,x)+(\tilde{B}_2x,x)+(Dx,x)} \geq -1 + c_4h^{p/2}$$

for $c_4 \leq c_3/2$, and, if $\omega \geq 1$, then

$$\frac{(1-\omega)(Ax,x)+(\tilde{B}_1x,x)+(Dx,x)}{(Ax,x)+(\tilde{B}_2x,x)+(Dx,x)} \geq -1 + c_4h^{p/2} - (\omega-1)c_6h^{-p},$$

showing that the lower bounds for inequalities (4.12) and (4.13) are valid.

Q.E.D.

If our objective is to minimize $\rho([(1-\omega)A+B_1][A+B_2]^{-1})$ and $\rho([(1-\omega)A+B_2][A+B_1]^{-1})$ in the hope that this will minimize $\rho(I-\omega M_\omega^{-1}A)$ and lead to an effective stationary iteration, then, based upon equations (4.12) and (4.13), we should take $\omega = 1$ and $p = \frac{4}{3}$. For future reference, we restate Theorem 4.7 for these particular values of ω and p.

Theorem 4.8: Assume that

1. $\omega = 1$,

2. $\alpha_{j,k}^{(i)} = c_0h^{4/3}$ for some positive constant c_0,

3. $\beta_i \leq \gamma_{j,k}^{(i)} \leq \gamma_*$ for some constant $\gamma_* < \infty$ independent of the stepsize, h, and

4. $a_1 = a_2$.

Then any eigenvalue λ of either $B_1[A+B_2]^{-1}$ or $B_2[A+B_1]^{-1}$ satisfies

$$-1 + c_4h^{2/3} \leq \lambda \leq 1 - c_5h^{2/3},$$

whence

$$\rho(B_1[A+B_2]^{-1}) \leq 1 - c_7h^{2/3}$$

and

$$\rho(B_2[A+B_1]^{-1}) \leq 1 - c_7 h^{2/3},$$

where $c_7 = \min\{c_4, c_5\} > 0$.

5. Work Estimates: Conjectures and Discussion.

In this section, using the eigenvalue estimates from the previous section, we develop conjectures that both $\rho(I-M_1^{-1}A)$ and $\rho(I-S_1^{-1}A)$ are bounded by $1-ch^{2/3}$, for some positive constant c. If the conjecture for the SAD-DKR factorization is valid, then the number of iterations of (1.5) required to reduce the A-norm of the initial error by a factor of ε is $O(h^{-2/3}\log\frac{1}{\varepsilon})$ with the associated number of arithmetic operations being $O(h^{-2\frac{2}{3}}\log\frac{1}{\varepsilon})$. Moreover, if (1.5) is accelerated by the Chebyshev or conjugate gradient techniques, then the number of iterations is decreased to $O(h^{-1/3}\log\frac{1}{\varepsilon})$ with the associated number of arithmetic operations being $O(h^{-2\frac{1}{3}}\log\frac{1}{\varepsilon})$. If additional conjectures concerning the spectral structure of $M_1^{-1}A$ hold, then, for the AD-DKR factorization, similar work estimates are valid for the stationary iteration (1.5) and its Chebyshev acceleration. Although the work estimates in this section are not rigorous, numerical results presented in the next section strongly support our conjecture that they are accurate.

We begin by stating the two fundamental conjectures about $\rho(I-M_1^{-1}A)$ and $\rho(I-S_1^{-1}A)$ upon which the work estimates in this section are based.

Conjecture 5.1: If the assumptions of Theorem 4.8 hold, then $\rho(I-M_1^{-1}A) \leq 1-c_7h^{2/3}$. Moreover, the eigenvalues of $M_1^{-1}A$ lie in a very eccentric ellipse, the major-axis of which is contained in the interval

$[c_7h^{2/3}, 2-c_7h^{2/3}]$.

Discussion: If C_1 and C_2 are normal matrices, then

$$\rho(C_1C_2) \le \|C_1C_2\| \le \|C_1\|\cdot\|C_2\| = \rho(C_1)\rho(C_2).$$

Hence, if $B_1[A+B_2]^{-1}$ and $B_2[A+B_1]^{-1}$ were normal, then

$$\begin{aligned}\rho(I-M_1^{-1}A) &= \rho([A+B_1]^{-1}B_1[A+B_2]^{-1}B_2) \\ &= \rho(B_1[A+B_2]^{-1}B_2[A+B_1]^{-1}) \\ &\le \rho(B_1[A+B_2]^{-1})\rho(B_2[A+B_1]^{-1}),\end{aligned}$$

and the first statement of the conjecture would follow from Theorem 4.8. Moreover, if M_1 were symmetric, then the eigenvalues of $M_1^{-1}A$ would be real and would lie in the interval $[c_7h^{2/3}, 2-c_7h^{2/3}]$.

The conjecture is based upon the observation that, under the assumptions of Theorem 4.8, each of $B_1[A+B_2]^{-1}$, $B_2[A+B_1]^{-1}$, and M_1 is 'almost symmetric' in the interior of the grid Ω_h, by which we mean, for example, that

$$(B_1[A+B_2]^{-1}w)_{j,k} \sim ([A+B_2]^{-1}B_1w)_{j,k} \qquad (5.1)$$

whenever the grid-point (jh,kh) is not 'too close' to $\partial\Omega_h$. This follows from a simple calculation that shows that the matrices B_1B_2, AB_1, AB_2, DB_1, and DB_2 'almost commute' in the interior of the grid Ω_h. However, if (jh,kh) is 'close' to $\partial\Omega_h$, then (5.1) is a very poor approximation. Although it is possible to be more specific about what we mean by 'almost symmetric', this has not lead us to a more convincing justification of the conjecture. Therefore, we do not pursue this argument further at this time.

Conjecture 5.2: If the assumptions of Theorem 4.8 hold, then $\rho(I-S_1^{-1}A) \le 1-c_7h^{2/3}$ and the eigenvalues of $S_1^{-1}A$ lie in the interval $[c_7h^{2/3}, 2-c_7h^{2/3}]$.

Discussion: If C_1 and C_2 are normal matrices, then

$$\rho(C_1+C_2) \le \|C_1+C_2\| \le \|C_1\| + \|C_2\| = \rho(C_1) + \rho(C_2).$$

In addition, if Conjecture 5.1 holds, then $\rho(I-M_1^{-1}A) \le 1-c_7h^{2/3}$; the conjecture that $\rho(I-M_1^{-t}A) \le 1-c_7h^{2/3}$ can be defended in a similar manner. Hence, if $I-M_1^{-1}A$ and $I-M_1^{-t}A$ were normal, then

$$\rho(I-S_1^{-1}A) \le \tfrac{1}{2}\rho(I-M_1^{-1}A) + \tfrac{1}{2}\rho(I-M_1^{-t}A) \le 1-c_7h^{2/3}. \tag{5.2}$$

Furthermore, since S_1 is symmetric, the eigenvalues of $S_1^{-1}A$ are real. Hence, if (5.2) holds, then the eigenvalues of $S_1^{-1}A$ lie in the interval $[c_7h^{2/3}, 2-c_7h^{2/3}]$.

Although $I-M_1^{-1}A$ and $I-M_1^{-t}A$ are not in general normal, they are 'almost symmetric' in the interior of the grid Ω_h in the sense used in the discussion following Conjecture 5.1.

Theorem 5.3: If the assumptions of Theorem 4.8 hold and Conjecture 5.2 is valid, then, for the SAD-DKR factorization, the number of iterations of (1.5) required to reduce the A-norm of the initial error by a factor of ε is $O(h^{-2/3}\log\frac{1}{\varepsilon})$ and the associated number of arithmetic operations is $O(h^{-2\frac{2}{3}}\log\frac{1}{\varepsilon})$. Moreover, if the iteration (1.5) is accelerated by the Chebyshev or conjugate gradient techniques, then the number of iterations is decreased to $O(h^{-1/3}\log\frac{1}{\varepsilon})$ and the associated number of arithmetic operations is $O(h^{-2\frac{1}{3}}\log\frac{1}{\varepsilon})$.

Proof: If the assumptions of Theorem 4.8 hold and Conjecture 5.2 is valid, then $\rho = \rho(I-S_1^{-1}A) \leq 1-c_7h^{2/3}$. Moreover, $A^{1/2}S_1^{-1}A^{1/2}$ is normal, since S_1^{-1} is symmetric. Hence, by Lemma 3.1, the number of iterations of (1.5) required to reduce the A-norm of the initial error by a factor of ε is at most n+1, where

$$n = \log\frac{1}{\varepsilon} \;/\; \log\frac{1}{\rho} = O(h^{-2/3}\log\tfrac{1}{\varepsilon}).$$

Moreover, S_1 is symmetric and the eigenvalues of $S_1^{-1}A$ lie in the interval $[c_7h^{2/3}, 2-c_7h^{2/3}]$. Hence, if the iteration (1.5) is accelerated by the Chebyshev or conjugate gradient technique, then, by Lemmas 3.2 and 3.3, the number of iterations of (1.5) required to reduce the A-norm of the initial error by a factor of ε is at most n+1, where

$$n = \log\frac{2}{\varepsilon} \;/\; \log\frac{1}{r} = O(h^{-1/3}\log\tfrac{1}{\varepsilon}),$$

since, in this case, $a = 1-c_7h^{2/3}$ and

$$\frac{1}{r} = \frac{1+\sqrt{1-a^2}}{a} \geq 1+ch^{1/3}$$

for some positive constant c.

Since, for the SAD-DKR factorization, the number of multiplies needed to perform one iteration of (1.5) or its Chebyshev or conjugate gradient acceleration is proportional to the number of grid-points in the discretization, the number of multiplies per iteration is $O(h^{-2})$. Hence, the work estimates follow immediately from the bounds on the number of iterations. Q.E.D.

For the AD-DKR factorization, the work estimates are complicated slightly by the appearance of the constants c and q in equations (3.3) and (3.8) and the constant b in the expression for r (3.9). Clearly, these constants depend upon the matrices M_1 and A and, consequently, may grow as h->0. However, if they do not grow 'too fast' as h->0, a result similar to Theorem 5.3 holds for the AD-DKR factorization as well.

Theorem 5.4: If

1. the assumptions of Theorem 4.8 hold,

2. Conjecture 5.1 is valid, and

3. the constants c and q that appear in the inequality (3.3) satisfy $c = O(\varepsilon^{-k})$ and $q \leq Q$, for some constants k and Q independent of h,

then, for the AD-DKR factorization, the number of iterations of (1.5) required to reduce the A-norm of the initial error by a factor of ε is $O(h^{-2/3}\log\frac{1}{\varepsilon})$ and the associated number of arithmetic operations is $O(h^{-2\frac{2}{3}}\log\frac{1}{\varepsilon})$. Moreover, if the iteration (1.5) is accelerated by the Chebyshev technique and Assumption 3 is replaced by

$\tilde{3}$. the constants c and q that appear in the inequality (3.8) satisfy $c = O(\varepsilon^{-k})$ and $q \le Q$, for some constants k and Q independent of h, and

4. the constant b that appears in the expression for r (3.9) satisfies $b = O(h^{1/3})$,

then the number of iterations is decreased to $O(h^{-1/3}\log\frac{1}{\varepsilon})$ with the associated number of arithmetic operations being $O(h^{-2\frac{1}{3}}\log\frac{1}{\varepsilon})$.

Proof: If the assumptions of Theorem 4.8 hold and Conjecture 5.1 is valid, then

$$\rho = \rho(I-M_1^{-1}A) \le 1-c_7h^{2/3} \tag{5.3}$$

for some positive constant c_7. By Lemma 3.1, the number of iterations of (1.5) required to reduce the A-norm of the initial error by a factor of ε is at most n+1, where

$$(n-q)\ \log\frac{1}{\rho} - \log\binom{n}{q} = \log\frac{1}{\varepsilon} + \log c$$

Therefore, by Assumption 3, and (5.3), $n \le m$, where

$$(m-Q)c_7h^{2/3} - Q\log m = (k+1)\ \log\frac{C}{\varepsilon}.$$

for some constants Q, k, and C independent of h, whence $n = O(h^{-2/3}\log\frac{1}{\varepsilon})$.

By Assumptions 1, 2, and 4, the eigenvalues of $M_1^{-1}A$ lie in the ellipse

$$E = \{\ z \in C : z = 1 - a\cos\theta + i\,b\sin\theta,\ 0 \le \theta \le 2\pi\ \},$$

where $a = 1-c_7h^{2/3}$ and $b = O(h^{1/3})$, whence

$$\frac{1}{r} = \frac{1+\sqrt{1-a^2+b^2}}{a+b} \geq 1+ch^{1/3}$$

for some positive constant c. Therefore, by Assumption $\tilde{3}$, Lemma 3.2, and an argument similar to the one used above for the stationary iteration, if the iteration (1.5) is accelerated by the Chebyshev technique then the number of iterations required to reduce the initial error by a factor of ε is decreased to $O(h^{-1/3}\log\frac{1}{\varepsilon})$.

Since, for the AD-DKR factorization, the number of multiplies needed to perform one iteration of (1.5) or its Chebyshev acceleration is proportional to the number of grid-points in the discretization, the number of multiplies per iteration is $O(h^{-2})$. Hence, the work estimates follow immediately from the bounds on the number of iterations. Q.E.D.

We have not been able to establish the validity of Assumptions 3, $\tilde{3}$, and 4 for the AD-DKR factorization, although we believe that the violation of either Assumption 3 or $\tilde{3}$ is very unlikely in practice. On the other hand, the validity of Assumption 4 is questionable. For a few sample problems with coarse discretizations, we computed the eigenvalues of $M_1^{-1}A$ and found some of them to have small, but not insignificant, imaginary parts. However, the numerical results presented in the next section do not contradict the conclusion of Theorem 5.4, which lends support to our belief that the assumptions on which the theorem is based may be valid as well.

Finally, we re-emphasize that the class of problems of the form (1.2) to which our convergence results for the AD-DKR and SAD-DKR factorizations pertain is essentially the same as the class considered by Dupont, Kendall, and Rachford [5] for the DKR factorization, except for the added restriction that $a_1 = a_2$. Experimental results show that, if this restriction is violated, then the Alternating-Direction technique may not improve the rate of convergence of the iteration (1.5) or its acceleration. Furthermore, note that the parameters $\omega = 1$ and $\alpha_{j,k} = c_0 h^{4/3}$ recommended for use with the AD-DKR and SAD-DKR factorizations are substantially different from the corresponding parameters $\omega = O(h)$ and $\alpha_{j,k} = ch^2$ recommended by Dupont, Kendall, and Rachford [5] for the DKR factorization. Moreover, experimental evidence suggests that the AD-DKR and SAD-DKR factorizations do not achieve the substantially improved rates of convergence that we have observed if the parameters recommended for the DKR factorization are used. A more complete discussion of these observations is given in [3].

6. Numerical Results.

In this section, we present some numerical results that support the conjectures of the previous section.

For this experiment, we chose the Dirichlet problem with homogeneous boundary conditions for the two-dimensional elliptic equation (1.2) with coefficients

$$a_1(x,y) = a_2(x,y) = e^{xy}, \qquad q(x,y) = -1/(1+x+y)$$

on the L-shaped domain Ω having vertices $(0,0)$, $(1,0)$, $(1,\frac{1}{2})$, $(\frac{1}{2},\frac{1}{2})$, $(\frac{1}{2},1)$, (0,1). The domain was discretized with N+1 evenly spaced grid-lines in each direction; $h = \frac{1}{N}$. For N = 10, 20, 30,...,90, we discretized (1.2) using the standard five-point operator described in Section 2. We computed r, the right side of the resulting system of linear equations (1.1), so that the system has the solution

$$w_{j,k} = x_j(\tfrac{1}{2}-x_j)(1-x_j)y_k(\tfrac{1}{2}-y_k)(1-y_k),$$

where

$$x_j = jh \quad \text{and} \quad y_k = kh.$$

Starting from an initial guess of zero, we solved (1.1) by the iterative methods discussed in the previous section. Also included for comparison is the conjugate gradient acceleration of (1.5) based upon the DKR factorization. In each case, we recorded the number of iterations required to reduce the A-norm of the initial error by a factor of $\varepsilon = 10^{-5}$.

In Figure 6-1, the number of iterations required to solve (1.1) to the specified accuracy are listed for the methods

1. SIN, the stationary iteration (1.5) based upon the nonsymmetric AD-DKR factorization M_1 with $\alpha_{j,k} = h^{4/3}$ and iteration parameter $\omega = 1$,

2. SIS, the stationary iteration (1.5) based upon the symmetric SAD-DKR factorization S_1 with $\alpha_{j,k} = h^{4/3}$ and iteration parameter $\omega = 1$,

3. CHN, the Chebyshev acceleration of the stationary iteration (1.5) based upon the nonsymmetric AD-DKR factorization M_1 with $\alpha_{j,k} = h^{4/3}$

and iteration parameters chosen to minimize $P_n(z)$ on the interval $[h^{2/3}, 2-h^{2/3}]$,

4. CHS, the Chebyshev acceleration of the stationary iteration (1.5) based upon the symmetric SAD-DKR factorization S_1 with $\alpha_{j,k} = h^{4/3}$ and iteration parameters chosen to minimize $P_n(z)$ on the interval $[h^{2/3}, 2-h^{2/3}]$,

5. CGS, the conjugate gradient acceleration of the stationary iteration (1.5) based upon the symmetric SAD-DKR factorization S_1 with $\alpha_{j,k} = h^{4/3}$, and

6. CGDKR, the conjugate gradient acceleration of the stationary iteration (1.5) based upon the DKR factorization with $\alpha_{j,k} = h^2$.

For each method, both the modified (M) and unmodified (UM) DKR factorizations were used. Also listed in the last two lines of Figure 6-1 are the expected rate of convergence, E, and the observed rate, R, where R is computed by a least squares fit to

$$\log N = R \log (\text{NUMBER OF ITERATIONS}) + C$$

for N = 30,40,...,90.

For each of the methods, the numerical results for the modified and unmodified DKR factorizations are almost identical. Consequently, we have plotted the number of iterations for the methods based upon the unmodified DKR factorization only in Figures 6-2 and 6-3. The CGDKR method is included in each graph for comparison.

N	SIN		SIS		CHN		CHS		CGS		CGDKR	
	M	UM	M	UM	M	UM	M	UM	M	UM	M	UM
10	5	4	4	4	9	9	9	9	4	4	7	7
20	8	7	7	7	9	9	9	9	6	5	10	10
30	10	10	10	10	11	11	11	11	7	7	12	12
40	12	12	12	12	13	13	12	12	8	8	14	14
50	14	14	14	14	14	14	13	13	8	8	16	16
60	15	15	16	16	15	15	14	13	9	9	17	17
70	17	17	18	18	17	16	15	15	9	9	19	19
80	18	18	19	19	17	17	15	15	10	10	20	20
90	20	20	21	20	18	18	17	16	10	10	21	21
E	2/3	2/3	2/3	2/3	1/3	1/3	1/3	1/3	1/3	1/3	1/2	1/2
R	.614	.614	.677	.652	.444	.430	.373	.337	.325	.325	.513	.513

Figure 6-1: The number of iterations required to reduce the A-norm of the error by a factor of $\varepsilon = 10^{-5}$ for the stationary iteration (1.5) and its Chebyshev and conjugate gradient accelerations based upon the AD-DKR, SAD-DKR, and DKR factorizations.

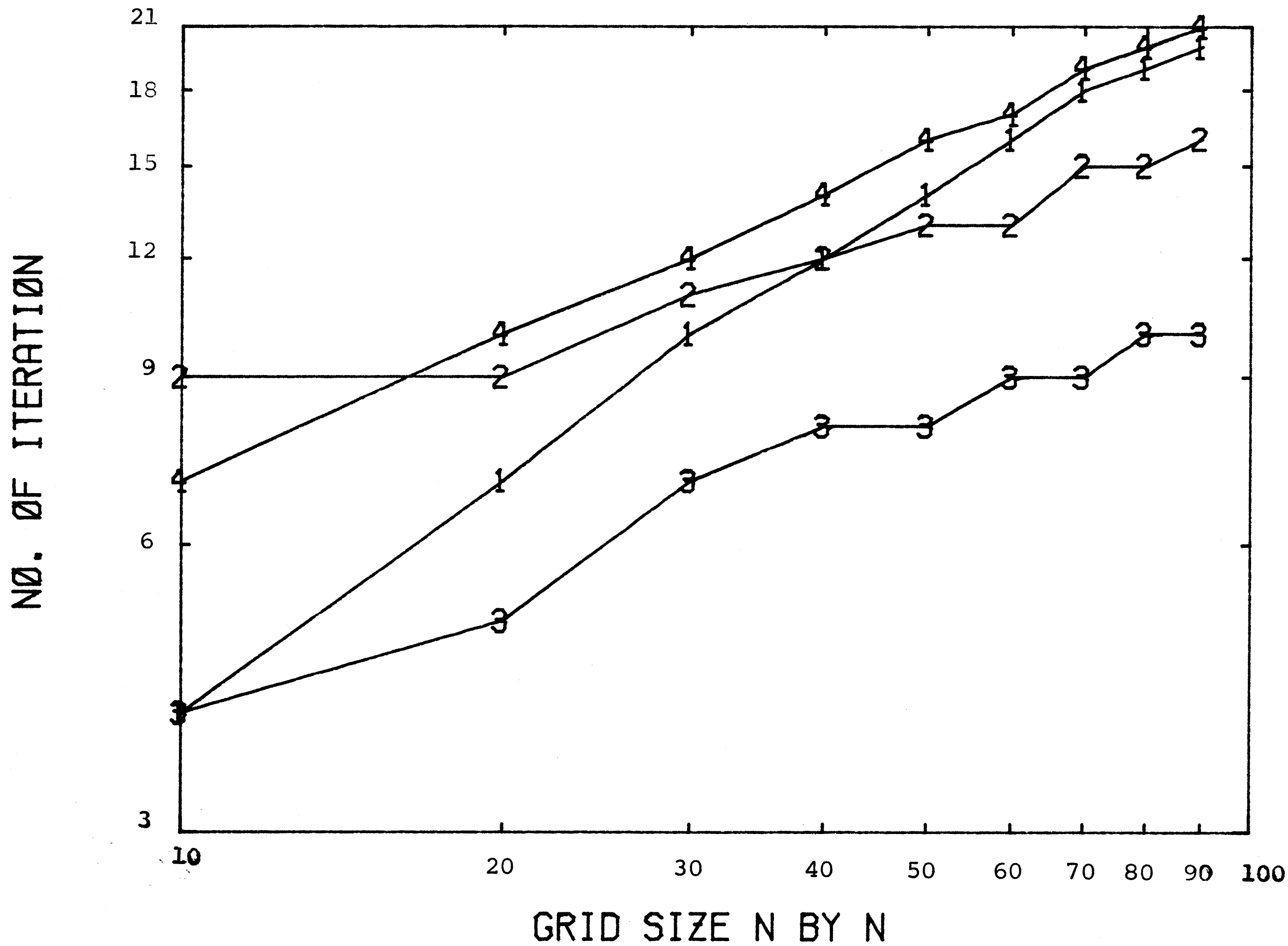

Figure 6-2: The number of iterations required by the methods SIS (1), CHS (2), CGS (3), and CGDKR (4) to reduce the A-norm of the error by a factor of $\varepsilon = 10^{-5}$.

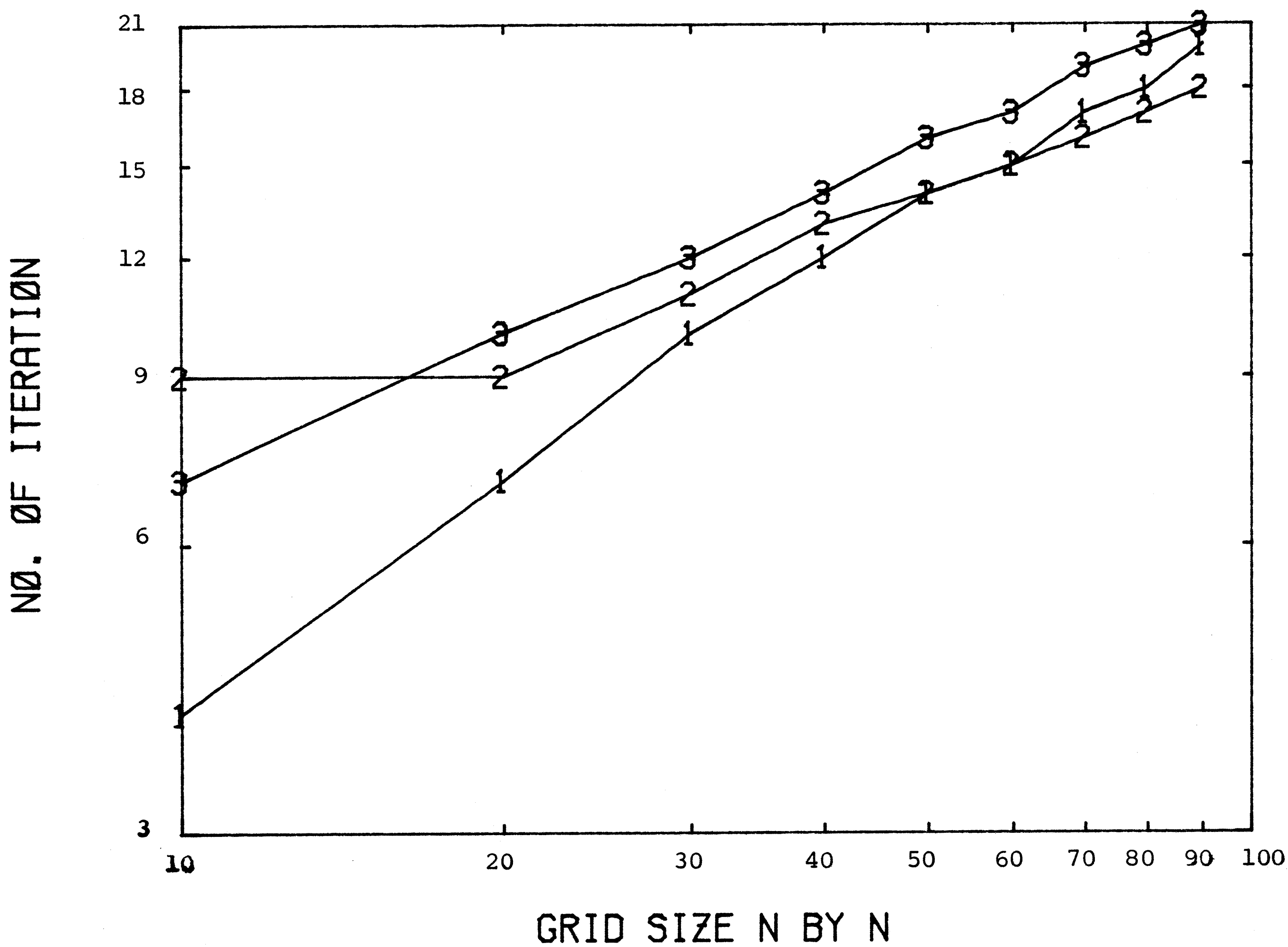

Figure 6-3: The number of iterations required by the methods SIN (1), CHN (2), and CGDKR (3) to reduce the A-norm of the error by a factor of $\varepsilon = 10^{-5}$.

The rate of convergence of the methods, with the possible exception of CHN, agrees very well with the rate predicted by the analysis in the previous section. The reason for the discrepancy for CHN is not clear to us, but it could be that Assumption 4 of Theorem 5.4 is violated or that the parameters that we chose for the Chebyshev iteration are not optimal. This question requires further investigation.

Although the principal aim of this paper is to present asymptotic work estimates for several ADIF methods and not to compare the efficiency of various algorithms for solving (1.1), we conclude with a few observations about the efficiency of CGS. Even on coarse grids, the number of iterations required to solve this test problem by CGS is about half the number required by CGDKR. Moreover, this ratio decreases with N, as the theory predicts. However, straightforward implementations of CGS and CGDKR require $16(N-1)^2$ and $44(N-1)^2$, respectively, multiply-adds per iteration. Hence, for these implementations, this problem, and the grids considered, CGDKR requires less work than CGS to solve the problem. But the relative efficiency of these two methods is problem dependent: for the Laplacian on a unit square with the same sequence of grids and implementations, we found that CGS requires slightly less work than CGDKR on the fine grids. In addition, Eisenstat [6] has shown that CGDKR can be implemented in $10(N-1)^2$ multiply-adds per iteration. Some of his techniques are applicable to CGS as well, and it is our hope that the work per iteration for this method can be significantly reduced. We intend to consider the question of efficient implementation of ADIF methods in [3], as well as the comparison of these methods with others.

Acknowledgments. We thank Howard Elman for several helpful discussions during the preparation of this paper and for instructing us in the use of his conjugate gradient package. In addition, we appreciate the encouragement of Stan Eisenstat and Martin Schultz and the pleasant working environment at Yale that they helped to provide for us.

References

[1] O. Axelsson. Solution of linear systems of equations: iterative methods. In V. A. Barker, Editor, Sparse Martix Techniques, Lecture Notes in Mathematics, 572, Springer-Verlag, New York, NY, 1977, pp. 1-51.

[2] N. I. Buleev. A numerical method for the solution of two-dimensional and three-dimensional equations of diffusion. Mat. Sb. 51:227-238, 1960.

[3] T. F. Chan, K. R. Jackson, and B. Zhu. Experimental results for alternating-direction incomplete factorizations. 1981. In preparation.

[4] T. Dupont. A factorization procedure for the solution of elliptic difference equations. SIAM J. Numer. Anal. 5:753-782, 1968.

[5] T. Dupont, R. P. Kendall, and H. H. Rachford, Jr. An approximate factorization procedure for solving self-adjoint elliptic difference equations. SIAM J. Numer. Anal. 5:559-573, 1968.

[6] S. C. Eisenstat. Efficient implementation of a class of preconditioned conjugate gradient methods. SIAM J. Sci. Stat. Comput. 2:1-4, 1981.

[7] I. Gustafsson. A class of first order factorization methods. BIT 18:142-156, 1978.

[8] T. A. Manteuffel. The Tchebychev iteration for nonsymmetric linear systems. Numer. Math. 28:307-327, 1977.

[9] T. A. Manteuffel. Shifted incomplete Cholesky Factorization. In I. S. Duff and G. W. Stewart, Editors, Sparse Matrix Proceedings, SIAM, Philadelphia, PA, 1978, pp. 41-61.

[10] T. A. Manteuffel. An incomplete factorization technique for positive definite linear systems. Math. Comp. 34:473-497, 1980.

[11] J. A. Meijerink and H. A. van der Vorst. An iterative solution method for linear systems of which the coefficient matrix is a symmetric M-matrix. Math. Comp. 31:148-162, 1977.

[12] A. Oliphant. An extrapolation process for solving linear systems. Quart. Appl. Math. 20:257-267, 1962.

[13] H. L. Stone. Iterative solution of implicit approximations of multidimensional partial differential equations. SIAM J. Numer. Anal. 5:530-558, 1968.

[14] R. S. Varga. Matrix Iterative Analysis. Prentice-Hall, Englewood Cliffs, NJ, 1962.