SIAM J. MATRIX ANAL. APPL.
Vol. 30, No. 4, pp. 1500–1519

ANALYSIS OF THE TRUNCATED SPIKE ALGORITHM*

CARL CHRISTIAN KJELGAARD MIKKELSEN† AND MURAT MANGUOGLU†

Abstract. The truncated SPIKE algorithm is a parallel solver for linear systems which are banded and strictly diagonally dominant by rows. There are machines for which the current implementation of the algorithm is faster and scales better than the corresponding solver in ScaLAPACK (PDDBTRF/PDDBTRS). In this paper we prove that the SPIKE matrix is strictly diagonally dominant by rows with a degree no less than the original matrix. We establish tight upper bounds on the decay rate of the spikes as well as the truncation error. We analyze the error of the method and present the results of some numerical experiments which show that the accuracy of the truncated SPIKE algorithm is comparable to LAPACK and ScaLAPACK.

Key words. direct methods, banded, and row diagonally dominant linear systems

AMS subject classification. 65F05

DOI. 10.1137/080719571

1. Introduction.

A matrix $A = [a_{ij}]$ is diagonally dominant by rows if

$$\sum_{i \neq j} |a_{ij}| \leq |a_{ii}| \tag{1.1}$$

for all i. If the inequality is sharp, then A is *strictly* diagonally dominant by rows.

The truncated SPIKE algorithm is a parallel solver for linear systems which are banded and strictly diagonally dominant by rows. Polizzi and Sameh demonstrated [10], [11] that there are parallel machines for which the algorithm is faster and scales better than the algorithm which is implemented in ScaLAPACK (PDDBTRF/PDDBTRS) [1]. We present the algorithm in section 2 and prove certain key properties of the truncated SPIKE algorithm in section 3. We analyze the error in section 4. We present the results of some experiments which supplement our theoretical analysis, and we compare the accuracy of the truncated SPIKE algorithm and ScaLAPACK in section 5.

The SPIKE algorithms are designed to solve banded systems on a parallel machine. The basic idea was introduced by Sameh and Kuck [12] who considered the tridiagonal case and Chen, Kuck, and Sameh [2] who studied the triangular case. Lawrie and Sameh [8] applied the algorithm to the symmetric positive definite systems, while Dongarra and Sameh [4] considered the strictly diagonally dominant case. Variations of the SPIKE algorithms for tridiagonal systems were introduced by Sun, Zhang, and Ni [13], who also analyzed the truncation error for tridiagonal systems which are evenly diagonally dominant. The truncation error for tridiagonal Toeplitz systems, which are also strictly diagonally dominant, as well as symmetric or skew symmetric was considered by Sun [14]. Another variation of the SPIKE algorithm for strictly diagonally dominant systems was studied by Larriba-Pey, Jorba, and Navarro

*Received by the editors March 31, 2008; accepted for publication (in revised form) by R.-C. Li August 25, 2008; published electronically December 3, 2008. This research is supported by the National Science Foundation (NSF-CCF-0635169), the Air Force Research Laboratory (FA8750-06-1-0233), and the Intel Corporation.

http://www.siam.org/journals/simax/30-4/71957.html

†Department of Computer Science, Purdue University, West Lafayette, IN 47907 (cmikkels@cs.purdue.edu, mmanguog@cs.purdue.edu).

[7]. Polizzi and Sameh have extended the SPIKE algorithms to the general banded case, and they developed the SPIKE package.

If A is nonsingular and diagonally dominant by rows, then the diagonal entries are nonzero and the dominance factor [5] ϵ is defined as follows:

$$\epsilon = \max_i \left\{ \frac{\sum_{i \neq j} |a_{ij}|}{|a_{ii}|} \right\}. \tag{1.2}$$

If $\epsilon > 0$, then the degree of diagonal dominance d is given by

$$d = \epsilon^{-1}. \tag{1.3}$$

The degree of diagonal dominance is central to the analysis of the truncated SPIKE algorithm.

2. The algorithm. Consider the nonsingular linear system

$$Ax = f,$$

where A is a n by n banded matrix which is strictly diagonally dominant by rows.

We assume that the number of superdiagonals k is equal to the number of subdiagonals and that the matrix is narrow banded, i.e., $k \ll n$. Let p denote the number of processors. We assume for simplicity that p divides n. Let the system be partitioned into the block diagonal form shown below

$$Ax = \begin{bmatrix} A_1 & \overline{B}_1 & & & \\ \overline{C}_2 & A_2 & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & \overline{B}_{p-1} \\ & & & \overline{C}_p & A_p \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ \vdots \\ x_p \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ \vdots \\ f_p \end{bmatrix}, \tag{2.1}$$

where $A_i, i = 1, 2, \ldots, p$ is a banded matrix of order $\mu = n/p$ and bandwidth $2k + 1$ (just like A),

$$\overline{B}_i = \begin{bmatrix} 0 & 0 \\ B_i & 0 \end{bmatrix}, \quad \text{and} \quad \overline{C}_{i+1} = \begin{bmatrix} 0 & C_{i+1} \\ 0 & 0 \end{bmatrix}, \quad i = 1, 2, \ldots p-1,$$

in which B_i and C_i are lower and upper triangular matrices, respectively, each of order k. Let D denote the main block diagonal D, i.e.,

$$D = \text{diag}\{A_1, A_2, \ldots, A_p\}.$$

The matrix D is nonsingular because A is strictly diagonally dominant. If we premultiply both sides of (2.1) by D^{-1}, we obtain a system $Sx = g$ of the form

$$\begin{bmatrix} I_\mu & \overline{V}_1 & & & \\ \overline{W}_2 & I_\mu & \overline{V}_2 & & \\ & \ddots & \ddots & \ddots & \\ & & \overline{W}_{p-1} & I_\mu & \overline{V}_{p-1} \\ & & & \overline{W}_p & I_\mu \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{p-1} \\ x_p \end{bmatrix} = \begin{bmatrix} g_1 \\ g_2 \\ \vdots \\ g_{p-1} \\ g_p \end{bmatrix}, \tag{2.2}$$

where

$$\overline{V}_i = \begin{bmatrix} V_i & 0 \end{bmatrix}, \quad \text{and} \quad \overline{W}_i = \begin{bmatrix} 0 & W_i \end{bmatrix},$$

in which V_i and W_i are matrices with k columns given by

$$V_i = A_i^{-1} \begin{bmatrix} 0 \\ B_i \end{bmatrix}, \quad \text{and} \quad W_i = A_i^{-1} \begin{bmatrix} C_i \\ 0 \end{bmatrix},$$

and will in general be full. Let the tall and narrow matrices V_i and W_i be partitioned, in turn, as follows

$$V_i = \begin{bmatrix} V_i^{(t)} \\ V_i^{(m)} \\ V_i^{(b)} \end{bmatrix}, \quad \text{and} \quad W_i = \begin{bmatrix} W_i^{(t)} \\ W_i^{(m)} \\ W_i^{(b)} \end{bmatrix},$$

where $V_i^{(t)}$, $V_i^{(b)}$, $W_i^{(t)}$, and $W_i^{(b)} \in \mathbb{R}^{k \times k}$. The superscripts t, m, and b are abbreviations of the words top, middle, and bottom, respectively. Let x_i and g_i be partitioned conformally, i.e.,

$$x_i = \begin{bmatrix} x_i^{(t)} \\ x_i^{(m)} \\ x_i^{(b)} \end{bmatrix}, \quad \text{and} \quad g_i = \begin{bmatrix} g_i^{(t)} \\ g_i^{(m)} \\ g_i^{(b)} \end{bmatrix}.$$

As an illustration we show the system (2.2) for $p = 3$, where $\nu = \mu - 2k$,

$$\begin{bmatrix} I_k & & & V_1^{(t)} & & & & & \\ & I_\nu & & V_1^{(m)} & & & & & \\ & & I_k & V_1^{(b)} & & & & & \\ & & W_2^{(t)} & I_k & & & V_2^{(t)} & & \\ & & W_2^{(m)} & & I_\nu & & V_2^{(m)} & & \\ & & W_2^{(b)} & & & I_k & V_2^{(b)} & & \\ & & & & & W_3^{(t)} & I_k & & \\ & & & & & W_3^{(t)} & & I_\nu & \\ & & & & & W_3^{(t)} & & & I_k \end{bmatrix} \begin{bmatrix} x_1^{(t)} \\ x_1^{(m)} \\ x_1^{(b)} \\ x_2^{(t)} \\ x_2^{(m)} \\ x_2^{(b)} \\ x_3^{(t)} \\ x_3^{(m)} \\ x_3^{(b)} \end{bmatrix} = \begin{bmatrix} g_1^{(t)} \\ g_1^{(m)} \\ g_1^{(b)} \\ g_2^{(t)} \\ g_2^{(m)} \\ g_2^{(b)} \\ g_3^{(t)} \\ g_3^{(m)} \\ g_3^{(b)} \end{bmatrix}.$$

It is from these narrow block columns or spikes extending from the main diagonal that the algorithm has derived its name. We will frequently refer to the V_i as the superdiagonal spikes and to the W_i as the subdiagonal spikes, and call the matrix S the SPIKE matrix.

Observe that the union of the k equations above and the k equations below the $p-1$ partition lines forms an independent subsystem of order $2k(p-1)$. We shall refer to this system as the "reduced" system and write $Rx_r = g_r$. It has the form

$$\begin{bmatrix} E_1 & F_1 & & & \\ G_2 & E_2 & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & F_{p-2} \\ & & & G_{p-2} & E_{p-1} \end{bmatrix} \begin{bmatrix} x_{r,1} \\ x_{r,2} \\ \vdots \\ x_{r,p-1} \end{bmatrix} = \begin{bmatrix} g_{r,1} \\ g_{r,2} \\ \vdots \\ g_{r,p-1} \end{bmatrix}, \tag{2.3}$$

where

$$E_i = \begin{bmatrix} I_k & V_i^{(b)} \\ W_{i+1}^{(t)} & I_k \end{bmatrix}, \quad F_i = \begin{bmatrix} 0 & 0 \\ 0 & V_{i+1}^{(t)} \end{bmatrix}, \quad \text{and} \quad G_i = \begin{bmatrix} W_i^{(b)} & 0 \\ 0 & 0 \end{bmatrix},$$

and

$$x_{r,i} = \begin{bmatrix} x_i^{(b)} \\ x_{i+1}^{(t)} \end{bmatrix}, \quad \text{and} \quad g_{r,i} = \begin{bmatrix} g_i^{(b)} \\ g_{i+1}^{(t)} \end{bmatrix}.$$

The subscript r is an abbreviation of the word "reduced". Dongarra and Sameh [4] noted that the reduced system is strictly diagonally dominant by rows and solved the reduced system using a parallel implementation of the Jacobi iteration. In Theorem 3.3 we show that the reduced system is strictly diagonally dominant by rows with a degree no less than the original matrix.

Once the reduced system has been solved

$$z_i = g_i - W_i x_{i-1}^{(b)} - V_i x_{i+1}^{(t)},$$

where x_0, x_{p+1}, W_1, and V_p are undefined and should be taken to zero in this equation. If the calculations are carried out using exact arithmetic, then z is the solution of $Ax = f$.

In general the reduced system is block tridiagonal. However, Polizzi and Sameh [10] noted that the off diagonal blocks are often negligible and can be dropped, yielding a truncated reduced system $Tx_{tr} = g_r$, which is block diagonal,

$$\begin{bmatrix} E_1 & & & & \\ & E_2 & & & \\ & & \ddots & & \\ & & & \ddots & \\ & & & & E_{p-1} \end{bmatrix} \begin{bmatrix} x_{tr,1} \\ x_{tr,2} \\ \vdots \\ x_{tr,p-1} \end{bmatrix} = \begin{bmatrix} g_{r,1} \\ g_{r,2} \\ \vdots \\ g_{r,p-1} \end{bmatrix}. \tag{2.4}$$

The subscript tr is an abbreviation of the words "truncated" and "reduced". In Theorem 3.8 we establish a tight upper bound on the size of the off diagonal blocks in terms of the degree of diagonal dominance of the original matrix and the size of the partitions. Polizzi and Sameh [10] showed that it is possible to compute the truncated reduced system without assembling the entire SPIKE system. Let $\mathcal{A}$ denote one of the diagonal blocks and consider the problem of computing the bottom $\mathcal{V}^{(b)}$ of the corresponding spike $\mathcal{V}$, given by

$$\mathcal{A}\mathcal{V} = \begin{bmatrix} 0 \\ \mathcal{B} \end{bmatrix}, \tag{2.5}$$

where $\mathcal{B}$ is a k by k dense matrix. It is not important here that $\mathcal{B}$ is lower triangular. We can exploit the remaining structure as follows. Let $\mathcal{A} = LU$ be the LU factorization of $\mathcal{A}$. Partition L and Y conformally with the right-hand side,

$$\begin{bmatrix} L_{11} & 0 \\ L_{21} & L_{22} \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} 0 \\ \mathcal{B} \end{bmatrix},$$

where L_{22} is a k by k lower unit triangular matrix. Since $L_{11}Y_1 = 0$, we have $Y_1 = 0$, and the problem reduces to solving $L_{22}Y_2 = \mathcal{B}$. Then we solve $U\mathcal{V} = Y$. Partition U and $\mathcal{V}$ conformally with Y and the right-hand side,

$$\begin{bmatrix} U_{11} & U_{12} \\ 0 & U_{22} \end{bmatrix} \begin{bmatrix} \mathcal{V}_1 \\ \mathcal{V}_2 \end{bmatrix} = \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix}.$$

Since U is upper triangular we can compute $\mathcal{V}^{(b)} = \mathcal{V}_2 = U_{22}^{-1}Y_2$ without computing $\mathcal{V}_1$. Similarly if $A_i = U_i'L_i'$ is a UL factorization of A_i, then it is possible to extract the top of the subdiagonal spikes without computing the entire spike.

Polizzi and Sameh [10] found experimentally that it is faster to extract the truncated reduced system using the LU/UL combinations than it is to compute the entire SPIKE matrix using the LU factorizations only. This is true on machines where arithmetic operations require much less time than memory references. The UL/LU strategy has greater data locality: computing the spikes is a BLAS2 operation, whereas computing the LU/UL factorizations is a BLAS3 operation.

The original equation is equivalent to

$$A_i x_i = f_i - C_i x_{i-1}^{(b)} - B_i x_{i+1}^{(t)}, \quad i = 1, 2, \dots p, \tag{2.6}$$

where $x_0^{(b)}$, $x_{p+1}^{(t)}$, C_1, and B_p are undefined and should be taken to zero in this equation.

These observations led to the truncated SPIKE algorithm by Polizzi and Sameh [10]. The algorithm consist of four stages.

Stage 1. Processor i computes the LU/UL factorizations

$$A_i = L_i U_i \quad \text{and} \quad A_i = U_i' L_i', \quad i = 1, 2, \dots p.$$

Stage 2. Processor i solves

$$A_i g_i = f_i, \quad i = 1, 2, \dots p,$$

using the LU factorization. Processor i computes $V_i^{(b)}$ using (L_i, U_i), $i = 1, 2, \dots p-1$. Processor i computes $W_i^{(t)}$ using (U_i', L_i') , $i = 2, 3 \dots p$.

Stage 3. Processor $i+1$ sends $W_{i+1}^{(t)}$ and $g_{i+1}^{(t)}$ to processor i, $i = 1, 2, \dots p-1$. Processor i solves one block of the truncated reduced system, specifically

$$\begin{bmatrix} I_k & V_i^{(b)} \\ W_{i+1}^{(t)} & I_k \end{bmatrix} \begin{bmatrix} x_i^{(b)} \\ x_{i+1}^{(t)} \end{bmatrix} = \begin{bmatrix} g_i^{(b)} \\ g_{i+1}^{(t)} \end{bmatrix}, \quad i = 1, 2, \dots p-1,$$

using Gaussian elimination without pivoting.

Stage 4. Processor i sends $x_i^{(b)}$ to processor $i+1$, for $i = 1, 2, \dots, p-1$, and processor i sends $x_i^{(t)}$ to processor $i-1$ for $i = 2, 3, \dots, p$. Then processor i solves

$$A_i y_i = f_i - C_i x_{i-1}^{(b)} - B_i x_{i+1}^{(t)}, \quad i = 1, 2, \dots, p$$

using the LU factorization, where $x_0^{(b)}$, $x_{p+1}^{(t)}$, C_1, and B_p are undefined and should be taken to zero in this equation. The vector y is an approximation of the solution to $Ax = f$.

3. The matrices S, R, and T. In this section we prove that the matrices S, R, and T in (2.2), (2.3), and (2.4) are strictly diagonally dominant by rows with degree no less than A, and we establish an upper bound on their condition number. The degree of diagonal dominance is defined by (1.3). We bound the truncation error, i.e., the difference between R and T, and show that all our bounds are tight.

The general estimates for the decay rates of the inverse of a banded matrix discovered by Demko, Moss, and Smith [3] are not suitable in our situation because it is necessary to exploit the relationship between the matrices and the right-hand sides which determine the spikes, in order to obtain estimates which are tight.

LEMMA 3.1. *Let $n \le m$ and let A be any n by m matrix which is strictly diagonally dominant by rows with degree $d > 1$. Let $A = LU$ be the LU factorization which is obtained by applying Gaussian elimination without pivoting to A. Then U is strictly diagonally dominant by rows with degree no less than d.*

Proof. Gaussian elimination produces a chain of matrices $A^{(j)}$, where the first $j-1$ columns of $A^{(j)}$ are lower triangular, $A = A^{(1)}$ and $A^{(n)} = U$. Due to the recursive nature of Gaussian elimination, it suffices to consider the transition from $A = A^{(1)}$ to $B = A^{(2)}$. Let $B = [b_{ij}]$. We must show the following equalities

$$|b_{kk}| \ge d \sum_{j \notin \{1,k\}} |b_{k,j}|, \quad k = 2, 3, \ldots, m.$$

Now, since $d \ge 1$ and $|a_{11}| \ge d \sum_{j=2}^{m} |a_{1j}|$ we have

$$\begin{aligned} |a_{kk}| &\ge d \sum_{j \ne k} |a_{kj}| \ge |a_{k1}| + d \sum_{j \notin \{1,k\}} |a_{kj}| \\ &\ge |a_{k1}| \frac{d \sum_{j=2}^{m} |a_{1j}|}{|a_{11}|} + d \sum_{j \notin \{1,k\}} |a_{kj}| \\ &\ge |a_{k1}| \frac{|a_{1k}|}{|a_{11}|} + d \sum_{j \notin \{1,k\}} \left(|a_{kj}| + \frac{|a_{k1}|}{|a_{11}|} |a_{1j}| \right). \end{aligned}$$

Now, since

$$b_{ij} = a_{ij} - \frac{a_{i1}}{a_{11}} a_{1j}, \quad i = 2, 3, \ldots, n, \quad j = 2, 3, \ldots, m,$$

the previous inequality implies

$$|b_{kk}| \ge |a_{kk}| - \frac{|a_{k1}|}{|a_{11}|} |a_{1k}| \ge d \sum_{j \notin \{1,k\}} \left(|a_{kj}| + \frac{|a_{k1}|}{|a_{11}|} |a_{1j}| \right) \ge d \sum_{j \notin \{1,k\}} |b_{kj}|. \qquad \square$$

COROLLARY 3.2. *Let A be an n by n matrix, and let F be an n by m matrix. If the matrix $\begin{bmatrix} A, & F \end{bmatrix}$ is strictly diagonally dominant by rows with degree $d > 1$, then the matrix $\begin{bmatrix} I, & A^{-1}F \end{bmatrix}$ is strictly diagonally dominant by rows with degree no less than d.*

Proof. We use Gaussian elimination without pivoting to reduce the n by $n+m$ matrix $\begin{bmatrix} A, & F \end{bmatrix}$ to upper triangular form, $U = [u_{ij}]$. By Lemma 3.1, U is diagonally dominant by rows with degree no less than d, and using back substitution we have a formula for the entries g_{ij} of the n by m matrix $G = A^{-1}F$, namely,

$$g_{n-t,j} = \frac{1}{u_{n-t,n-t}} \left(u_{n-t,n+j} - \sum_{s=n-t+1}^{n} u_{n-t,s} g_{s,j} \right),$$

for $j = 1, 2, \ldots m$ and $t = 0, 1, 2 \ldots n-1$. Let $\epsilon = d^{-1}$ and let $\Omega \subseteq \{0, 1, \ldots n-1\}$ be given by

$$t \in \Omega \Leftrightarrow \sum_{j=1}^{m} |g_{n-t,j}| \leq \epsilon.$$

We will prove that $\Omega = \{0, 1, \ldots n-1\}$. First, $0 \in \Omega$ because U is strictly diagonally dominant by rows with degree no less than d, and if $\{0, 1, 2, \ldots t-1\} \subset \Omega$ with $t < n$, then

$$\begin{aligned}\sum_{j=1}^{m} |g_{n-t,j}| &\leq \frac{1}{|u_{n-t,n-t}|} \sum_{j=1}^{m} \left(|u_{n-t,n+j}| + \sum_{s=n-t+1}^{n} |u_{n-t,s}||g_{s,j}| \right) \\ &= \frac{1}{|u_{n-t,n-t}|} \left(\sum_{j=1}^{m} |u_{n-t,n+j}| + \sum_{s=n-t+1}^{n} |u_{n-t,s}| \sum_{j=1}^{m} |g_{s,j}| \right) \\ &\leq \frac{1}{|u_{n-t,n-t}|} \left(\sum_{j=1}^{m} |u_{n-t,n+j}| + \sum_{s=n-t+1}^{n} |u_{n-t,s}|\epsilon \right) \leq \epsilon,\end{aligned}$$

which implies $t \in \Omega$. Therefore $\Omega = \{0, 1, 2 \ldots n-1\}$ and the proof is complete. ∎

THEOREM 3.3. *Let A be strictly diagonally dominant by rows with degree $d > 1$. Then the matrices S, R, and T are strictly diagonally dominant by rows with degree no less than d, specifically*

$$d \leq d(S) \leq d(R) \leq d(T),$$

with equality possible. The condition numbers share a common bound, namely

$$\max\{\, \kappa_\infty(S),\, \kappa_\infty(R),\, \kappa_\infty(T)\} \leq \frac{d+1}{d-1},$$

with the possibility of

$$\kappa_\infty(S) = \kappa_\infty(R) = \kappa_\infty(T) = \frac{d+1}{d-1}.$$

Proof. If S is strictly diagonally dominant by rows, then it is clear that T and R are strictly diagonally dominant by rows and $d(S) \leq d(R) \leq d(T)$. By applying Lemma 3.2 to the matrices $\begin{bmatrix} A_i, & F_i \end{bmatrix}$ where

$$F_1 = \begin{bmatrix} 0 \\ B_1 \end{bmatrix}, \quad F_i = \left[\begin{bmatrix} 0 \\ B_i \end{bmatrix}, \begin{bmatrix} C_i \\ 0 \end{bmatrix} \right], \; i = 2, \ldots p-1, \quad \text{and} \quad F_p = \begin{bmatrix} C_p \\ 0 \end{bmatrix},$$

we see that S is strictly diagonally dominant by rows with degree no less than d. Since $S_{ii} = 1$, we have $\|S - I\|_\infty \leq \epsilon < 1$ which allows us to treat S as a small perturbation of the identity matrix and estimate

$$\|S^{-1}\|_\infty \leq \frac{1}{1-\epsilon}, \quad \text{and} \quad \kappa_\infty(S) \leq \frac{1+\epsilon}{1-\epsilon} = \frac{d+1}{d-1},$$

and similarly for R and T.

It remains to be seen that our bounds are tight. To this end we consider a special case of the original problem, (2.1), where the diagonal blocks satisfy $A_i = I_\mu$ and the off-diagonal blocks are given by

$$\overline{B}_i = \begin{bmatrix} 0 & 0 \\ \epsilon J_k & 0 \end{bmatrix}, \quad \text{and} \quad \overline{C}_{i+1} = \begin{bmatrix} 0 & \epsilon J_k \\ 0 & 0 \end{bmatrix}, \quad i = 1, 2, \ldots p-1,$$

where J_k is the k by k antidiagonal identity matrix, and $\epsilon \in (0, 1)$. The matrix A is diagonally dominant by rows with degree $d = \epsilon^{-1}$. The upper and the lower bandwidths are equal to k. The main block diagonal is equal to the identity matrix, which implies $A = S$. The reduced system is block diagonal which implies $T = R$. It follows that

$$d(T) = d(R) = d(S) = d.$$

Computing S^{-1} reduces to inverting the 2 by 2 matrix $\left[\begin{smallmatrix} 1 & \epsilon \\ \epsilon & 1 \end{smallmatrix}\right]$. Direct computation establishes that

$$\kappa_\infty(T) = \kappa_\infty(R) = \kappa_\infty(S) = \frac{1+\epsilon}{1-\epsilon} = \frac{d+1}{d-1}. \qquad \square$$

We now study the truncation error, i.e., $\|R - T\|_\infty$. Let $\mathcal{A}$ denote one of the diagonal blocks of A, and let $\mathcal{V}$ be the corresponding superdiagonal spike given by (2.5). We are especially interested in the size of the elements at the top of the spike, i.e., the submatrix $\mathcal{V}^{(t)}$, which is given by

$$\mathcal{V}^{(t)} = \mathcal{V}(1:k, 1:k).$$

There is no loss of generality in limiting the analysis to the first diagonal block, rather there is a slight notational advantage, because the numbering of the elements of A and $\mathcal{A}$ coincide. We will use μ to denote the size of the first diagonal block.

We begin by estimating the size of the elements located in the bottom of $\mathcal{V}$; i.e., the submatrix $\mathcal{V}^{(b)}$ given by

$$\mathcal{V}^{(b)} = \mathcal{V}(\mu - k + 1 : \mu, 1 : k).$$

LEMMA 3.4. *Let A be strictly diagonally dominant by rows with degree $d > 1$. Let $\mathcal{V}$ be a superdiagonal spike. Then the submatrix $\mathcal{V}^{(b)}$ satisfies*

$$\|\mathcal{V}^{(b)}\|_\infty \leq \epsilon,$$

where $\epsilon = d^{-1}$.

Proof. Reduce the first μ by n block row to upper triangular form U. Since Gaussian elimination without pivoting preserves the upper bandwidth and does not decrease the degree of diagonal dominance, we have the following set of inequalities

$$\sum_{j=1}^{k} |u_{\mu-t,\mu-t+j}| \leq \epsilon |u_{\mu-t,\mu-t}|, \qquad t = 0, 1, \ldots k-1. \tag{3.1}$$

Our goal is to show that $\|\mathcal{V}^{(b)}\|_\infty \leq \epsilon$ or equivalently

$$\sum_{j=1}^{k} |v_{\mu-t,j}| \leq \epsilon, \quad t = 0, 1, \ldots k-1. \tag{3.2}$$

To this end we define the set $\Omega \subseteq \{0, 1, 2, \ldots, k-1\}$ by

$$t \in \Omega \Leftrightarrow \sum_{j=1}^{k} |v_{n-t,j}| \leq \epsilon.$$

We claim that $\Omega = \{0, 1, 2, \ldots, k-1\}$. Clearly $0 \in \Omega$, because

$$\sum_{j=1}^{k} |u_{\mu,\mu+j}| \leq \epsilon |u_{\mu,\mu}|, \quad \text{and} \quad v_{\mu,j} = \frac{u_{\mu,\mu+j}}{u_{\mu,\mu}}, \quad j = 1, 2, \ldots k.$$

Now suppose $\{0, 1, 2, \ldots t-1\} \subset \Omega$ with $t < k$. We wish to show that $t \in \Omega$. By back substitution we find that

$$v_{\mu-t,j} = \frac{1}{u_{\mu-t,\mu-t}} \left(u_{\mu-t,\mu+j} - \sum_{s=1}^{t} u_{\mu-t,\mu-t+s} v_{\mu-t+s,j} \right), \quad j = 1, 2, \ldots, (k-t),$$

and

$$v_{\mu-t,j} = -\frac{1}{u_{\mu-t,\mu-t}} \sum_{s=1}^{t} u_{\mu-t,\mu-t+s} v_{\mu-t+s,j}, \quad j = (k-t)+1, \ldots, k.$$

It follows that

$$\begin{aligned}
\sum_{j=1}^{k} |v_{\mu-t,j}| &\leq \frac{1}{|u_{\mu-t,\mu-t}|} \left(\sum_{j=1}^{k-t} |u_{\mu-t,\mu+j}| + \sum_{j=1}^{k} \sum_{s=1}^{t} |u_{\mu-t,\mu-t+s} v_{\mu-t+s,j}| \right) \\
&= \frac{1}{|u_{\mu-t,\mu-t}|} \left(\sum_{j=1}^{k-t} |u_{\mu-t,\mu+j}| + \sum_{s=1}^{t} |u_{\mu-t,\mu-t+s}| \sum_{j=1}^{k} |v_{\mu-t+s,j}| \right) \\
&\leq \frac{1}{|u_{\mu-t,\mu-t}|} \left(\sum_{j=1}^{k-t} |u_{\mu-t,\mu+j}| + \epsilon \sum_{s=1}^{t} |u_{\mu-t,\mu-t+s}| \right) \\
&\leq \frac{1}{|u_{\mu-t,\mu-t}|} \sum_{j=1}^{k} |u_{\mu-t,\mu-t+j}| \leq \epsilon,
\end{aligned}$$

which implies $t \in \Omega$. It follows that $\Omega = \{0, 1, 2, \ldots k-1\}$ and $\|\mathcal{V}^{(b)}\|_\infty \leq \epsilon$. ∎

We continue with the following lemma which relates the size of the elements in a specific row of $\mathcal{V}$ to the infinity norm of the k by k submatrix which lies directly below the row.

LEMMA 3.5. *Let μ denote the dimension of the diagonal block $\mathcal{A}$ and let $i \geq \mu - k$. Then*

$$\sum_{j=1}^{k} |v_{i,j}| \leq \epsilon \|\mathcal{V}(i+1 : i+k, 1 : k)\|_\infty.$$

Proof. We have

$$\mathcal{V} = \mathcal{A}^{-1} \begin{bmatrix} 0 \\ \mathcal{B} \end{bmatrix}$$

for the appropriate k by k matrix $\mathcal{B}$. We use Gaussian elimination without pivoting to reduce the matrix

$$\begin{bmatrix} \mathcal{A}, & \begin{bmatrix} 0 \\ \mathcal{B} \end{bmatrix} \end{bmatrix},$$

to upper triangular form $U = [u_{ij}]$. By Lemma 3.2 U is strictly diagonally dominant by rows with degree no less than d. Since the original matrix A was banded and no pivoting was applied, it follows that $u_{ij} = 0$ for all i and j such that $j > \mu$ and $i \geq \mu - k$. It follows by back substitution that

$$v_{i,j} = -\frac{1}{u_{i,i}} \sum_{s=i+1}^{i+k} u_{i,s} v_{s,j},$$

which implies

$$\sum_{j=1}^{k} |v_{i,j}| \leq \frac{1}{|u_{i,i}|} \sum_{s=i+1}^{i+k} |u_{i,s}| \sum_{j=1}^{k} |v_{s,j}|.$$

By definition

$$\max_{s=i+1,\ldots i+k} \sum_{j=1}^{k} |v_{s,j}| = \|\mathcal{V}(i+1:i+k, 1:k)\|_\infty,$$

and since U is strictly diagonally dominant by rows with degree no less than d, we have

$$\frac{1}{|u_{i,i}|} \sum_{s=i+1}^{i+k} |u_{i,s}| \leq \epsilon,$$

which completes the proof. ∎

The following corollary is an immediate consequence.

COROLLARY 3.6. *Let $\mathcal{V}'$ and $\mathcal{V}''$ be two k by k submatrices of the superdiagonal spike $\mathcal{V}$, such that $\mathcal{V}'$ lies directly on top of $\mathcal{V}''$. Then*

$$\|\mathcal{V}'\|_\infty \leq \epsilon \|\mathcal{V}''\|_\infty.$$

This corollary establishes a chain of inequalities leading from the bottom to the top of the spike which together with Lemma 3.4 implies the following theorem.

THEOREM 3.7. *Let d denote the degree of diagonal dominance of A, let μ denote the dimension of one of the diagonal blocks, and $q = \lfloor \mu/k \rfloor$ is the largest integer less than or equal to μ/k. The top of the corresponding superdiagonal spike $\mathcal{V}$ satisfies the inequality*

$$\|\mathcal{V}^{(t)}\|_\infty \leq \epsilon^q.$$

Is this estimate for the decay rate of the spikes tight or not? Let $\epsilon \in (0, 1)$ and consider the upper triangular matrix A given by $a_{ii} = 1$, $a_{ij} = \epsilon$ for $i = j - k$, and $a_{ij} = 0$ in all other cases. Now consider a partition of a certain size μ. Write $\mu = qk + r$, where $q = \lfloor \mu/k \rfloor$, and the remainder r satisfies $0 \leq r < k$. If $r > 0$, then by back substitution we find that the corresponding spike is given by

$$\mathcal{V} = \begin{bmatrix} \mathcal{V}_{q+1}^T & \mathcal{V}_q^T & \ldots & \mathcal{V}_1^T \end{bmatrix}^T,$$

where

$$V_j = (-1)^{j-1}\epsilon^j I_k \quad \text{for } j = 1, 2, \ldots, q,$$

and $V_{q+1} = (-1)^q \epsilon^{q+1} E_r$, where I_k is the k by k identity matrix and E_r consists of the last r rows of I_k. If $r = 0$, then the term V_{q+1} does not appear. Regardless of the value of the remainder r, we have

$$\left\|\mathcal{V}^{(t)}\right\|_\infty = \epsilon^q.$$

In short, if we limit ourselves to matrices A which are strictly diagonally dominant by rows with degree d and upper bandwidth k, then the estimate given in Theorem 3.7 is tight.

The following theorem is an immediate consequence of Theorem 3.7.

THEOREM 3.8. *Let A be a n by n narrow banded matrix with upper and lower bandwidth k, and strictly diagonally dominant by rows with degree d. Then the truncation error satisfies*

$$\|R - T\|_\infty \leq \max_{i=1,\ldots p} d^{-q_i},$$

where $q_i = \lfloor \mu_i / k \rfloor$, and μ_i is the size of the ith partition.

A better bound exists in the special case in which A is a tridiagonal, evenly diagonally dominant matrix [13], or when A is a tridiagonal Toeplitz matrix, which is also strictly diagonally dominant, as well as symmetric or skew symmetric [14].

Now, consider for the sake of simplicity, the case when the partitions have the same size μ. Then Theorem 3.8 reduces to the statement

$$\|R - T\|_\infty \leq d^{-q},$$

where $q = \lfloor \mu / k \rfloor$. Let S_T denote the matrix obtained by eliminating the tips of the spikes from the spike matrix S. Then the reduced system matrix for S_T is equal to T. The truncation error effectively replaces A with the matrix $A_T = DS_T$, for which we have

$$\|A - A_T\|_\infty \leq \|D\|_\infty \|S - S_T\|_\infty \leq \|A\|_\infty \|R - T\|_\infty \leq d^{-q}\|A\|_\infty, \tag{3.3}$$

or equivalently $A_T = A + \Delta A$, where $\|\Delta A\|_\infty \leq d^{-q}\|A\|_\infty$. We see that the effect of the truncation is to introduce a normwise relative backward error which is bounded by d^{-q}.

We have already seen that the estimate of Theorem 3.8 is tight, but which matrices exhibit the slowest possible decay rate? We can answer this question for tridiagonal matrices.

THEOREM 3.9. *Let $\{(a_i, b_i, c_i)\}_{i=1}^n$ be a finite sequence, such that $a_i \neq 0$, and*

$$\max_{i=1,\ldots,n} \frac{|b_i| + |c_i|}{|a_i|} = \epsilon < 1.$$

If the vector $x = (x_1, x_2, \ldots, x_n)^T$ given by

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} a_1 & b_1 & & \\ c_2 & \ddots & \ddots & \\ & \ddots & & b_{n-1} \\ & & c_n & a_n \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ \vdots \\ 0 \\ b_n \end{bmatrix}$$

exhibits the smallest possible decay rate, i.e., if $|x_1| = \epsilon^n, \epsilon = d^{-1}$, *then* $c_i = 0$, *for* $i = 1, 2, 3, \ldots, n$ *and* $|b_i| = \epsilon |a_i|$ *for* $i = 1, 2, \ldots n$.

Proof. Mikkelsen [9] gives a direct proof. □

4. Error analysis. In this section we do an error analysis of the truncated SPIKE algorithm. We begin by deriving a few results on Gaussian elimination for systems which are strictly diagonally dominant by rows with degree $d > 1$.

We assume that the original problem has been scaled by rows such that $a_{ii} = 1$. Such a scaling preserves the degree of diagonal dominance, and allows us to estimate

$$\|A_i\|_\infty \leq 1 + d^{-1}, \quad \|A_i^{-1}\|_\infty \leq \frac{1}{1 - d^{-1}}, \quad \text{and} \quad \kappa_\infty(A_i) \leq \frac{d+1}{d-1}.$$

Let u denote the unit roundoff error on the machine, and following Higham [6], we define

$$\gamma_j = \frac{ju}{1 - ju}, \tag{4.1}$$

when $ju < 1$. If A is any matrix, then $B = |A|$ is the matrix given by $b_{ij} = |a_{ij}|$. If A, B are matrices of the same dimension, then we write $A \leq B$, if $a_{ij} \leq b_{ij}$ for all i and j.

If A is a banded matrix with upper and lower bandwidth k, which is diagonally dominant by rows, and if Gaussian elimination runs to completion, then the computed solution $\hat{x}$ to $Ax = f$ satisfies

$$(A + \Delta A)\hat{x} = f, \quad |\Delta A| \leq \gamma_{3k+2} |\hat{L}||\hat{U}|,$$

where $\hat{L}$ and $\hat{U}$ are the computed LU factors.

Now, how large is $\|\Delta A\|_\infty$? If A is any n by n matrix and if $A = LU$ is the exact LU factorization, then

$$|L||U| = |AU^{-1}||U| \leq |A||U^{-1}||U|.$$

If U is diagonally dominant by rows, then by Lemma 8.8 [6]

$$\||U^{-1}||U|\|_\infty \leq (2n - 1). \tag{4.2}$$

This estimate is tight. However, if A is strictly diagonally dominant by rows with degree $d > 1$, then we may be able to improve upon it. By Theorem 3.1 U is strictly diagonally dominant by rows with degree no less than d. Write $U = DV$, where D is the main diagonal of U, then

$$|U^{-1}||U| = |V^{-1}D^{-1}||DV| = |V^{-1}||V|,$$

which allows us to estimate

$$\||U^{-1}||U|\|_\infty \leq \frac{d+1}{d-1}, \tag{4.3}$$

because $\|I - V\|_\infty \leq d^{-1} < 1$.

It is important to realize that neither (4.2) nor (4.3) need apply to the computed LU factorization, because, while $\hat{L}\hat{U}$ is the exact LU factorization of the matrix $A + \Delta A$, this matrix need not be diagonally dominant! However, since $\hat{L} \to L$, and $\hat{U} \to U$

as $u \to 0$, then $A + \Delta A$ will be strictly diagonally dominant by rows with degree close to d, for u sufficiently small, and then we may estimate

$$\|\Delta A\|_\infty \leq \gamma_{3k+2} \| |\hat{L}||\hat{U}| \|_\infty \lesssim \gamma_{3k+2} \frac{d+1}{d-1} \|A\|_\infty.$$

In the following we assume that we may estimate

$$\|\Delta A\|_\infty \leq \gamma_{3k+2} \frac{d+1}{d-1} \|A\|_\infty.$$

Now, what can be said about the solution $\hat{X}$ to the equation $AX = F$ where X and F have m columns? We have

$$(A + \Delta A_j)\hat{x}_j = f_j, \quad |\Delta A_j| \leq \gamma_{3k+2} |\hat{L}||\hat{U}|, \quad j = 1, 2, \dots m,$$

where the perturbations ΔA_j depend on j, but share a common bound which is independent of j. Now, if the unit roundoff error is sufficiently small, specifically if

$$\alpha = \gamma_{3k+2} \left(\frac{d+1}{d-1} \right)^2 < 1, \tag{4.4}$$

then $I + A^{-1}\Delta A_j$ and $I + \Delta A_j A^{-1}$ are both invertible and we may write

$$\hat{x}_j = \sum_{i=0}^{\infty} \left(-A^{-1}\Delta A_j \right)^i x_j = A^{-1} \sum_{i=0}^{\infty} \left(-(\Delta A_j)A^{-1} \right)^i f_j,$$

from which it follows immediately that

$$|\hat{x}_j - x_j| \leq E_1 |x_j|, \quad E_1 = \sum_{i=1}^{\infty} \left(\gamma_{3k+2} |A^{-1}||\hat{L}||\hat{U}| \right)^i,$$

$$|A\hat{x}_j - f_j| \leq E_2 |f_j|, \quad E_2 = \sum_{i=1}^{\infty} \left(\gamma_{3k+2} |\hat{L}||\hat{U}||A^{-1}| \right)^i,$$

which implies

$$|\hat{X} - X| \leq E_1 |X|, \quad \text{and} \quad |A\hat{X} - F| \leq E_2 |F|.$$

The two operators, E_1 and E_2, share a common bound, namely,

$$\|E_1\|_\infty \leq \frac{\alpha}{1-\alpha}, \quad \text{and} \quad \|E_2\|_\infty \leq \frac{\alpha}{1-\alpha},$$

where α is defined by (4.4). It follows that

$$\|\hat{X} - X\|_\infty \leq \frac{\alpha}{1-\alpha} \|X\|_\infty, \quad \text{and} \quad \|A\hat{X} - F\|_\infty \leq \frac{\alpha}{1-\alpha} \|F\|_\infty. \tag{4.5}$$

Stage 1 Each matrix A_i has dimension μ and is strictly diagonally dominant by rows. The computed LU factorization satisfies

$$A_i + \Delta A_i = \hat{L}_i \hat{U}_i, \quad |\Delta A_i| \leq \gamma_{k+1} \left| \hat{L}_i \right| \left| \hat{U}_i \right|,$$

where

$$\left\|\left|\hat{L}_i\right|\left|\hat{U}_i\right|\right\|_\infty \lesssim \frac{d+1}{d-1}\|A_i\|_\infty,$$

when the unit roundoff error u is sufficiently small. We have the same type of estimate for the computed UL factorizations.

Stage 2 In the truncated SPIKE algorithm, we do not compute the entire SPIKE matrix but stop substituting as soon as the truncated reduced system matrix has been computed. However, in order to estimate the error, it is convenient to consider the computation of the entire SPIKE matrix S.

By applying (4.5) repeatedly to the individual block rows we find

$$\|\hat{S}-S\|_\infty \le \frac{2\alpha}{1-\alpha}\|S-I\|_\infty, \quad \left\|D\hat{S}-A\right\|_\infty \le \frac{2\alpha}{1-\alpha}\|A-D\|_\infty.$$

The extra factor of 2 is introduced because we have to treat the superdiagonal and the subdiagonal spikes separately.

Similarly we find for the computation of the modified right-hand side that

$$\|\hat{g}-g\|_\infty \le \frac{\alpha}{1-\alpha}\|g\|_\infty, \quad \text{and} \quad \left\|D\hat{g}-f\right\|_\infty \le \frac{\alpha}{1-\alpha}\|f\|_\infty.$$

It is clear that since $\hat{T}-T$ is a submatrix of $\hat{S}-S$ we have

$$\left\|\hat{T}-T\right\|_\infty \le \left\|\hat{S}-S\right\|_\infty \le \frac{2\alpha}{1-\alpha}\|S-I\|_\infty \le \frac{2\alpha}{1-\alpha}d^{-1}.$$

Stage 3 By Theorem 3.8 the truncated reduced system is a good approximation of the reduced system if d is not too close to 1 and if the partitions are not too small. By Theorem 3.3 the truncated reduced system is strictly diagonally dominant by rows with a degree no less than the original system. It consists of $p-1$ independent systems which are of dimension $2k$. By Theorem 9.3 [6] it follows that if Gaussian elimination runs to completion, then the computed solution $\hat{x}_{tr}$ of the computed truncated reduced system $\hat{T}x_{tr}=\hat{g}_r$ satisfies

$$\left(\hat{T}+\Delta\hat{T}\right)\hat{x}_{tr}=\hat{g}_r, \quad \left|\Delta\hat{T}\right| \le \gamma_{6k}\left|\hat{L}_t\right|\left|\hat{U}_t\right|,$$

where $\hat{L}_t\hat{U}_t$ is the computed LU factorization of the computed truncated reduced system matrix $\hat{T}$. It follows that

$$\|\hat{x}_{tr}-x_{tr}\|_\infty \le \frac{\beta}{1-\beta}\|x_{tr}\|_\infty \quad \text{and} \quad \left\|\hat{T}\hat{x}_{tr}-\hat{g}_r\right\| \le \frac{\beta}{1-\beta}\|\hat{g}_r\|_\infty,$$

provided the unit round off error is so small that

$$\beta=\gamma_{6k}\left(\frac{d+1}{d-1}\right)^2<1.$$

Stage 4 Adjusting the original right-hand side, i.e., computing

$$h_i=f_i-C_ix_{i-1}^{(b)}-B_ix_{i+1}^{(t)},$$

introduces a small forward error. Notice that C_i affects only the top of f_i and B_i affects only the bottom of f_i. The componentwise relative forward error is no more than

$$|\hat{h}_i-h_i| \le \gamma_{k+1}\left(|f_i|+|C_i||x_{i-1}^{(b)}|+|B_i||x_{i+1}^{(t)}|\right),$$

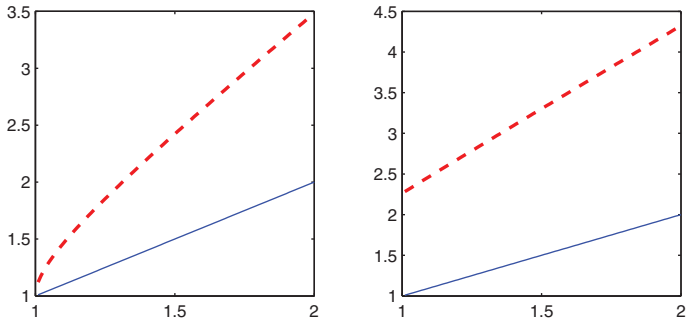

FIG. 5.1. *The degree of diagonal dominance for the matrix $S^{(k)}$ as a function of the degree of diagonal dominance of the original matrices: $A^{(k)}$ (left), and $B^{(k)}$ (right). The matrices are defined by equation* (5.1). *The red dotted line is the experimental result and the solid blue line is the theoretical lower bound.*

regardless of the order in which the scalar products are evaluated. This is an overestimate which does not take into account that the central components of f_i are not changed at all. The solution of the final set of linear equations is identical to stage 2 and generates a normwise relative residual of at most $\frac{\alpha}{1-\alpha}$, as well as a normwise relative forward error of at most $\frac{\alpha}{1-\alpha}$; cf. (4.5).

In short, if d is not too close to 1 and if the partitions are not too small, then the errors at every stage of the algorithm are small. We found that the simplest way to evaluate the overall error was to calculate the residual and estimate

$$\|x-y\|_\infty \leq \|A^{-1}\|_\infty \|f-Ay\|_\infty \leq \frac{1}{1-d^{-1}}\|f-Ay\|_\infty,$$

which turned out to be fairly effective as long as d is not too close to 1.

5. Numerical experiments. We ran experiments to verify the main results of this paper as well as compare the accuracy of the truncated SPIKE algorithm with the algorithm implemented in ScaLAPACK.

5.1. The matrices S, R, and T. We wanted to verify that the degree of diagonal dominance of the SPIKE matrix S was no less than that of the original matrix A. We selected two sequences of matrices with $(n, k_u, k_l) = (10^6, 5, 5)$:

$$A_{ij}^{(k)} = \begin{cases} 1+0.01k & \text{for } i=j \\ -0.1 & \text{for } 0<|i-j|\leq 5, \\ 0 & \text{otherwise} \end{cases} \qquad B_{ij}^{(k)} = \begin{cases} 1+0.01k & \text{for } i=j \\ 0.1 & \text{for } 0<|i-j|\leq 5, \\ 0 & \text{otherwise} \end{cases} \tag{5.1}$$

for $k = 1, 2, \ldots, 100$. We selected $p = 8$ partitions and a uniform block size of $1.25\cdot 10^5$. We explicitly computed the entire SPIKE matrix S and the excess $\|S-I\|_\infty$ for each of these 200 matrices, from which we determined the degree of diagonal dominance as $d(S) = 1/\|S-I\|_\infty$. Our results are displayed in Figure 5.1. We found that not only is the degree of diagonal dominance preserved, i.e., $d(S) \geq d(A)$, but there can be a substantial increase in diagonal dominance as well.

We extracted the truncated reduced system matrix T from each of the 200 matrices and computed the condition number in the infinity norm by explicitly inverting T and calculating $\|T^{-1}\|_\infty$. We then plotted the condition number of T as a function of the degree of diagonal dominance of A. The results are displayed in Figure 5.2.

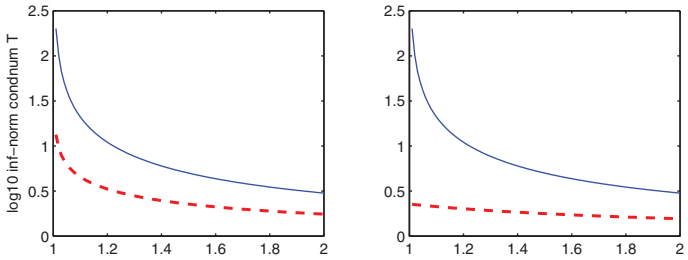

FIG. 5.2. *The condition number of the truncated reduced system as a function of the degree of diagonal dominance of the original system matrices: $A^{(k)}$ (left), and $B^{(k)}$ (right). The matrices are defined by equation* (5.1). *The dotted red line is the experimental result and the solid blue line is the theoretical upper bound.*

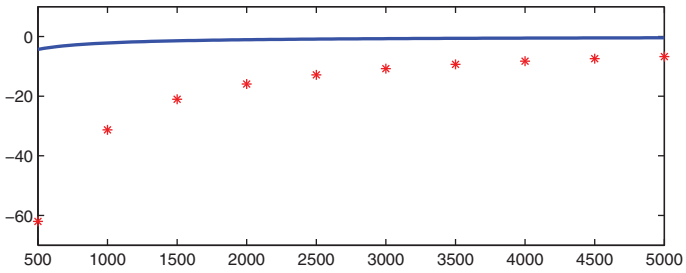

FIG. 5.3. *The infinity norm of the truncation error as a function of the number of partitions. Solid blue line is the theoretical upper bound, while red dots are experimental results. The matrix has* $d = 1.01$ *and is tridiagonal.*

The theoretical upper bound is given by $\frac{d+1}{d-1}$ where $d = d(A)$ is the degree of diagonal dominance of A. We found that the truncated reduced system was even better conditioned than expected.

We wanted to investigate the size of the truncation error as a function of the degree of diagonal dominance of the original matrix A and the number of partitions p. We selected a tridiagonal Toeplitz matrix with $n = 5 \cdot 10^5$ and 1.01 on the main diagonal and 0.5 on the off-diagonal elements. We choose $p = 500j$, for $j = 1, 2, \ldots, 10$ and computed the truncation error explicitly. The theoretical upper bound is given by d^{-q} where $d = 1.01$ and $q = \lfloor 5 \cdot 10^5/p \rfloor$. The results are displayed in Figure 5.3. The truncation error is much smaller than the theoretical upper bound and it is smaller than the unit roundoff error $u = 2^{-53} \approx 1.1 \cdot 10^{-16}$ as long as $p \leq 2000$.

5.2. The error analysis. We wanted to verify the bounds presented in section 4. We constructed matrices which were diagonally dominant by rows with different degrees and ran them through our implementation of the truncated SPIKE algorithm. The matrices all had $(n, k_l, k_u) = (10^6, 10, 10)$ with every diagonal entry equal to 1. The nonzero, off-diagonal entries where positive and constant for each matrix, such that the degree of diagonal dominance varied from 1.1 for the first matrix to 2.0 for the last matrix, with steps of 0.1. The right-hand side side was generated from the solution which was selected as $x = (1, 1, \ldots, 1)^T$. Our results are listed as Table 5.1 and Table 5.2. The bounds were computed as follows:

1. The modified right-hand side,

$$\|D\hat{g} - f\|_\infty \leq \frac{\alpha}{1-\alpha}\|f\|_\infty.$$

TABLE 5.1
A comparison of certain measurable quantities and their bounds for 10 *different matrices distinguished by their degree of diagonal dominance.*

d	α	$\|D\hat{g} - f\|_\infty$		$\left\|D\hat{S} - A\right\|_\infty$	
		measured	bound	measured	bound
1.1	1.57e − 12	9.58e − 16	2.99e − 12	6.94e − 17	2.85e − 12
1.2	4.30e − 13	1.25e − 15	7.88e − 13	5.90e − 17	7.16e − 13
1.3	2.09e − 13	1.57e − 15	3.69e − 13	7.05e − 17	3.21e − 13
1.4	1.28e − 13	1.51e − 15	2.19e − 13	6.94e − 17	1.83e − 13
1.5	8.88e − 14	1.64e − 15	1.48e − 13	5.11e − 17	1.18e − 13
1.6	6.67e − 14	9.78e − 16	1.08e − 13	5.55e − 17	8.34e − 14
1.7	5.29e − 14	1.51e − 15	8.39e − 14	2.93e − 17	6.22e − 14
1.8	4.35e − 14	1.47e − 15	6.77e − 14	4.47e − 17	4.84e − 14
1.9	3.69e − 14	1.75e − 15	5.63e − 14	4.27e − 17	3.88e − 14
2.0	3.20e − 14	1.30e − 15	4.80e − 14	4.16e − 17	3.20e − 14

TABLE 5.2
A comparison of certain measurable quantities and their bounds for 10 *different matrices distinguished by their degree of diagonal dominance.*

d	$\left\|\hat{T}\hat{x}_{tr} - g_r\right\|_\infty$		$\|A\hat{x} - f\|_\infty$	$\|\hat{x} - x\|_\infty$	
	measured	bound	measured	measured	bound
1.1	7.77e − 16	1.40e − 13	8.88e − 16	8.88e − 16	9.77e − 15
1.2	7.77e − 16	7.33e − 14	1.33e − 15	1.11e − 15	7.99e − 15
1.3	5.55e − 16	5.11e − 14	1.55e − 15	1.55e − 15	5.44e − 15
1.5	4.44e − 16	3.33e − 14	1.55e − 15	1.55e − 15	4.66e − 15
1.6	7.77e − 16	2.89e − 14	8.88e − 16	8.88e − 16	2.37e − 15
1.7	5.55e − 16	2.57e − 14	1.55e − 15	1.22e − 15	3.77e − 15
1.8	5.55e − 16	2.33e − 14	1.55e − 15	1.55e − 15	3.50e − 15
1.9	6.66e − 16	2.15e − 14	1.78e − 15	1.55e − 15	3.75e − 15
2.0	5.55e − 16	2.00e − 14	1.33e − 15	1.33e − 15	2.66e − 15

2. The SPIKE matrix,

$$\left\|D\hat{S} - A\right\|_\infty \leq 2\frac{\alpha}{1-\alpha}d^{-1}.$$

3. The computed truncated reduced system,

$$\left\|\hat{T}\hat{x}_{tr} - \hat{g}_r\right\|_\infty \leq \gamma_{6k}\frac{d+1}{d-1}\|\hat{x}_{tr}\|_\infty.$$

4. The overall error,

$$\|\hat{x} - x\|_\infty \leq \frac{1}{1-d^{-1}}\|A\hat{x} - f\|_\infty.$$

We see that the modified right-hand side g is computed with a small residual and that the bound becomes increasingly accurate as d becomes larger. The SPIKE matrix is computed with a very small residual and the bound is between 10^3 and 10^5 times too large. The computed reduced system is solved with a very small residual and the bound is between 10^2 and 10^3 times larger. Finally we see that using the residual to estimate the error is very reliable, leading to estimates that are accurate within one order of magnitude.

5.3. Comparisons with ScaLAPACK. We began by comparing the errors in the truncated SPIKE algorithm to ScaLAPACK (PDDBTRF/PDDBTRS) for four

TABLE 5.3

The 2-norm of the absolute error for ScaLAPACK (Sca) (PDDBTRF/PDDBTRS) and the truncated SPIKE (T.S) algorithm for four different banded matrices and different numbers of partitions. The results from LAPACK (DGBTRF/DGBTRS) are listed at the bottom of the table.

	(n, k_l, k_u)							
	$(2e4, 10, 10)$		$(1e5, 10, 10)$		$(1e5, 50, 50)$		$(1e6, 10, 10)$	
p	Sca	T.S	Sca	T.S	Sca	T.S	Sca	T.S
2	4.98e − 10	5.02e − 10	5.33e − 9	5.34e − 9	1.33e − 8	1.33e − 8	2.10e − 7	2.10e − 7
4	4.98e − 10	5.02e − 10	5.33e − 9	5.33e − 9	1.33e − 8	1.33e − 8	2.10e − 7	2.10e − 7
8	4.97e − 10	5.02e − 10	5.32e − 9	5.33e − 9	1.33e − 8	1.33e − 8	2.10e − 7	2.10e − 7
12	4.97e − 10	5.01e − 10	5.32e − 9	5.33e − 9	1.33e − 8	1.34e − 8	2.10e − 7	2.10e − 7
16	4.97e − 10	5.02e − 10	5.32e − 9	5.33e − 9	1.33e − 8	1.33e − 8	2.10e − 7	2.10e − 7
24	4.95e − 10	5.00e − 10	5.32e − 9	5.33e − 9	1.33e − 8	1.34e − 8	2.10e − 7	2.10e − 7
32	4.95e − 10	5.02e − 10	5.32e − 9	5.33e − 9	1.32e − 8	1.33e − 8	2.10e − 7	2.10e − 7
48	4.90e − 10	4.98e − 10	5.31e − 9	5.32e − 9	1.33e − 8	1.33e − 8	2.10e − 7	2.10e − 7
64	4.88e − 10	4.95e − 10	5.30e − 9	5.32e − 9	1.32e − 8	1.33e − 8	2.10e − 7	2.10e − 7
128	4.81e − 10	4.88e − 10	5.28e − 9	5.30e − 9	1.33e − 8	1.34e − 8	2.10e − 7	2.10e − 7
256	N/A	1.43e − 7	5.23e − 9	5.26e − 9	1.33e − 8	7.64e − 2	2.10e − 7	2.10e − 7
LA	4.99e − 10		5.33e − 9		1.33e − 8		2.10e − 7	

different matrices with

$$(n, k_l, k_u) \in \left\{ \left(2.0 \cdot 10^4, 10, 10\right), \left(10^5, 10, 10\right), \left(10^5, 50, 50\right), \left(10^6, 10, 10\right) \right\}.$$

Every diagonal entry was 1 and all other entries within the band were 10^{-2}. The right-hand side was constructed from the solution which was selected as $(1, 2, \ldots, n)^T$. The number of partitions were 2, 4, 8, 16, 24, 32, 48, 64, 128, and 256. The calculations were carried out in IEEE double precision arithmetic. We measured the 2-norm of the absolute error. Our results are displayed in Table 5.3. In our experiments ScaLAPACK did slightly better than the truncated SPIKE algorithm, but the difference between the two algorithms decreased, as the problems became larger. We would like to draw attention to the case of $p = 256$. In this case ScaLAPACK cannot be applied to the first matrix where $n = 20{,}000$, because the matrix is too small and the bandwidth is large compared to the number of partitions, and the routine issues the appropriate error message. The truncated SPIKE algorithm had a large error for the first and the third matrix. This is due to the fact that the infinity norm of the truncation error was very large: for the first matrix it was $1.62 \cdot 10^{-12}$, while for the third matrix it was $1.52 \cdot 10^{-7}$. In all other cases we found that the infinity norm of the truncation error was either less than machine ϵ or much smaller than the unit round off error u. The experiments with $p = 256$ emphasize the fact that the truncated SPIKE algorithm should not be applied to problems where the partitions are either too small or where the diagonal blocks are not diagonally dominant enough. The first matrix is diagonally dominant with degree $d = 5$, and for $p = 256$ the dimension of the smallest partition was 78. In this case Theorem 3.7 gives an upper bound for the infinity norm of the truncations error of $5^{-7} \approx 1.28 \cdot 10^{-5}$. In other words, we knew in advance that the result might not be accurate. Theorem 3.7 does not apply to the third matrix, which is not strictly diagonally dominant.

We found nine matrices that were diagonally dominant at Matrix Market. They were all quite small, with dimensions no larger than 5000. We extracted narrow banded matrices from these matrices by choosing $k = \lceil 0.01n \rceil$. We ran the examples through LAPACK (DGBTRF/DGBTRS), ScaLAPACK (PDDBTRF/PDDBTRS), our own implementation of the truncated SPIKE algorithm, as well as the SPIKE package itself (TU0). The matrices were scaled such that the main diagonals were 1

TABLE 5.4

The 2-norm of the absolute error for nine different matrices from Matrix Market. The results are given for LAPACK (dgbtrf/dgbtrs) and ScaLAPACK (PDDBTRF/PDDBTRS). The results are given for 2, 4, and 8 partitions.

matrix	n	LA	ScaLAPACK		
			2	4	8
dwb512	512	3.27e − 15	3.14e − 15	3.14e − 15	3.14e − 15
gr_30_30	900	0.00e + 00	0.00e + 00	1.57e − 16	2.94e − 16
jpwh_991	991	1.37e − 15	2.04e − 15	2.03e − 15	2.01e − 15
nos6	675	0.00e + 00	3.05e − 15	3.07e − 15	3.10e − 15
orsirr_1	1030	4.40e − 15	4.44e − 15	4.42e − 15	4.36e − 15
orsirr_2	886	4.11e − 15	4.10e − 15	4.11e − 15	4.12e − 15
orsreg_1	2205	7.08e − 15	7.12e − 15	7.30e − 15	6.93e − 15
sherman3	5005	1.92e − 12	1.99e − 12	1.99e − 12	1.99e − 12
sherman4	1104	2.53e − 15	2.59e − 15	2.58e − 15	2.58e − 15

TABLE 5.5

The 2-norm of the absolute error for nine different matrices from Matrix Market. The results are given for our implementation (T.S) of the truncated SPIKE algorithm, as well as the current implementation of the SPIKE package (TU0). The results are given for 2, 4, and 8 partitions.

matrix	T.S			T.U		
	2	4	8	2	4	8
dwb512	3.30e − 15	3.30e − 15	3.30e − 15	3.04e − 15	3.09e − 15	3.11e − 15
gr_30_30	0.00e + 00	0.00e + 00	0.00e + 00	0.00e + 00	1.11e − 16	5.09e − 16
jpwh_991	2.00e − 15	1.97e − 15	1.95e − 15	2.03e − 15	2.06e − 15	2.01e − 15
nos6	0.00e + 00	0.00e + 00	0.00e + 00	3.03e − 15	3.12e − 15	3.01e − 15
orsirr_1	4.39e − 15	4.38e − 15	4.31e − 15	4.40e − 15	4.26e − 15	4.15e − 15
orsirr_2	4.07e − 15	4.10e − 15	4.10e − 15	4.16e − 15	3.94e − 15	4.06e − 15
orsreg_1	7.16e − 15	7.16e − 15	7.00e − 15	6.49e − 15	6.79e − 15	6.18e − 15
sherman3	1.99e − 12	1.99e − 12	1.99e − 12	1.98e − 12	1.98e − 12	1.98e − 12
sherman4	2.49e − 15	2.50e − 15	2.51e − 15	2.44e − 15	2.41e − 15	2.48e − 15

and the right-hand side was generated from the solution which was selected as $x = (1, 1, \ldots, 1)^T$. We measured the 2-norm of the absolute error. Our results are listed in Table 5.4 and Table 5.5. We found no substantial difference in the accuracy of the four different routines.

6. Conclusion. We have shown that the SPIKE matrix is diagonally dominant by rows with a degree no less than that of the original matrix. We have derived a tight upper bound on the truncation error for the general case. We showed that the error committed at each stage is small, and we found that our bounds are probably pessimistic. We compared the truncated SPIKE algorithm to the corresponding algorithm in ScaLAPACK (PDDBTRF/PDDBTRS) and found no substantial difference between the accuracy of the two methods. The advantage of the truncated SPIKE algorithm is that if the matrix is diagonally dominant by rows with degree $d > 1$ and the partitions are sufficiently large, then the reduced system is essentially block diagonal and can be solved with a constant amount of communication, with all but one processor contributing equally to the solution of the reduced system. In ScaLAPACK (PDDTRS) the reduced system is solved recursively with the number of active processors being cut in half at each iteration.

Acknowledgments. The authors wish to thank the referees as well as the editor for their comments and suggestions which improved the presentation. The authors also want to thank their advisor Ahmed Sameh for his continued support.

REFERENCES

[1] P. Arbenz, A. Cleary, J. Dongarra, and M. Hegland, *A comparison of parallel solvers for diagonally dominant and general narrow banded linear systems* II, in EuroPar '99 Parallel Processing, P. Amestoy, Ph. Berger, M. Dayd, I. Duff, V. Frayss, L. Giraud, D. Ruiz, eds., Springer, Berlin, 1999, pp. 1078–1087.

[2] S.C. Chen, D.J. Kuck, and A. Sameh, *Practical parallel band triangular system solvers*, ACM Trans. Math. Software, 4 (1978), pp. 270–277.

[3] S. Demko, W.F. Moss, and Ph.W. Smith, *Decay rates for inverses of band matrices*, Math. Comput., 43 (1984), pp. 491–499.

[4] J.J. Dongarra and A. Sameh, *On some parallel banded system solvers*, Parallel Comput., 1 (1984), pp. 223–235.

[5] A. George and Kh. Ikramov, *Gaussian elimination is stable for the inverse of a diagonally dominant matrix*, Math. Comp., 73 (2003), pp. 653–657.

[6] N.J. Higham, *Accuracy and Stability of Numerical Algorithms*, 2nd ed., SIAM, Philadelphia, PA, 2002.

[7] J.-L. Larriba-Pey, À. Jorba, and J.J. Navarro, *Spike algorithm with savings for strictly diagonal dominant tridiagonal systems*, Microprocessing and Microprogramming, 39 (1993), pp. 125–128.

[8] D.H. Lawrie and A. Sameh, *The computation and communication complexity of a parallel banded system solver*, ACM Trans. Math. Software, 10 (1984), pp. 185–195.

[9] C.C.K. Mikkelsen, *The Decay Rate of the Solution to a Tridiagonal Linear System with a Very Special Right Hand Side*, Technical report, CSD TR #08-021, Computer Science Department, Purdue University, West Lafayette, IN, 2008.

[10] E. Polizzi and A. Sameh, *A parallel hybrid banded system solver: The SPIKE algorithm*, Parallel Comput., 32 (2006), pp. 177–194.

[11] E. Polizzi and A. Sameh, *SPIKE: A parallel enviroment for solving banded linear systems*, Comput. Fluids, 36 (2007), pp. 113–120.

[12] A.H. Sameh and D.J. Kuck, *On stable parallel linear system solvers*, J. ACM, 25 (1978), pp. 81–91.

[13] X.-H. Sun, H. Zhang, and L.M. Ni, *Efficient tridiagonal solvers on multicomputers*, IEEE Trans. Comput., 41 (1992), pp. 286–296.

[14] X.-H. Sun, *Application and accuracy of the parallel diagonal dominant algorithm*, Parallel Comput., 21 (1995), pp. 1241–1267.