1. Introduction. This is a short note which deals with a detail in the analysis of the truncated SPIKE algorithm [2], [3] for systems which are strictly diagonally dominant by rows. It contains the proof of Theorem 3.9 [1].

2. The main result. There is only one result namely the following theorem

THEOREM 2.1. 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 \( x \) given by

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

exhibits the smallest possible decay rate, i.e. if

\[
|x_1| = \epsilon^n
\] (2.1)

then

\[
c_i = 0, \quad \text{and} \quad |b_i| = \epsilon |a_i|,
\] (2.2)

for \( i = 1, 2, \ldots, n \).

Proof. We prove the theorem using the Thomas algorithm [4], which is designed to solve tridiagonal systems of the form

\[c_i x_{i-1} + a_i x_i + b_i x_{i+1} = f_i, \quad i = 1, 2, \ldots, n,\]

where \( x_0 \), and \( x_{n+1} \) are given in advance. If the system is strictly diagonally dominant by rows with degree \( d = \epsilon^{-1} > 1 \) then the solution can be computed as follows

\[x_i = p_i x_{i+1} + q_i, \quad i = 1, \ldots, n,\]

where the coefficient \( p_i \) and \( q_i \) are given by

\[p_0 = 0, \quad p_i = \frac{-b_i}{a_i + c_i p_{i-1}}, \quad i = 1, 2, \ldots, n,\]

and

\[q_0 = x_0, \quad q_i = \frac{f_i - c_i q_{i-1}}{a_i + c_i p_{i-1}}, \quad i = 1, 2, \ldots, n.\]

---

*Department of Computer Science, Purdue University (cmikkels@cs.purdue.edu)
We claim that \(|p_i| \leq \epsilon\), for \(i = 0, 1, 2, \ldots, n\). If \(b_i = 0\) then \(p_i = 0\), and there is nothing to show. Assuming \(|p_{i-1}| \leq \epsilon < 1\), and \(b_i \neq 0\), we have

\[
|p_i| = \frac{|b_i|}{|a_i + c_i p_{i-1}|} \leq \frac{|b_i|}{|a_i| - |c_i| \epsilon} \leq \frac{|b_i|}{\epsilon^{-1}(|b_i| + |c_i|) - |c_i| \epsilon} = \frac{|b_i|}{\epsilon^{-1}|b_i| + (\epsilon^{-1} - \epsilon)|c_i|} \leq \epsilon, \tag{2.3}
\]

because \(\epsilon \leq 1\), implies \((\epsilon^{-1} - \epsilon)|c_i| \geq 0\).

In our case \(x_0 = x_{n+1} = 0\), and \(f_i = 0\) for \(i = 1, 2, \ldots, n - 1\), while \(f_n = b_n\). It follows that

\[q_i = 0, \quad i = 0, 1, 2, \ldots, n - 1,\]

while

\[q_n = \frac{b_n}{a_n + c_n p_{n-1}}, \quad \text{and} \quad |q_n| \leq \epsilon.\]

It follows that

\[x_n = q_n, \quad x_i = (\prod_{j=i}^{n-1} p_j) q_n,\]

which implies that

\[|x_i| \leq \epsilon^{n-i+1}.\]

Now suppose \(|x_1|\) assumes the largest possible value, namely

\[|x_1| = \epsilon^n\]

then we must have

\[|p_i| = \epsilon, \quad i = 1, 2, \ldots, n - 1, \quad \text{and} \quad |q_n| = \epsilon.\]

Now, we claim that this can only happen if \(c_i = 0\), for \(i = 1, 2, \ldots, n\). From (2.3) we see that we actually have

\[\frac{|b_i|}{|a_i| - |c_i| \epsilon} = \epsilon,\]

for \(i = 1, 2, \ldots, n - 1\), as well as \(i = n\). It follows, that

\[\epsilon^2 |c_i| = \epsilon |a_i| - |b_i|.
\]

However, \(\epsilon |a_i| \geq |b_i| + |c_i|\), leaving us with

\[\epsilon^2 |c_i| = \epsilon |a_i| - |b_i| \geq |c_i|,
\]

from which we deduce \(|c_i| = 0\), because \(\epsilon < 1\). \(\Box\)

In short, if a tridiagonal matrix which is strictly diagonally dominant by rows, exhibits the slowest possible decay rate, then it is actually bidiagonal, and the ratio \(|b_i|/|a_i|\) is fixed. In our experience the spikes always decay much faster than the worst case.
REFERENCES