RATIONAL ITERATIVE METHODS FOR THE MATRIX SIGN FUNCTION*

CHARLES KENNEY† AND ALAN J. LAUB†

Abstract. In this paper an analysis of rational iterations for the matrix sign function is presented. This analysis is based on Padé approximations of a certain hypergeometric function and it is shown that local convergence results for "multiplication-rich" polynomial iterations also apply to these rational methods. Multiplication-rich methods are of particular interest for many parallel and vector computing environments. The main diagonal Padé recursions, which include Newton's and Halley's methods as special cases, are globally convergent and can be implemented in a multiplication-rich fashion which is computationally competitive with the polynomial recursions (which are not globally convergent). Other rational iteration schemes are also discussed, including Laurent approximations, Cayley power methods, and globally convergent eigenvalue assignment methods.

Key words. Padé approximation, matrix sign function, Riccati equations, rational iterations

AMS(MOS) subject classifications. 15A24, 65D99, 65F99

1. Introduction. It is a classical result that the algebraic Riccati equation can be solved by using an invariant subspace of an associated Hamiltonian matrix. This motivated the introduction, by Roberts [21] in 1971, of the matrix sign function as a means of finding the positive and negative invariant subspaces of any matrix $X$ which does not have eigenvalues on the imaginary axis. This and subsequent work [9] showed that the matrix sign function could be used to solve many problems in control theory.

The sign of $X$ can be defined constructively as the limit of the Newton sequence

$$
X_{n+1} = \frac{1}{2} (X_n + X_n^{-1}), \quad X_0 = X,
$$

$$
\text{sgn}(X) = \lim_{n \to +\infty} X_n.
$$

Newton's method has the pleasant feature that it is globally convergent; if $X$ has no eigenvalues on the imaginary axis then the limit in (1.2) exists. As a definition, however, (1.2) does not reveal many of the important properties of the sign function. Because of this, it is useful to have an equivalent definition based on the Jordan canonical form of $X$ (see [4], [7]). For a complex scalar $z$ with $\text{Re} \, z \neq 0$, define the sign of $z$ by

$$
\text{sgn} \, z = \begin{cases} 
1 & \text{if } \text{Re} \, z > 0, \\
-1 & \text{if } \text{Re} \, z < 0.
\end{cases}
$$

For a complex matrix $X$ such that $\Lambda(X) \subset \mathbb{C}^+ \cup \mathbb{C}^-$ (i.e., $X$ has no eigenvalues on the imaginary axis) let $T$ take $X$ to Jordan form:

$$
X = T^{-1} \begin{bmatrix} P & 0 \\ 0 & N \end{bmatrix} T,
$$

---

* Received by the editors September 28, 1989; accepted for publication (in revised form) November 15, 1989. This research was supported by National Science Foundation (and Air Force Office of Scientific Research) grant ECS87-18897, National Science Foundation grant DMS88-00817, and Air Force Office of Scientific Research contract AFOSR-89-0167.

† Department of Electrical and Computer Engineering, University of California, Santa Barbara, California 93106 (laub@lanczos@hub.ucsb.edu).

273
where $P$ and $N$ are in block diagonal Jordan form with, respectively, positive and negative real part eigenvalues. Then the sign of $X$ is given by

$$\text{sgn} (X) = T^{-1} \begin{bmatrix} I & 0 \\ 0 & -I \end{bmatrix} T,$$

where $I$ and $-I$ in (1.5) have the same dimensions as $P$ and $N$ in (1.4). This shows immediately that the sign of $X$ is a square root of the identity which commutes with $X$:

$$S^2 = I, \quad XS = SX,$$

where $S = \text{sgn} (X)$.

Using (1.4) in (1.1), shows that the eigenvalues $\lambda_j^{(n)}$ of $X_n$ are decoupled from each other and obey the scalar recursions

$$\lambda_j^{(n+1)} = \frac{1}{2} \left( \lambda_j^{(n)} + \frac{1}{\lambda_j^{(m)}} \right), \quad \lambda_j^{(0)} = \lambda_j (X),$$

with $\lim_{n \to +\infty} \lambda_j^{(n)} = \text{sgn} (\lambda_j)$. This decoupling greatly simplifies the analysis of methods like (1.1).

Because of the need for pivoting, matrix inversions are sometimes not as amenable to parallel or vector implementation as matrix multiplications. Thus, a current trend in evaluating $\text{sgn} (X)$ and related functions such as the polar decomposition [5], [11], [12] is to favor algorithms which are “multiplication-rich,” such as the Newton–Schulz iteration

$$X_{n+1} = \frac{1}{2} X_n (3I - X_n^2).$$

(The recursion (1.8) is obtained from (1.1) by using Schulz’s approximation $X_n^{-1} \approx X_n + (I - X_n^2) X_n$ as suggested in [12].) This method avoids the matrix inversion in (1.1) and is quadratically convergent provided

$$\| I - X^2 \| < 1,$$

where $\| \cdot \|$ is any reasonable matrix norm (see Theorem 5.2). If (1.9) is not satisfied then a starter method such as (1.1) must be used until $\| I - X_n^2 \| < 1$.

Higher-order polynomial recursions for the polar decomposition of a nonsingular matrix were developed independently by Kovarik [17] and Leipnik [18] and are applicable to the matrix sign function. These methods are based on polynomial approximations of the hypergeometric function

$$(1 - \xi)^{-1/2} = 1 + \frac{1}{2} \xi + \frac{3}{8} \xi^2 + \cdots,$$

and generate convergent matrix sequences provided that (1.9) is satisfied. The motivation for studying this function is that for nonzero real $x$, $\text{sgn} x = x/|x| = x/(1 - \xi)^{1/2}$ where $\xi = 1 - x^2$. In § 3, we show that the sufficient condition (1.9) actually provides a rather good approximation to the true region of convergence for these methods. Consequently, we might feel that loss of global convergence is the price that must be paid in order to use multiplication-rich algorithms. Rather surprisingly, this is not the case.

For example, recursions based on rational (Padé) approximations of $(1 - \xi)^{-1/2}$ have much larger regions of convergence. In fact, the main diagonal approximations (those for which the degree $m$ of the denominator is equal to or one greater than the
degree $k$ of the numerator) lead to globally convergent iterations that satisfy an elegant error formula:

\[(S-X_n)(S+X_n)^{-1} = (S-X_0)^{\gamma^n}(S+X_0)^{-\gamma^n},\]

where $\gamma = k + m + 1$ is the order of the approximation. (For Newton’s method, a similar result was proved by Balzer [3, eq. (39)] and by Roberts [21, § 1.3].) These methods are easily modified to allow exact one-step convergence of specified eigenvalues (much like the eigenvalue assignment schemes of Balzer in [3]) while still remaining globally convergent. An analysis of the Halley family of algorithms of Gander [10] for the polar decomposition shows that these methods belong to this class of assignment procedures. The work in [10] can also be adapted to give a local convergence theory for general sign function iterations of the form $X_{n+1} = F(X_n)$.

A second family of globally convergent multiplication-rich methods is based on the Cayley transform

\[(1.12) \quad Y = (I-X)(I+X)^{-1},\]

which takes the positive real part eigenvalues of $X$ inside the unit circle and the negative real part eigenvalues of $X$ outside the unit circle. If $Y$ is multiplied by itself repeatedly, then these eigenvalues move toward zero and infinity, respectively. Transforming back to get $X_n$,

\[(1.13) \quad X_n = (I - Y^n)(I + Y^n)^{-1}\]

moves these eigenvalues very near one and minus one, respectively. (If $X$ has $-1$ as an eigenvalue, then $I + X$ is singular and a modified version of (1.12), (1.13) must be used.) A fascinating correspondence between the Cayley power method and the Padé approximation method is that if the power $\nu$ in (1.13) is equal to $\gamma^n$ in (1.11), then $X_n$ is equal to $\tilde{X}_n$. This does not mean, however, that these two methods should be viewed as identical because in this case the Padé method requires $n$ matrix inversions while the Cayley method requires only two. Similar equivalency results for different members of the Padé method can also be proved (see Theorem 3.4). An interesting sidelight on the Cayley power method is that (1.12) can be replaced by any transformation which is a rational or analytic function of $X$ that takes the right- and left-half complex planes inside and outside the unit disk, respectively. For example, if $Y = e^{-X}$ then $Y^n$ is just the fundamental solution matrix to $\hat{Y} = \hat{X} \hat{Y}$ at time $\nu$: $Y^n = e^{-\nu X}$ and $\tilde{X}_n = (I - e^{-\nu X})(I + e^{-\nu X})^{-1}$. Note in this case that $I + e^{-\nu X}$ is never singular, since the eigenvalues of $X$ are not on the imaginary axis.

In the next section we present the theory of the Padé approximants of $(1-\xi)^{-1/2}$ for $k \geq m - 1$, which is based on well-known results for hypergeometric functions. This theory is then used to analyze scalar sign function recursions in § 3, where we also show how it can be adapted to give globally convergent eigenvalue assignment iterations. In § 4 we consider other rational iterations including Laurent methods. These scalar results are useful because matrix convergence is predicated on the scalar convergence of the eigenvalues of $X$ (§ 5). This leads to local convergence results for $k \geq m - 1$, and global convergence for the main diagonal approximants $k = m$ and $k = m - 1$.

2. Padé approximations to $(1-\xi)^{-1/2}$. Let $(\alpha)_n = (\alpha)(\alpha + 1)\cdots(\alpha + n - 1)$ with $(\alpha)_0 = 1$, and define the family of hypergeometric functions

\[(2.1) \quad {}_2F_1(\alpha, \beta, \gamma, \xi) = \sum_{n=0}^{+\infty} \frac{(\alpha)_n(\beta)_n}{n!(\gamma)_n} \xi^n.\]
From [1],

\[(2.2) \quad (1 - \xi)^{-1/2} = \F1_2\left(\frac{1}{2}, 1, 1, \xi\right) = f(\xi).\]

In general, the \([k/m]\) Padé approximant to \(f\) is a rational function \(P_{km}/Q_{km}\) where \(\deg(P_{km}) = k\), \(\deg(Q_{km}) = m\), and

\[(2.3) \quad f(\xi) - \frac{P_{km}(\xi)}{Q_{km}(\xi)} = O(\xi^{k + m + 1}).\]

Because \(f\) is a hypergeometric function, a great deal is known about \(P_{km}\) and \(Q_{km}\) [1]. First of all [13], \(Q_{km}\) is related to the set of orthogonal polynomials over \([0, 1]\) defined with respect to the weight function \(\omega(\xi) = (\xi^{-1/2}/\pi)(1 - \xi)^{-1/2}\xi^{k+1-m}\) for \(k \geq m - 1\).

If \(\psi_m\) is the \(m\)th such polynomial with \(\psi_m(1) = 1\), then

\[(2.4) \quad Q_{km}(\xi) = \xi^m \psi_m(\xi^{-1}),\]

and \(Q_{km}(0) = 1\). From (2.4), the zeros of \(Q_{km}\) are just the reciprocals of the zeros of \(\psi_m\). Since the zeros of \(\psi_m\) are simple [22] and lie in \((0, 1)\), the zeros of \(Q_{km}\) are also simple and lie in \((1, \infty)\). (This result could have been anticipated from another point of view since \((1 - \xi)^{-1/2}\) has a natural branch cut along \((1, \infty)\) and, as noted in [2, pp. 51-57], the zeros and poles of a Padé approximant tend to fall along the branchcuts of the functions they approximate.) Denoting the zeros of \(Q_{km}\) by \(z_1 < z_2 < \cdots < z_m\), we may write

\[(2.5) \quad Q_{km}(\xi) = \prod_{i=1}^{m} \frac{(z_i - \xi)}{z_i}.\]

This identity is useful for convergence analysis, but a more convenient form [1] is

\[(2.6) \quad Q_{km}(\xi) = \F1_2\left(-m, -\frac{1}{2} - k, -k - m, \xi\right) = \sum_{n=0}^{m} \frac{(-m)_n (-\frac{1}{2} - k)_n \xi^n}{n! (-k - m)_n} = \sum_{n=0}^{m} q_{kn} \xi^n.\]

From [13], \(P_{km}\) is given by

\[(2.7) \quad P_{km}(\xi) = \sum_{n=0}^{k} p_{km}^{n} \xi^n = \sum_{n=0}^{k} \frac{(\frac{1}{2} - m)_n (n - k - m)_m \xi^n}{n! (-k - m)_m (n + \frac{1}{2} - m)_m} = \sum_{n=0}^{k} \frac{(-m)_n (-\frac{1}{2} - k)_n \xi^n}{n! (-k - m)_n}.\]

The key to the local error analysis of Padé recursions is the following theorem, which was proved by Leipnik [18, Thm. 1] and stated by Kovarik [17, lemma following Thm. 2] for the polynomial case \(m = 0\).

**THEOREM 2.1.** For \(k \geq m - 1\),

\[(2.8) \quad Q_{km}(\xi) - (1 - \xi) P_{km}^2(\xi) = \xi^{k + m + 1} \left( \sum_{i=1}^{n} c_i \xi^i \right),\]
where \(c_i = c_i(k, m) > 0\) for \(0 \leq i \leq \mu = \max(2k+1, 2m) - (k+m+1)\), and

\[
Q_{km}^2(1) = \sum_{i=1}^{\mu} c_i.
\]

**Proof.** From (2.3) and the fact that \(Q_{km}^2(\xi) - (1 - \xi)P_{km}^2(\xi)\) is a polynomial of order \(\mu + k + m + 1\),

\[
Q_{km}^2(\xi) - (1 - \xi)P_{km}^2(\xi) = \frac{Q_{km}^2}{f^2} \left( f - \frac{P_{km}}{Q_{km}} \right) \left( f + \frac{P_{km}}{Q_{km}} \right)
\]

\[
= \xi^{k+m+1} \left( \sum_{i=1}^{\mu} c_i \xi^i \right),
\]

for some constants \(c_0, c_1, \cdots, c_{\mu}\). Setting \(\xi = 1\) gives (2.9). It remains to show that the coefficients \(c_i\) are positive. The idea of the proof is best illustrated by considering the diagonals, \(m = k - t\), for \(t = -1, 0, 1, \cdots, k\) in the Padé table. (For example, see Table 1.) For the first main diagonal, \(t = -1, \mu = 0\) and multiplying out the left side of (2.8) gives \(c_0 = (q_{k+1}^{kk})^2 > 0\). For the second main diagonal, \(t = 0, \mu = 0, c_0 = (P_{k}^{kk})^2 > 0\). For the first superdiagonal, \(t = 1, \mu = 1, c_1 = 1\), and

\[
c_0 = (P_{k}^{kk})^2 > 0, \quad c_1 = P_{k}^{kk} - (P_{k}^{kk} - 1) + P_{k}^{kk} - 1 P_{k}^{kk} - 1.
\]

In general, for \(t \geq 0, \mu = t, \) and the coefficients, \(c_s\) can be written as the sum of terms of the form

\[
P_{k-r}^{kk-t}(P_{k+r-s}^{kk-t} - P_{k+r-s+1}^{kk-t}),
\]

and

\[
P_{k-r}^{kk-t} P_{k+r-s}^{kk-t},
\]

where

\[
0 \leq r \leq s \leq t \leq k.
\]

We complete the proof of the theorem by showing that each term of the type (2.10) or (2.11) is positive. From (2.7),

\[
P_{k-r}^{kk-t} = \frac{1}{(k+r-s)!} \frac{(k-r-s-t)(k-r-s-t-k)(k-r-s-t-k)}{(k+r-s-t-k)(k+r-s-t-k+1)}.
\]

Since both \(P_{k-r}^{kk-t}\) and \(P_{k-r}^{kk-t}\) have sign \((-1)^{k-t}\),

\[
P_{k-r}^{kk-t} P_{k+r-s}^{kk-t} > 0.
\]

Using (2.13),

\[
P_{k-r}^{kk-t} (P_{k+r-s}^{kk-t} - P_{k+r-s+1}^{kk-t}) = P_{k-r}^{kk-t} P_{k+r-s}^{kk-t} \left( 1 - \frac{(s-r)(s-r+k-t)}{(s-r+k-t)(k+r-s+1)} \right) > 0
\]

by (2.14) and (2.12) because \((s-r)/(s-r+k-t) \leq 1\) and

\[
(t-s+r+\frac{1}{2})/(k-s+r+1) < 1.
\]

(Note that the degenerate case \(k = t = s = r\) does not cause a problem because (2.10) then reduces to \(P_0^{kk} P_k^{kk}\), which is positive by (2.14).) \(\square\)
3. Scalar Padé recursions. As we show in §5, the convergence of the matrix sequence \( \{X_n\} \) is determined by the convergence of the scalar sequences for the eigenvalues of \( X_0 \). The scalar Padé recursions have the form

\[
x_{n+1} = x_n \frac{P_{km}(1-x_n^2)}{Q_{km}(1-x_n^2)},
\]

where \( P_{km}/Q_{km} \) is the \([k/m] \) Padé approximant to \((1-\xi)^{-1/2}\). Table 1 gives the expressions for the right-hand side of (3.1) for \( k \) and \( m \) between zero and three. For example, the case \( k = 0, m = 1 \) gives

\[
x_{n+1} = \frac{2x_n}{1+x_n^2},
\]

which might be called the “inverse” Newton method for solving the equation \( x^2 - 1 = 0 \) since the values \( x_1, x_2, \cdots \) generated by (3.2) are the inverses of those generated by the “regular” Newton method

\[
x_{n+1} = \frac{1}{2}\left(x_n + \frac{1}{x_n}\right).
\]

The case \( k = 1, m = 1 \) gives Halley’s method (see [10] for a related application). The next theorem generalizes the local convergence results of Leipnik [18] and Kovarik [17].

**Theorem 3.1.** Let \(|1-x_0| < 1\) for \( x_0 \in \mathbb{C} \) and define \( \{x_n\} \) by (3.1) for \( k \geq m - 1 \). Then

\[
|1-x_n^2| \leq |1-x_0^2|^{(k+m+1)^n},
\]

and

\[
\lim_{n \to +\infty} x_n = \text{sgn}(x_0).
\]

**Proof.** By (3.1),

\[
1-x_n^2 = \frac{Q_{km}(\xi) - (1-\xi)P_{km}(\xi)}{Q_{km}(\xi)},
\]

where \( \xi = 1-x_0^2 \). But \( Q_{km} \) has zeros \( z_1, \cdots, z_m \) in \((1, +\infty)\), so by (2.5)

\[
|Q_{km}(\xi)| = \prod_{i=1}^{m} \left| \frac{z_i-\xi}{z_i} \right| \leq \prod_{i=1}^{m} \frac{|z_i-\xi|}{z_i} = Q_{km}(1).
\]

### Table 1
Padé recursions for the matrix sign function.

<table>
<thead>
<tr>
<th>( m )</th>
<th>( k = 0 )</th>
<th>( k = 1 )</th>
<th>( k = 2 )</th>
<th>( k = 3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( m = 0 )</td>
<td>( x )</td>
<td>( \frac{x}{2}(3-x^2) )</td>
<td>( \frac{x}{8}(15-10x^2+3x^4) )</td>
<td>( \frac{x}{16}(35-35x^2+21x^4-5x^6) )</td>
</tr>
<tr>
<td>( m = 1 )</td>
<td>( \frac{2x}{1+x^2} )</td>
<td>( \frac{x(3+x^2)}{1+3x^2} )</td>
<td>( \frac{x(15+10x^2-x^4)}{4(1+5x^2)} )</td>
<td>( \frac{x(35+35x^2-7x^4+x^6)}{8} )</td>
</tr>
<tr>
<td>( m = 2 )</td>
<td>( \frac{8x}{3+6x^2-x^4} )</td>
<td>( \frac{4x(1+x^2)}{1+6x^2+x^4} )</td>
<td>( \frac{x(5+10x^2+x^4)}{1+10x^2+5x^4} )</td>
<td>( \frac{x(35+105x^2+21x^4-x^6)}{2} )</td>
</tr>
<tr>
<td>( m = 3 )</td>
<td>( \frac{16x}{5+15x^2-5x^4+x^6} )</td>
<td>( \frac{8x(3+5x^2)}{5+45x^2+15x^4-x^6} )</td>
<td>( \frac{2x(3+10x^2+3x^4)}{1+15x^2+15x^4+x^6} )</td>
<td>( \frac{x(7+35x^2+21x^4+x^6)}{1+21x^2+35x^4+7x^6} )</td>
</tr>
</tbody>
</table>
Using Theorem 2.1 in (3.6) gives
\[
|1 - x_n^2| \leq |\xi|^{k+m+1} \left( \sum_{i=1}^{\mu} c_i |\xi|^i \right) / |Q_{km}(\xi)|^2
\]
\[
\leq |1 - x_0^2|^{k+m+1} \left( \sum_{i=1}^{\mu} c_i \right) / |Q_{km}(\xi)|^2
\]
\[
\leq |1 - x_0^2|^{k+m+1} Q_{km}(1) / |Q_{km}(\xi)|^2
\]
\[
\leq |1 - x_0^2|^{k+m+1}
\]
by (2.9) and (3.7). Repeating this argument gives (3.4). From (3.4), \(x_n^2 \to 1\). To see that \(x_n \to \text{sgn} (x_0)\), let \(h(x) = x P_{km}(1 - x^2)/Q_{km}(1 - x^2)\). Since the only poles of \(h\) lie on the imaginary axis, \(h\) is continuous on the set
\[
(3.8) \quad S = \{x: |1 - x^2| < 1\} = S_+ \cup S_-,
\]
where \(S_+ = \{x \in S: \text{Re } x > 0\}\), \(S_- = \{x \in S: \text{Re } x < 0\}\). By (3.4), \(h\) takes \(S\) into \(S\). Since \(S_+ \cap S_- = \emptyset\) and each is a connected set, \(h(S_+)\) must lie entirely in \(S_+\) or \(S_-\), because the continuous image of a connected set is connected. But \(1 \in S_+\) and \(h(1) = 1 \in S_+\), so \(h(S_+) \subset S_+\). Similarly, \(h(S_-) \subset S_-\). Thus if \(x_0 \in S_+\) then \(x_n \in S_+\) for all \(n\), and by (3.4),
\[
\lim_{n \to +\infty} x_n = \text{sgn} (x_0).
\]

In order to assess how well the set \(S\) in (3.8) approximates the region of convergence for the recursions in (3.1), we define the basins of attraction for the fixed points \(\pm 1\) of \(h\):
\[
(3.9) \quad B_+ = \{x: \lim_{n \to +\infty} x_n = 1\}, \quad B_- = \{x: \lim_{n \to +\infty} x_n = -1\}.
\]
The Julia set \([6], [19]\) for the recursion (3.1) is the boundary of the basin of attraction of \(+1:\)
\[
(3.10) \quad J_{km} = \partial B_+.
\]
Because of the unusual properties associated with Julia sets, \(J_{km}\) is also the boundary of the basin of attraction for \(-1:\)
\[
(3.11) \quad J_{km} = \partial B_-.
\]
(See [19] for a very readable introduction to Julia sets and the properties of rational recursions such as (3.1); for a deeper study, see [6].)

Computationally, \(J_{km}\) can be approximated by starting with (almost) any point \(z_0 \in \mathbb{C}\) and then reversing (3.1) to solve for the predecessors of \(z_0:\)
\[
(3.12) \quad z_n = z_{n+1} P_{km}(1 - z_{n+1}^2)/Q_{km}(1 - z_{n+1}^2),
\]
where \(z_n = x_{-n}\) in (3.1). Since (3.12) can be written as a polynomial in \(z_{n+1}\) of order \(\mu_1 = \max (2k + 1, 2m)\), there are \(\mu_1\) solutions \(z_{n+1}^{(i)}\) to (3.12), one of which is selected at random to continue the iteration. This scheme takes advantage of the fact that for the forward recursion (3.1), the Julia set is repulsive; points near \(J_{km}\) move to \(\pm 1\). In reverse, under (3.12), the Julia set becomes attractive and nearly all orbits of points are dense in \(J_{km}\) (see [6, Thm. 2.5]). Thus by plotting \(\{z_n^{(i)}\}\) for \(n > 30\) (to allow the initial points time to approach the Julia set) we obtain a good graphical approximation of \(J_{km}\) and
thus can assess easily the real region of convergence of (3.1) as compared to the set $|1 - x^2| < 1$. This was done for each of the recursions given in Table 1 (excluding the globally convergent main diagonal recursions), and the results are displayed in Figs. 1–9, along with the set $|1 - x^2| = 1$ for comparison (this set looks like an "infinity" symbol centered at zero). In each of these figures, the principal domains of attraction of $±1$ are the largest connected regions, inside the Julia set, which contain $±1$, respectively. The other connected regions nested within the Julia set map onto these principal domains after a finite number of steps in (3.1). For the multiplication-rich polynomial recursions ($m = 0$), the set $|1 - x^2| < 1$ provides a rather good approximation to the actual region of convergence. However, as $m$ increases toward $k$, that is, as we move toward the main diagonals $k = m$ or $k = m - 1$, the region of convergence becomes much larger than $|1 - x^2| < 1$.

We now show that along the main diagonals, the regions of convergence are as large as possible and we have, in fact, global convergence. That is, if $x_0$ is not on the imaginary axis then $\lim_{n \to +\infty} x_n = \text{sgn} (x_0)$.

First note a rather remarkable property of (3.1) for $k = m$ and $k = m - 1$: the polynomials $-xP_{km}(1 - x^2)$ and $Q_{km}(1 - x^2)$ are, respectively, the odd and even parts of $(1 - x)^{k+m+1}$. This makes it very easy to write down the appropriate recursion. For example, if $k = m = 2$, then

$$(1 - x)^{k+m+1} = 1 - 5x + 10x^2 - 10x^3 + 5x^4 - x^5,$$

so $-xP_{22}(1 - x^2) = -5x - 10x^3 - x^5$, and $Q_{22}(1 - x^2) = 1 + 10x^2 + 4x^4$. Thus the $[2/2]$ recursion is $x_{n+1} = x_n(5 + 10x_n^2 + x_n^4)/(1 + 10x_n^2 + 5x_n^4)$. This property can be proved either by manipulating the series (2.6), (2.7) or by starting with $-xP_{km}(1 - x^2)$ and $Q_{km}(1 - x^2)$ as the odd and even parts of $(1 - x)^{k+m+1}$ and then showing that (2.3) is satisfied.

**Theorem 3.2.** Let $x_0 \in \mathbb{C}^+ \cup \mathbb{C}^-$ and let $\{x_n\}$ be defined by (3.1) for $k = m$ or $k = m - 1$. Then

$$(3.13) \quad \frac{1 - x_n}{1 + x_n} = \left(\frac{1 - x_0}{1 + x_0}\right)^{(k+m+1)n} \quad \text{for } x_0 \in \mathbb{C}^+$$

and

$$(3.14) \quad \frac{1 + x_n}{1 - x_n} = \left(\frac{1 + x_0}{1 - x_0}\right)^{(k+m+1)n} \quad \text{for } x_0 \in \mathbb{C}^-. $$

In either case, for $s = \text{sgn} (x_0)$,

$$(3.15) \quad \frac{s - x_n}{s + x_n} = \left(\frac{s - x_0}{s + x_0}\right)^{(k+m+1)n}.$$

**Proof.** Equations (3.13) and (3.14) are identical except for being inverses of each other to avoid division by zero when $x_0 = ±1$. Let $x_0 \in \mathbb{C}^+$ for convenience and set $x_1 = x_0 P_{km}(1 - x_0^2)/Q_{km}(1 - x_0^2).$ By the preceding remarks, for any $x$, and $k = m$ or $m - 1$,

$$(3.16) \quad Q_{km}(1 - x^2) - xP_{km}(1 - x^2) = (1 - x)^{k+m+1}. $$

Replacing $x$ by $-x$ gives

$$(3.17) \quad Q_{km}(1 - x^2) + xP_{km}(1 - x^2) = (1 + x)^{k+m+1}. $$
Fig. 1. Padé convergence region for $k = 1$, $m = 0$.

Fig. 2. Padé convergence region for $k = 2$, $m = 0$.

Fig. 3. Padé convergence region for $k = 3$, $m = 0$. 
Fig. 4. Padé convergence region for $k = 0, m = 2$.

Fig. 5. Padé convergence region for $k = 2, m = 1$.

Fig. 6. Padé convergence region for $k = 3, m = 1$. 
Fig. 7. Padé convergence region for $k = 0, m = 3$.

Fig. 8. Padé convergence region for $k = 1, m = 3$.

Fig. 9. Padé convergence region for $k = 3, m = 2$. 
Thus,
\[
1 - x_1 = \frac{(Q_{km}(1 - x_0^2) - x_0 P_{km}(1 - x_0^2))}{Q_{km}(1 - x_0^2)} \cdot (1 - x_0)^{k+m+1}/Q_{km}(1 - x_0^2),
\]
and
\[
1 + x_1 = (1 + x_0)^{k+m+1}/Q_{km}(1 - x_0^2).
\]
Dividing, we obtain (3.13 for \( n = 1 \). Repeat to get the general statement. □

From Theorem 3.2, we immediately get Theorem 3.3.

**THEOREM 3.3 (Global Convergence).** If \( x_0 \in C^+ \cup C^- \), then for \( k = m \) or \( k = m - 1 \), with \( m \geq 1 \), \( \lim_{n \to +\infty} x_n = \text{sgn} (x_0) \).

**Proof.** By Theorem 3.2, we need only show that \( |(1 - x_0)/(1 + x_0)| < 1 \) for \( x_0 \in C^+ \) or \( |(1 + x_0)/(1 - x_0)| < 1 \) for \( x_0 \in C^- \). Let \( x_0 = \rho \cos \theta \in C^+ \) with \( -\pi/2 < \theta < \pi/2 \). Then \( |(1 - x_0)/(1 + x_0)|^2 = (1 - 2\rho \cos \theta + \rho^2)/(1 + 2\rho \cos \theta + \rho^2) < 1 \), and similarly for \( x_0 \in C^- \). □

From Theorem 3.2, we see that the distance measure from \( x_0 \) to 1 given by \( d_+(x_0) = |(1 - x_0)/(1 + x_0)| \) and its counterpart for \(-1, d_-(x_0) = |(1 + x_0)/(1 - x_0)| \), are more natural than \( |1 - x| \) and \( |1 + x| \), respectively. For example, \( x_0 = 10^{-6} \) and \( 1/x_0 = 10^6 \) are equidistant from 1 under the regular Newton method (since (1.2) is symmetric with respect to \( x_0 \) and \( 1/x_0 \)) but \( |1 - x_0| \approx 1 \) while \( |1 - 1/x_0| \approx 10^6 \). (See [3] and [15].)

Theorem 3.2 is also useful in establishing the equivalence of certain methods in the Padé table. For example, if \( x_0 \in C^+ \), then two steps of the inverse Newton method \((k = 0, m = 1)\) give
\[
(3.18) \quad \frac{1 - x_2}{1 + x_2} = \left( \frac{1 - x_0}{1 + x_0} \right)^4.
\]
However, if \( \tilde{x}_1 \) denotes the result of taking one step from \( x_0 \) with the recursion \((k = 1, m = 2)\), then
\[
(3.19) \quad \frac{1 - \tilde{x}_1}{1 + \tilde{x}_1} = \left( \frac{1 - x_0}{1 + x_0} \right)^4.
\]
Solving for \( x_2 \) and \( \tilde{x}_1 \), we find \( x_2 = \tilde{x}_1 \). Similarly, if we take one step with \((k = 0, m = 1)\) followed by a step with \((k = 3, m = 3)\) the result would be the same as one step with \((k = 6, m = 7)\).

**THEOREM 3.4 (Equivalency).** Let \( x_0 \in C^+ \cup C^- \) and let \( x_r \) be the result of applying \( r \) steps of the (possibly different) main diagonal Padé recursions \([k_1/m_1], \ldots, [k_r/m_r]\). Then \( x_r = \tilde{x}_r \), where \( \tilde{x}_r \) is obtained by \( r \) main diagonal steps \([\tilde{k}_1/\tilde{m}_1], \ldots, [\tilde{k}_r/\tilde{m}_r]\), provided that both are of the same order, i.e.,
\[
(3.20) \quad \prod_{i=1}^{r} (k_i + m_i + 1) = \prod_{i=1}^{r} (\tilde{k}_i + \tilde{m}_i + 1) = \rho.
\]

**Proof.** Applying Theorem 3.2 for each individual step,
\[
\left( \frac{1 - x_r}{1 + x_r} \right) = \left( \frac{1 - x_0}{1 + x_0} \right)^\rho = \left( \frac{1 - \tilde{x}_r}{1 + \tilde{x}_r} \right).
\]
Solving for \( x_r \) and \( \tilde{x}_r \) gives \( x_r = \tilde{x}_r \). If \( x_0 \in C^- \), use (3.14). □
4. Other rational methods. In this section we consider other rational iterations, including eigenvalue assignment methods, Cayley transform methods, and Laurent series methods. Eigenvalue assignment methods were introduced by Balzer [3], in the form of scaled Newton methods which move specified real eigenvalues to $x = 1$ in one step. These methods were shown to be globally but not quadratically convergent. By using the methods of Theorem 3.2, it is easy to construct globally convergent methods of arbitrarily high order that will move any selected set $\{\lambda_i\}$ of real or complex conjugate eigenvalues to $x = 1$ in one step.

For example, if we want a fourth-order method which assigns $\lambda_1 = 2$, $\lambda_2 = 1 + i$, and $\lambda_3 = 1 - i$ to $x = 1$, then we let $-xp(x^2)$ and $q(x^2)$ be, respectively, the odd and even terms in the expansion of $(1 - x)^4(2 - x)(1 + i - x)(1 - i - x)$:

$$(1 - x)^4(2 - x)(1 + i - x)(1 - i - x) = 4 - 22x + 52x^2 - 69x^3 + 56x^4 - 28x^5 + 8x^6 - x^7.$$

Then

$$xp(x^2) = 22x + 69x^3 + 28x^5 + x^7 = x(22 + 69x^2 + 28x^4 + x^6),$$

$$q(x^2) = 4 + 52x^2 + 56x^4 + 8x^6 = 4(1 + 13x^2 + 14x^4 + 2x^6),$$

and the desired iteration is

$$x_{n+1} = \frac{x_n(22 + 69x_n^2 + 28x_n^4 + x_n^6)}{4(1 + 13x^2 + 14x^4 + 2x^6)}.$$

In order to prove global convergence for these assignment methods, we need the following lemma.

**Lemma 4.1.** Let $\Re z > 0$, $\Re \lambda > 0$, and $r > 0$. Then

$$\left| \frac{r - z}{r + z} \right| < 1,$$

and

$$\left| \frac{\lambda - z}{\lambda + z} \right| < 1.$$

**Proof.** If we set $x = z/r$, then $\Re x > 0$ and

$$\left| \frac{r - z}{r + z} \right| = \left| \frac{1 - x}{1 + x} \right| < 1,$$

as in the proof of Theorem 3.3. Now say $\lambda = re^{i\theta}$, $z = pe^{i\phi}$ where $\phi, \theta \in (-\pi/2, \pi/2)$. Then

$$\left| (\lambda - z)(\bar{\lambda} - z) \right|^2 = (r^2 - 2pr \cos \theta \cos \phi + r^2 \cos 2\phi)^2 + \sin^2 \phi(2r^2 \cos \phi - 2pr \cos \theta)^2$$

$$< (r^2 + 2pr \cos \theta \cos \phi + r^2 \cos 2\phi)^2 + \sin^2 \phi(2r^2 \cos \phi + 2pr \cos \theta)^2$$

$$= |(\lambda + z)(\bar{\lambda} + z)|^2.$$
THEOREM 4.2. Let \( \{\lambda_1, \lambda_2, \cdots, \lambda_\mu\} \) be a conjugate symmetric set in the open right-half plane and \( -xp(x^2) \) and \( q(x^2) \) be, respectively, the odd and even parts of \( (1-x)^\gamma(\lambda_1 - x) \cdots (\lambda_\mu - x) \). Then the iterative method

\[
x_{n+1} = \frac{x_n p(x_n^2)}{q(x_n^2)}
\]

is globally convergent of order \( \gamma \) and takes \( \{\lambda_1, \lambda_2, \cdots, \lambda_\mu\} \) to \( x = 1 \) in one step. Moreover, for \( s = \text{sgn} (x_0) \),

\[
\frac{s - x_{n+1}}{s + x_{n+1}} = \frac{(s - x_n)^\gamma (\lambda_1 s - x) \cdots (\lambda_\mu s - x)}{(s + x_n)^\gamma (\lambda_1 s + x) \cdots (\lambda_\mu s + x)},
\]

and

\[
\frac{|s - x_{n+1}|}{|s + x_{n+1}|} \leq \frac{|s - x_n|^\gamma}{|s + x_n|},
\]

Proof. We shall prove (4.4) and (4.5) for the case \( s = 1 \), since the case \( s = -1 \) follows immediately. From (4.3),

\[
1 - x_{n+1} = \frac{q(x_n^2) - x_n p(x_n^2)}{q(x_n^2) + x_n p(x_n^2)}
\]

which proves (4.4). Inequality (4.5) then follows from (4.4) and Lemma 4.1.

Remark 1. Since \( xp(x^2)/q(x^2) \) in (4.3) is an odd function, it also moves \( \{-\lambda_1, -\lambda_2, \cdots, -\lambda_\mu\} \) to \(-1\) in one step.

Remark 2. In\[10\], Gander gives a family of quadratically convergent methods which depend on a parameter \( f \):

\[
x_{n+1} = x_n \frac{2f - 3 + x_n^2}{2f - 2 + f x_n^2}.
\]

In Theorem 2 of\[10\], it is shown that (4.6) is globally convergent for \( f > 2 \) and for \( f = 3 \) gives Halley’s method, which is cubically convergent. For \( f < 2 \), prescaling must be done to ensure convergence. We can interpret Gander’s method as a second-order method which makes one real eigenvalue assignment. Expand

\[(1-x)^2(\lambda - x) = \lambda - (2\lambda + 1)x + (2 + \lambda)x^2 - x^3,\]

and use the method of Theorem 4.2 to obtain the iteration

\[
x_{n+1} = \frac{x_n (2\lambda + 1 + x_n^2)}{\lambda + (2 + \lambda)x_n^2}.
\]

This is the same as (4.6) for \( \lambda = f - 2 \). Thus the condition \( f > 2 \) for global convergence in (4.6) is just the requirement that the real eigenvalue \( \lambda \), which gets mapped to \( x = 1 \), must be in the right-half plane as in Theorem 4.2. Moreover, \( f = 3 \) corresponds to \( \lambda = 1 \) being triply assigned to \( x = 1 \), so that the iteration is cubically convergent (Halley’s method).
Remark 3. Allowing some of the eigenvalues \( \lambda \) in Theorem 4.2 to be multiple results in methods in which \( \lambda \) is mapped to \( x = 1 \) and points near \( \lambda \) are taken at least quadratically to \( x = 1 \). For example, expanding \((1 - x)^2(2 - x)^2\) gives the second-order method in which \( \lambda = 2 \) is doubly assigned to one:

\[
x_{n+1} = \frac{x_n(12 + 6x_n^2)}{4 + 13x_n^2 + x_n^4}.
\]

If \( x_0 = 2.1 \), then \( x_1 = 0.99985 \cdots \).

As indicated in the Introduction, another family of methods can be based on the Cayley transform. For \( x \neq -1 \), let

\[
y = \frac{1 - x}{1 + x}.
\]

Let \( \tilde{x} \) denote the result of multiplying \( y \) by itself \( \nu \) times and then transforming back:

\[
\tilde{x}_\nu = \frac{1 - y^\nu}{1 + y^\nu}.
\]

From this we see that

\[
\frac{1 - \tilde{x}_\nu}{1 + \tilde{x}_\nu} = y^\nu = \left(\frac{1 - x}{1 + x}\right)^\nu.
\]

Now suppose that \( x_n \) is defined by (3.1) for one of the main diagonal \((k = m \text{ or } k = m - 1)\) Padé recursions where \( \nu = (k + m + 1)^n \). By the Equivalency Theorem 3.4, we must have

\[
x = x_n.
\]

Thus the Cayley transform method and the Padé recursions produce exactly the same results, except that the arithmetic operations of inversion and multiplication have been rearranged. It was pointed out earlier that this can have a significant effect in the matrix case, since the Cayley transform approach is multiplication-rich compared to the Padé methods. We now extend the Cayley transform method to the case where \( x = -1 \) or where \(-1\) is an eigenvalue of \( X \) in the matrix case.

From (4.8) and (4.9),

\[
\tilde{x}_\nu = \frac{1 - y^\nu}{1 + y^\nu}.
\]

The next lemma shows that the right-hand side of (4.12) is well defined for any \( x \) which is not on the imaginary axis.

**Lemma 4.3.** Let \( x \in \mathbb{C}^+ \cup \mathbb{C}^- \). Then \((1 + x)^\nu + (1 - x)^\nu \neq 0 \) for any positive integer \( \nu \).

**Proof.** Suppose to the contrary that \((1 + x)^\nu + (1 - x)^\nu = 0\). Then \( x \neq 1 \), so \((1 + x)^\nu/(1 - x)^\nu = -1\). This means that \((1 + x)/(1 - x)\) is the \( \nu \)th root of \(-1\): \((1 + x)/(1 - x) = e^{i\theta}\) where \( \theta \) is not an odd multiple of \( \pi \) (else \( x = +\infty \)). Solving for \( x \) we find \( x = (\sin \theta/(1 + \cos \theta))i \notin \mathbb{C}^+ \cup \mathbb{C}^- \), which is a contradiction. \( \square \)
We end this section with a short discussion of Laurent methods, which are polynomial iterations in \( x \) and \( x^{-1} \) of the form

\[
x_{n+1} = \sum_{j=-\nu}^{\nu} b_j x^n.
\]

These methods are motivated by a desire to generate a "multiplication-rich" iteration once \( X^{-1} \) has been computed. For example, Newton's method is of this form with \( \nu = 1 \), \( b_{-1} = \frac{1}{2} = b_1 \). If we let

\[
L(x) = \sum_{j=-\nu}^{\nu} b_j x^j,
\]

then the coefficients \( b_j \) can be determined from \( L(1) = 1, L'(1) = 0, \ldots, L^{(2\nu-1)}(1) = 0 \). (Other conditions which assign specified eigenvalues to \( x = 1 \) can be used as well.)
Because of symmetry reasons we generally want $L$ to be an odd function, $L(-x) = -L(x)$, so that $v$ should be odd and $b_j = 0$ whenever $j$ is even. After Newton’s method ($v = 1$) the next two methods ($v = 3$ and $v = 5$) are of order four and six, respectively, and take the form

$$x_{n+1} = \frac{1}{16} \left(-\frac{1}{x_n} + \frac{9}{x_n^3} + 9x_n - x_n^3\right)$$

for $v = 3$,

$$x_{n+1} = \frac{1}{7552} \left(\frac{73}{x_n^5} - \frac{660}{x_n^3} + \frac{4270}{x_n} + 4580x_n - 815x_n^3 + 104x_n^5\right)$$

for $v = 5$.

These methods are multiplication-rich in the sense that they require one matrix inversion and $v + 1$ multiplies per step. However, they are not globally convergent and, in fact, the region of convergence for these two methods does not even include the set $|x^2 - 1| < 1$, as do the Padé methods. This is illustrated in Figs. 10 and 11, where the set $|x^2 - 1| = 1$ is included for comparison.

5. Matrix convergence. In this section we show that convergence in the matrix case is determined by the scalar convergence of the eigenvalues. This allows us to apply the scalar convergence results of the previous sections to the matrix case.

The following general result is the key to this process.

**Lemma 5.1.** Let $R = R(x)$ be an odd rational function such that $R(1) = 1$ and $R'(1) = 0$. Let $x_0 \in \mathbb{C}^+ \cup \mathbb{C}^-$ such that $\lim_{n \to +\infty} x_n = \operatorname{sgn}(x_0)$, where $x_{n+1} = R(x_n)$. Let $X_0$ be a Jordan block of the form

$$X_0 = \begin{bmatrix}
0 & 1 & 0 & \cdots & 0 \\
0 & 0 & 1 & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & \cdots & \cdots & \cdots & 1 \\
x_0 & & & & 
\end{bmatrix}.$$  

Then the matrix sequence defined by $X_{n+1} = R(X_n)$ satisfies $\lim_{n \to +\infty} X_n = \operatorname{sgn}(X_0)$.

**Proof.** Let $R_1(x) = R(x)$, $R_2(x) = R(R(x))$, and in general $R_{n+1}(x) = R(R_n(x))$. Because $X_0$ is a Jordan block,

$$X_n = R_n(X_0) = \begin{bmatrix}
a_1 & a_2 & \cdots & a_v \\
a_1 & \cdots & \cdots & \cdots \\
0 & \cdots & a_2 & a_1 \\
\end{bmatrix},$$

where $v$ is the order of $X_0$ and

$$a_j = a_j(n) = \frac{1}{(j-1)!} \left. \frac{d^{j-1}}{dx^{j-1}} R_n \right|_{x_0}.$$  

Thus $a_1(n) = R_n(x_0) = x_n \to \operatorname{sgn}(x_0)$ by assumption. For $j = 2$,

$$a_2(n) = \left. \frac{dR_n}{dx} \right|_{x_0} = \left. \frac{dR}{dx} \right|_{R_{n-1}(x_0)} \left. \frac{dR_{n-1}}{dx} \right|_{x_0} = \left. \frac{dR}{dx} \right|_{x_{n-1}} a_2(n-1),$$

where $x_{n-1} = R_{n-1}(x_0)$, and so on.
by the chain rule. But \( \lim_{n \to +\infty} dR/dx|_{x_{n-1}} = dR/dx|_{\text{sgn}(x_0)} = 0 \) since \( \text{sgn}(x_0) = \pm 1 \) and \( dR/dx(\pm 1) = 0 \) by assumption. Thus \( a_2(n) \to 0 \).

As an induction hypothesis suppose that \( a_j(n) \to 0 \) for \( 2 \leq j \leq i - 1 \). Then by the chain rule

\[
a_i(n) = \frac{dR}{dx}|_{x_{n-1}} a_i(n-1) + r_n,
\]

where \( r_n \) has a fixed form, independent of \( n \), involving sums and products of \( a_j \) for \( 2 \leq j \leq i - 1 \). Thus \( r_n \to 0 \) by the induction hypothesis. Since \( dR/dx|_{x_{n-1}} \) also tends to zero we have \( \lim_{n \to +\infty} a_i(n) = 0 \). This means that \( \lim_{n \to +\infty} x_n = \text{sgn}(x_0)I = \text{sgn}(x_0) \).

Using Lemma 5.1, we obtain the matrix analogues of Theorems 3.1-3.4 and the Cayley power method.

**Theorem 5.2.** Let \( k \geq m - 1 \) and assume that the eigenvalues of \( X_0 \) lie in \( \mathbb{C}^+ \cup \mathbb{C}^- \). Assume that \( \| I - X_0^2 \| < 1 \) and define

\[
X_{n+1} = -X_nP_{km}(I - X_n^2)Q^{-1}_{km}(I - X_n^2).
\]

Then

\[
\| I - X_n^2 \| < \| I - X_0^2 \|^{(k + m + 1)},
\]

and

\[
\lim_{n \to +\infty} X_n = \text{sgn}(x_0).
\]

**Proof.** The condition \( \| I - X_0^2 \| < 1 \) ensures that \( |1 - \lambda^2| < 1 \) for any eigenvalue \( \lambda \) of \( X_0 \). Hence by Theorem 3.1, the eigenvalues \( \lambda_{n,i} \) for \( X_n \) converge to \( \text{sgn}(\lambda_{0,i}) \). By Lemma 5.1 and the definition of \( \text{sgn}(X_0) \) in terms of its Jordan form, (5.3) is true. The matrix inequality (5.2) can be obtained by using the matrix analogue of the arguments in the proof of Theorem 3.1.

**Theorem 5.3.** Let \( \Lambda(X_0) \subset \mathbb{C}^+ \cup \mathbb{C}^- \) and assume that \( k = m \) or \( k = m - 1 \) in (4.3). Then for \( \gamma = k + m + 1 \)

\[
\lim_{n \to +\infty} X_n = \text{sgn} X_0 = S,
\]

(5.5) \((S - X_n)(S + X_n)^{-1} = [(S - X_0)(S + X_0)^{-1}]^\gamma\),

and

\[
X_n = (A^\gamma - B^\gamma)(A^\gamma + B^\gamma)^{-1},
\]

where

\[
A = I + X_0 \quad \text{and} \quad B = I - X_0.
\]

**Proof.** By Theorem 3.3 the eigenvalues of \( X_0 \) converge under (3.1) to the appropriate value of \( \pm 1 \). By Lemma 5.1, this means that \( \lim_{n \to +\infty} X_n = \text{sgn}(X_0) \). Equation (5.5) is obtained by considering the individual Jordan blocks and using (3.15). Similarly, use Lemma 4.3 to see that (5.6) is true for each Jordan block and hence for \( X_n \) itself.

**6. Conclusion.** In this paper, we have presented a theory of rational recursions for the matrix sign function, including Padé, Laurent, Cayley transform, and eigenvalue assignment methods. Of particular interest are the globally convergent main diagonal Padé iterations and their multiplication-rich Cayley transform equivalents.
Several important aspects concerning the numerical evaluation of sign function iterations have been treated elsewhere and so have not been discussed here. For example, scaling can significantly increase the speed of convergence of $X_n$ to $\text{sgn} \ (X)$ as noted in [3] and [4]; for scaling related to the polar decomposition, see [11]. The choice of optimal and nearly optimal scaling constants for Newton’s method is discussed at length in [15] and it is not hard to adapt these results to the main diagonal Padé recursions. Similarly, the problem of estimating the sensitivity of the sign of a matrix is considered in [16], based on the work in [8], [14], and [20].

REFERENCES