Chapter 5
Matrix Sign Function

The scalar sign function is defined for $z \in \mathbb{C}$ lying off the imaginary axis by

$$\operatorname{sign}(z) = \begin{cases} 1, & \operatorname{Re} z > 0, \\ -1, & \operatorname{Re} z < 0. \end{cases}$$

The matrix sign function can be obtained from any of the definitions in Chapter 1. Note that in the case of the Jordan canonical form and interpolating polynomial definitions, the derivatives $\operatorname{sign}^{(k)}(z)$ are zero for $k \geq 1$. Throughout this chapter, $A \in \mathbb{C}^{n\times n}$ is assumed to have no eigenvalues on the imaginary axis, so that $\operatorname{sign}(A)$ is defined. Note that this assumption implies that A is nonsingular.

As we noted in Section 2.4, if $A = ZJZ^{-1}$ is a Jordan canonical form arranged so that $J = \operatorname{diag}(J_1, J_2)$, where the eigenvalues of $J_1 \in \mathbb{C}^{p\times p}$ lie in the open left half-plane and those of $J_2 \in \mathbb{C}^{q\times q}$ lie in the open right half-plane, then

$$\operatorname{sign}(A) = Z \begin{bmatrix} -I_p & 0 \\ 0 & I_q \end{bmatrix} Z^{-1}. \tag{5.1}$$

Two other representations have some advantages. First is the particularly concise formula (see (5.5))

$$\operatorname{sign}(A) = A(A^2)^{-1/2}, \tag{5.2}$$

which generalizes the scalar formula $\operatorname{sign}(z) = z/(z^2)^{1/2}$. Recall that $B^{1/2}$ denotes the principal square root of B (see Section 1.7). Note that A having no pure imaginary eigenvalues is equivalent to A^2 having no eigenvalues on $\mathbb{R}^-$. Next, $\operatorname{sign}(A)$ has the integral representation (see Problem 5.3)

$$\operatorname{sign}(A) = \frac{2}{\pi} A \int_0^\infty (t^2 I + A^2)^{-1}\, dt. \tag{5.3}$$

Some properties of $\operatorname{sign}(A)$ are collected in the following theorem.

Theorem 5.1 (properties of the sign function). *Let $A \in \mathbb{C}^{n\times n}$ have no pure imaginary eigenvalues and let $S = \operatorname{sign}(A)$. Then*

(a) *$S^2 = I$ (S is involutory);*

(b) *S is diagonalizable with eigenvalues ± 1;*

(c) *$SA = AS$;*

(d) *if A is real then S is real;*

(e) *$(I+S)/2$ and $(I-S)/2$ are projectors onto the invariant subspaces associated with the eigenvalues in the right half-plane and left half-plane, respectively.*

Proof. The properties follow from (5.1)–(5.3). Of course, properties (c) and (d) hold more generally for matrix functions, as we know from Chapter 1 (see Theorem 1.13 (a) and Theorem 1.18). □

Although $\operatorname{sign}(A)$ is a square root of the identity matrix, it is not equal to I or $-I$ unless the spectrum of A lies entirely in the open right half-plane or open left half-plane, respectively. Hence, in general, $\operatorname{sign}(A)$ is a nonprimary square root of I. Moreover, although $\operatorname{sign}(A)$ has eigenvalues ± 1, its norm can be arbitrarily large.

The early appearance of this chapter in the book is due to the fact that the sign function plays a fundamental role in iterative methods for matrix roots and the polar decomposition. The definition (5.2) might suggest that the sign function is a "special case" of the square root. The following theorem, which provides an explicit formula for the sign of a block 2×2 matrix with zero diagonal blocks, shows that, if anything, the converse is true: the square root can be obtained from the sign function (see (5.4)). The theorem will prove useful in the next three chapters.

Theorem 5.2 (Higham, Mackey, Mackey, and Tisseur). *Let $A, B \in \mathbb{C}^{n\times n}$ and suppose that AB (and hence also BA) has no eigenvalues on $\mathbb{R}^-$. Then*

$$\operatorname{sign}\left(\begin{bmatrix} 0 & A \\ B & 0 \end{bmatrix}\right) = \begin{bmatrix} 0 & C \\ C^{-1} & 0 \end{bmatrix},$$

where $C = A(BA)^{-1/2}$.

Proof. The matrix $P = \left[\begin{smallmatrix} 0 & A \\ B & 0 \end{smallmatrix}\right]$ cannot have any eigenvalues on the imaginary axis, because if it did then $P^2 = \left[\begin{smallmatrix} AB & 0 \\ 0 & BA \end{smallmatrix}\right]$ would have an eigenvalue on $\mathbb{R}^-$. Hence $\operatorname{sign}(P)$ is defined and

$$\begin{aligned} \operatorname{sign}(P) = P(P^2)^{-1/2} &= \begin{bmatrix} 0 & A \\ B & 0 \end{bmatrix} \begin{bmatrix} AB & 0 \\ 0 & BA \end{bmatrix}^{-1/2} \\ &= \begin{bmatrix} 0 & A \\ B & 0 \end{bmatrix} \begin{bmatrix} (AB)^{-1/2} & 0 \\ 0 & (BA)^{-1/2} \end{bmatrix} \\ &= \begin{bmatrix} 0 & A(BA)^{-1/2} \\ B(AB)^{-1/2} & 0 \end{bmatrix} =: \begin{bmatrix} 0 & C \\ D & 0 \end{bmatrix}. \end{aligned}$$

Since the square of the matrix sign of any matrix is the identity,

$$I = (\operatorname{sign}(P))^2 = \begin{bmatrix} 0 & C \\ D & 0 \end{bmatrix}^2 = \begin{bmatrix} CD & 0 \\ 0 & DC \end{bmatrix},$$

so $D = C^{-1}$. Alternatively, Corollary 1.34 may be used to see more directly that $CD = A(BA)^{-1/2}B(AB)^{-1/2}$ is equal to I. □

A special case of the theorem, first noted by Higham [274, 1997], is

$$\operatorname{sign}\left(\begin{bmatrix} 0 & A \\ I & 0 \end{bmatrix}\right) = \begin{bmatrix} 0 & A^{1/2} \\ A^{-1/2} & 0 \end{bmatrix}. \tag{5.4}$$

In addition to the association with matrix roots and the polar decomposition (Chapter 8), the importance of the sign function stems from its applications to Riccati equations (Section 2.4), the eigenvalue problem (Section 2.5), and lattice QCD (Section 2.7).

In this chapter we first give perturbation theory for the matrix sign function and identify appropriate condition numbers. An expensive, but stable, Schur method for $\mathrm{sign}(A)$ is described. Then Newton's method and a rich Padé family of iterations, having many interesting properties, are described and analyzed. How to scale and how to terminate the iterations are discussed. Then numerical stability is considered, with the very satisfactory conclusion that all sign iterations of practical interest are stable. Numerical experiments illustrating these various features are presented. Finally, best L_∞ rational approximation via Zolotarev's formulae, of interest for Hermitian matrices, is described.

As we will see in Chapter 8, the matrix sign function has many connections with the polar decomposition, particularly regarding iterations for computing it. Some of the results and ideas in Chapter 8 are applicable, with suitable modification, to the sign function, but are not discussed here to avoid repetition. See, for example, Problem 8.26.

5.1. Sensitivity and Conditioning

Associated with the matrix sign function is the *matrix sign decomposition*

$$A = SN, \qquad S = \mathrm{sign}(A), \quad N = (A^2)^{1/2}. \tag{5.5}$$

To establish the decomposition note that $N = S^{-1}A = SA$. Since S commutes with A, $N^2 = A^2$, and since the spectrum of SA lies in the open right half-plane, $N = (A^2)^{1/2}$.

The matrix sign factor N is useful in characterizing the Fréchet derivative of the matrix sign function.

Let $S + \Delta S = \mathrm{sign}(A + \Delta A)$, where the sign function is assumed to be defined in a ball of radius $\|\Delta A\|$ about A. The definition (3.6) of Fréchet derivative says that

$$\Delta S - L(A, \Delta A) = o(\|\Delta A\|), \tag{5.6}$$

where $L(A, \Delta A)$ is the Fréchet derivative of the matrix sign function at A in the direction ΔA. Now from $(A + \Delta A)(S + \Delta S) = (S + \Delta S)(A + \Delta A)$ we have

$$A\Delta S - \Delta S A = S\Delta A - \Delta A S + \Delta S \Delta A - \Delta A \Delta S = S\Delta A - \Delta A S + o(\|\Delta A\|), \tag{5.7}$$

since $\Delta S = O(\|\Delta A\|)$. Moreover, $(S + \Delta S)^2 = I$ gives

$$S\Delta S + \Delta S S = -\Delta S^2 = o(\|\Delta A\|).$$

Premultiplying (5.7) by S and using the latter equation gives

$$N\Delta S + \Delta S N = \Delta A - S\Delta A S + o(\|\Delta A\|). \tag{5.8}$$

Theorem 5.3 (Kenney and Laub). *The Fréchet derivative $L = L_{\mathrm{sign}}(A, \Delta A)$ of the matrix sign function satisfies*

$$NL + LN = \Delta A - S\Delta A S, \tag{5.9}$$

where $A = SN$ is the matrix sign decomposition.

Proof. Since the eigenvalues of N lie in the open right half-plane, the Sylvester equation (5.9) has a unique solution L which is a linear function of ΔA and, in view of (5.8), differs from $\Delta S = \operatorname{sign}(A+\Delta A) - S$ by $o(\|\Delta A\|)$. Hence (5.6) implies that $L = L(A, \Delta A)$. □

By applying the vec operator and using the relation (B.16) we can rewrite (5.9) as

$$P \operatorname{vec}(L) = (I_{n^2} - S^T \otimes S) \operatorname{vec}(\Delta A),$$

where

$$P = I \otimes N + N^T \otimes I.$$

Hence

$$\begin{aligned}\max_{\|\Delta A\|_F = 1} \|L(A, \Delta A)\|_F &= \max_{\|\Delta A\|_F = 1} \|P^{-1}(I_{n^2} - S^T \otimes S) \operatorname{vec}(\Delta A)\|_2 \\ &= \|P^{-1}(I_{n^2} - S^T \otimes S)\|_2.\end{aligned}$$

The (relative) condition number of $\operatorname{sign}(A)$ in the Frobenius norm is therefore

$$\kappa_{\text{sign}}(A) := \operatorname{cond}_{\text{rel}}(\operatorname{sign}, A) = \|P^{-1}(I_{n^2} - S^T \otimes S)\|_2 \frac{\|A\|_F}{\|S\|_F}. \tag{5.10}$$

If $S = I$, which means that all the eigenvalues of A are in the open right half-plane, then $\operatorname{cond}(S) = 0$, which corresponds to the fact that the eigenvalues remain in this half-plane under sufficiently small perturbations of A.

To gain some insight into the condition number, suppose that A is diagonalizable: $A = ZDZ^{-1}$, where $D = \operatorname{diag}(\lambda_i)$. Then $S = ZD_SZ^{-1}$ and $N = ZD_NZ^{-1}$, where $D_S = \operatorname{diag}(\sigma_i)$ and $D_N = \operatorname{diag}(\sigma_i\lambda_i)$, with $\sigma_i = \operatorname{sign}(\lambda_i)$. Hence

$$\kappa_{\text{sign}}(A) = \|(Z^{-T} \otimes Z) \cdot (I \otimes D_N + D_N \otimes I)^{-1}(I_{n^2} - D_S^T \otimes D_S) \cdot (Z^T \otimes Z^{-1})\|_2 \frac{\|A\|_F}{\|S\|_F}.$$

The diagonal matrix in the middle has elements $(1 - \sigma_i\sigma_j)/(\sigma_i\lambda_i + \sigma_j\lambda_j)$, which are either zero or of the form $2/|\lambda_i - \lambda_j|$. Hence

$$\kappa_{\text{sign}}(A) \le 2\kappa_2(Z)^2 \max\left\{ \frac{1}{|\lambda_i - \lambda_j|} : \operatorname{Re}\lambda_i \operatorname{Re}\lambda_j < 0 \right\} \frac{\|A\|_F}{\|S\|_F}. \tag{5.11}$$

Equality holds in this bound for normal A, for which Z can be taken to unitary. The gist of (5.11) is that the condition of S is bounded in terms of the minimum distance between eigenvalues across the imaginary axis and the square of the condition of the eigenvectors. Note that (5.11) is precisely the bound obtained by applying Theorem 3.15 to the matrix sign function.

One of the main uses of κ_{sign} is to indicate the sensitivity of $\operatorname{sign}(A)$ to perturbations in A, through the perturbation bound (3.3), which we rewrite here for the sign function as

$$\frac{\|\operatorname{sign}(A+E) - \operatorname{sign}(A)\|_F}{\|\operatorname{sign}(A)\|_F} \le \kappa_{\text{sign}}(A) \frac{\|E\|_F}{\|A\|_F} + o(\|E\|_F). \tag{5.12}$$

This bound is valid as long as $\operatorname{sign}(A + tE)$ is defined for all $t \in [0, 1]$. It is instructive to see what can go wrong when this condition is not satisfied. Consider the example, from [347, 1995],

$$A = \operatorname{diag}(1, -\epsilon^2), \qquad E = \operatorname{diag}(0, 2\epsilon^2), \qquad 0 < \epsilon \ll 1.$$

We have $\operatorname{sign}(A) = \operatorname{diag}(1, -1)$ and $\operatorname{sign}(A + E) = I$. Because A is normal, (5.11) gives $\kappa_{\mathrm{sign}}(A) = (2/(1 + \epsilon^2))\|A\|_F/\sqrt{2}$. Hence the bound (5.12) takes the form

$$\frac{2}{\sqrt{2}} \le \frac{2}{\sqrt{2}(1+\epsilon^2)} 2\epsilon^2 + o(\epsilon^2) = 2\sqrt{2}\epsilon^2 + o(\epsilon^2).$$

This bound is clearly incorrect. The reason is that the perturbation E causes eigenvalues to cross the imaginary axis; therefore $\operatorname{sign}(A + tE)$ does not exist for all $t \in [0, 1]$. Referring back to the analysis at the start of this section, we note that (5.7) is valid for $\|\Delta A\|_F < \|E\|_F/3$, but does not hold for $\Delta A = E$, since then $\Delta S \ne O(\|\Delta A\|)$.

Another useful characterization of the Fréchet derivative is as the limit of a matrix iteration; see Theorem 5.7.

Consider now how to estimate $\kappa_{\mathrm{sign}}(A)$. We need to compute a norm of $B = P^{-1}(I_{n^2} - S^T \otimes S)$. For the 2-norm we can use Algorithm 3.20 (the power method). Alternatively, Algorithm 3.22 can be used to estimate the 1-norm. In both cases we need to compute $L(A, E)$, which if done via (5.9) requires solving a Sylvester equation involving N; this can be done via a matrix sign evaluation (see Section 2.4), since N is positive stable. We can compute $L^\star(X, E)$ in a similar fashion, solving a Sylvester equation of the same form. Alternatively, $L(A, E)$ can be computed using iteration (5.23) or estimated by finite differences. All these methods require $O(n^3)$ operations.

It is also of interest to understand the conditioning of the sign function for $A \approx \operatorname{sign}(A)$, which is termed the asymptotic conditioning. The next result provides useful bounds.

Theorem 5.4 (Kenney and Laub). *Let $A \in \mathbb{C}^{n\times n}$ have no pure imaginary eigenvalues and let $S = \operatorname{sign}(A)$. If $\|(A - S)S\|_2 < 1$, then*

$$\frac{\|S\|_2^2 - 1}{2(1 + \|(A-S)S\|_2)} \le \frac{\kappa_{\mathrm{sign}}(A)}{\|A\|_F/\|S\|_F} \le \frac{\|S\|_2^2 + 1}{2(1 - \|(A-S)S\|_2)}. \qquad (5.13)$$

In particular,

$$\frac{\|S\|_2^2 - 1}{2} \le \kappa_{\mathrm{sign}}(S) \le \frac{\|S\|_2^2 + 1}{2}. \qquad (5.14)$$

Proof. We need to bound $\|L_{\mathrm{sign}}(A)\|_F = \kappa_{\mathrm{sign}}(A)\|S\|_F/\|A\|_F$. Let $\Delta S = L_{\mathrm{sign}}(A, \Delta A)$. Then by (5.9),

$$N\Delta S + \Delta S N = \Delta A - S\Delta A S,$$

where $N = SA = AS$. Defining $G = AS - S^2 = N - I$, we have

$$2\Delta S = \Delta A - S\Delta A S - G\Delta S - \Delta S G. \qquad (5.15)$$

Taking norms, using (B.7), leads to

$$\|\Delta S\|_F \le \frac{(\|S\|_2^2 + 1)\|\Delta A\|_F}{2(1 - \|G\|_2)},$$

which gives the upper bound.

Now let $\sigma = \|S\|_2$ and $Sv = \sigma u$, $u^*S = \sigma v^*$, where u and v are (unit-norm) left and right singular vectors, respectively. Putting $\Delta A = vu^*$ in (5.15) gives

$$2\Delta S = vu^* - Svu^*S - G\Delta S - \Delta S G = vu^* - \sigma^2 uv^* - G\Delta S - \Delta S G.$$

Hence

$$(\|S\|_2^2 - 1)\|\varDelta A\|_F = (\sigma^2 - 1)\|\varDelta A\|_F \le 2\|\varDelta S\|_F(1 + \|G\|_2),$$

which implies the lower bound.

Setting $A = S$ in (5.13) gives (5.14). □

Theorem 5.4 has something to say about the attainable accuracy of a computed sign function. In computing $S = \operatorname{sign}(A)$ we surely cannot do better than if we computed $\operatorname{sign}(fl(S))$. But Theorem 5.4 says that relative errors in S can be magnified when we take the sign by as much as $\|S\|^2/2$, so we cannot expect a relative error in our computed sign smaller than $\|S\|^2u/2$, whatever the method used.

5.2. Schur Method

The first method that we consider for computing $\operatorname{sign}(A)$ is expensive but has excellent numerical stability. Because the method utilizes a Schur decomposition it is not suitable for the applications in Sections 2.4 and 2.5, since those problems can be solved directly by the use of a Schur decomposition, without explicitly forming the sign function.

Let $A \in \mathbb{C}^{n\times n}$ have the Schur decomposition $A = QTQ^*$, where Q is unitary and T is upper triangular. Then $\operatorname{sign}(A) = Q\operatorname{sign}(T)Q^*$ (see Theorem 1.13 (c)). The problem therefore reduces to computing $U = \operatorname{sign}(T)$, and clearly U is upper triangular with $u_{ii} = \operatorname{sign}(t_{ii}) = \pm 1$ for all i. We will determine u_{ij} from the equation $U^2 = I$ when possible (namely, when $u_{ii} + u_{jj} \ne 0$), and from $TU = UT$ otherwise (in which case $t_{ii} \ne t_{jj}$), employing the Parlett recurrence (Algorithm 4.13) in this second case.

Algorithm 5.5 (Schur method). Given $A \in \mathbb{C}^{n\times n}$ having no pure imaginary eigenvalues, this algorithm computes $S = \operatorname{sign}(A)$ via a Schur decomposition.

```
Compute a (complex) Schur decomposition $A = QTQ^*$.
$u_{ii} = \operatorname{sign}(t_{ii})$, $i = 1\colon n$
for $j = 2\colon n$
    for $i = j-1\colon -1\colon 1$
        $u_{ij} = \begin{cases} -\dfrac{\sum_{k=i+1}^{j-1} u_{ik}u_{kj}}{u_{ii}+u_{jj}}, & u_{ii}+u_{jj} \ne 0, \\ t_{ij}\dfrac{u_{ii}-u_{jj}}{t_{ii}-t_{jj}} + \dfrac{\sum_{k=i+1}^{j-1}(u_{ik}t_{kj} - t_{ik}u_{kj})}{t_{ii}-t_{jj}}, & u_{ii}+u_{jj} = 0. \end{cases}$
    end
end
$S = QUQ^*$
```

Cost: $25n^3$ flops for the Schur decomposition plus between $n^3/3$ and $2n^3/3$ flops for U and $3n^3$ flops to form S: about $28\frac{2}{3}n^3$ flops in total.

It is worth noting that the sign of an upper triangular matrix T will usually have some zero elements in the upper triangle. Indeed, suppose for some $j > i$ that t_{ii}, $t_{i+1,i+1}$, …, t_{jj} all have the same sign, and let $T_{ij} = T(i\colon j, i\colon j)$. Then, since all the eigenvalues of T_{ij} have the same sign, the corresponding block $S(i\colon j, i\colon j)$ of $S = \operatorname{sign}(T)$ is $\pm I$. This fact could be exploited by reordering the Schur form so that the diagonal of T is grouped according to sign. Then $\operatorname{sign}(T)$ would have the form

$\left[\begin{smallmatrix}\pm I & W\\ 0 & \mp I\end{smallmatrix}\right]$, where W is computed by the Parlett recurrence. The cost of the reordering may or may not be less than the cost of (redundantly) computing zeros from the first expression for u_{ij} in Algorithm 5.5.

5.3. Newton's Method

The most widely used and best known method for computing the sign function is the Newton iteration, due to Roberts:

Newton iteration (matrix sign function):

$$X_{k+1} = \frac{1}{2}(X_k + X_k^{-1}), \qquad X_0 = A. \tag{5.16}$$

The connection of this iteration with the sign function is not immediately obvious, but in fact the iteration can be derived by applying Newton's method to the equation $X^2 = I$ (see Problem 5.8), and of course $\operatorname{sign}(A)$ is one solution of this equation (Theorem 5.1 (a)). The following theorem describes the convergence of the iteration.

Theorem 5.6 (convergence of the Newton sign iteration). *Let $A \in \mathbb{C}^{n\times n}$ have no pure imaginary eigenvalues. Then the Newton iterates X_k in* (5.16) *converge quadratically to $S = \operatorname{sign}(A)$, with*

$$\|X_{k+1} - S\| \le \frac{1}{2}\|X_k^{-1}\|\,\|X_k - S\|^2 \tag{5.17}$$

for any consistent norm. Moreover, for $k \ge 1$,

$$X_k = (I - G_0^{2^k})^{-1}(I + G_0^{2^k})S, \quad \textit{where } G_0 = (A - S)(A + S)^{-1}. \tag{5.18}$$

Proof. For $\lambda = re^{i\theta}$ we have $\lambda + \lambda^{-1} = (r + r^{-1})\cos\theta + i(r - r^{-1})\sin\theta$, and hence eigenvalues of X_k remain in their open half-plane under the mapping (5.16). Hence X_k is defined and nonsingular for all k. Moreover, $\operatorname{sign}(X_k) = \operatorname{sign}(X_0) = S$, and so $X_k + S = X_k + \operatorname{sign}(X_k)$ is also nonsingular.

Clearly the X_k are (rational) functions of A and hence, like A, commute with S. Then

$$\begin{aligned} X_{k+1} \pm S &= \frac{1}{2}\left(X_k + X_k^{-1} \pm 2S\right)\\ &= \frac{1}{2}X_k^{-1}\left(X_k^2 \pm 2X_kS + I\right)\\ &= \frac{1}{2}X_k^{-1}(X_k \pm S)^2, \end{aligned} \tag{5.19}$$

and hence

$$(X_{k+1} - S)(X_{k+1} + S)^{-1} = \left((X_k - S)(X_k + S)^{-1}\right)^2.$$

Defining $G_k = (X_k - S)(X_k + S)^{-1}$, we have $G_{k+1} = G_k^2 = \cdots = G_0^{2^{k+1}}$. Now $G_0 = (A - S)(A + S)^{-1}$ has eigenvalues $(\lambda - \operatorname{sign}(\lambda))/(\lambda + \operatorname{sign}(\lambda))$, where $\lambda \in \Lambda(A)$, all of which lie inside the unit circle since λ is not pure imaginary. Since $G_k = G_0^{2^k}$ and $\rho(G_0) < 1$, by a standard result (B.9) $G_k \to 0$ as $k \to \infty$. Hence

$$X_k = (I - G_k)^{-1}(I + G_k)S \to S \quad \text{as } k \to \infty. \tag{5.20}$$

The norm inequality (5.17), which displays the quadratic convergence, is obtained by taking norms in (5.19) with the minus sign. $\square$

Theorem 5.6 reveals quadratic convergence of the Newton iteration, but also displays in (5.18) precisely how convergence occurs: through the powers of the matrix G_0 converging to zero. Since for any matrix norm,

$$\|G_0^{2^k}\| \geq \rho(G_0^{2^k}) = \left(\max_{\lambda \in \Lambda(A)} \frac{|\lambda - \operatorname{sign}(\lambda)|}{|\lambda + \operatorname{sign}(\lambda)|}. \right)^{2^k}, \tag{5.21}$$

It is clear that convergence will be slow if either $\rho(A) \gg 1$ or A has an eigenvalue close to the imaginary axis. We return to the speed of convergence in Section 5.5. For the behaviour of the iteration when it does not converge, see Problem 5.11.

The Newton iteration provides one of the rare circumstances in numerical analysis where the explicit computation of a matrix inverse is required. One way to try to remove the inverse from the formula is to approximate it by one step of Newton's method for the matrix inverse, which has the form $Y_{k+1} = Y_k(2I - BY_k)$ for computing B^{-1}; this is known as the Newton–Schulz iteration [512, 1933] (see Problem 7.8). Replacing X_k^{-1} by $X_k(2I - X_k^2)$ in (5.16) (having taken $Y_k = B = X_k$) gives

Newton–Schulz iteration:

$$X_{k+1} = \frac{1}{2} X_k (3I - X_k^2), \qquad X_0 = A. \tag{5.22}$$

This iteration is multiplication-rich and retains the quadratic convergence of Newton's method. However, it is only locally convergent, with convergence guaranteed for $\|I - A^2\| < 1$; see Theorem 5.8.

The Newton iteration also provides a way of computing the Fréchet derivative of the sign function.

Theorem 5.7 (Kenney and Laub). *Let $A \in \mathbb{C}^{n\times n}$ have no pure imaginary eigenvalues. With X_k defined by the Newton iteration* (5.16)*, let*

$$Y_{k+1} = \frac{1}{2}(Y_k - X_k^{-1} Y_k X_k^{-1}), \qquad Y_0 = E. \tag{5.23}$$

Then $\lim_{k\to\infty} Y_k = L_{\text{sign}}(A, E)$.

Proof. Denote by B_k the Newton sign iterates (5.16) for the matrix $B = \left[\begin{smallmatrix} A & E \\ 0 & A \end{smallmatrix}\right]$, which clearly has no pure imaginary eigenvalues. It is easy to show by induction that $B_k = \left[\begin{smallmatrix} X_k & Y_k \\ 0 & X_k \end{smallmatrix}\right]$. By Theorem 5.6 and (3.16) we have

$$B_k \to \operatorname{sign}(B) = \begin{bmatrix} \operatorname{sign}(A) & L_{\text{sign}}(A, E) \\ 0 & \operatorname{sign}(A) \end{bmatrix}.$$

The result follows on equating the (1,2) blocks. $\square$

5.4. The Padé Family of Iterations

The Newton iteration is by no means the only rational matrix iteration for computing the matrix sign function. A variety of other iterations have been derived, with various aims, including to avoid matrix inversion in favour of matrix multiplication, to achieve a higher order of convergence, and to be better suited to parallel computation. Ad hoc manipulations can be used to derive new iterations, as we now indicate for the scalar case. By setting $y_k = x_k^{-1}$ in the Newton formula $x_{k+1} = (x_k + x_k^{-1})/2$, we obtain the "inverse Newton" variant

$$y_{k+1} = \frac{2y_k}{y_k^2 + 1}, \qquad y_0 = a, \tag{5.24}$$

which has quadratic convergence to $\mathrm{sign}(a)$. Combining two Newton steps yields $y_{k+2} = (y_k^4 + 6y_k^2 + 1)/(4y_k(y_k^2 + 1))$, and we can thereby define the quartically convergent iteration

$$y_{k+1} = \frac{y_k^4 + 6y_k^2 + 1}{4y_k(y_k^2 + 1)}, \qquad y_0 = a.$$

While a lot can be done using arguments such as these, a more systematic development is preferable. We describe an elegant Padé approximation approach, due to Kenney and Laub [343, 1991], that yields a whole table of methods containing essentially all those of current interest.

For non–pure imaginary $z \in \mathbb{C}$ we can write

$$\mathrm{sign}(z) = \frac{z}{(z^2)^{1/2}} = \frac{z}{(1 - (1 - z^2))^{1/2}} = \frac{z}{(1 - \xi)^{1/2}}, \tag{5.25}$$

where $\xi = 1 - z^2$. Hence the task of approximating $\mathrm{sign}(z)$ leads to that of approximating

$$h(\xi) = (1 - \xi)^{-1/2}, \tag{5.26}$$

where we may wish to think of ξ as having magnitude less than 1. Now h is a particular case of a hypergeometric function and hence much is known about $[\ell/m]$ Padé approximants $r_{\ell m}(\xi) = p_{\ell m}(\xi)/q_{\ell m}(\xi)$ to h, including explicit formulae for $p_{\ell m}$ and $q_{\ell m}$. (See Section 4.4.2 for the definition of Padé approximants.) Kenney and Laub's idea is to set up the family of iterations

$$x_{k+1} = f_{\ell m}(x_k) := x_k \frac{p_{\ell m}(1 - x_k^2)}{q_{\ell m}(1 - x_k^2)}, \qquad x_0 = a. \tag{5.27}$$

Table 5.1 shows the first nine iteration functions $f_{\ell m}$ from this family. Note that f_{11} gives Halley's method (see Problem 5.12), while f_{10} gives the Newton–Schulz iteration (5.22). The matrix versions of the iterations are defined in the obvious way:

Padé iteration:

$$X_{k+1} = X_k p_{\ell m}(I - X_k^2) q_{\ell m}(I - X_k^2)^{-1}, \qquad X_0 = A. \tag{5.28}$$

Two key questions are "what can be said about the convergence of (5.28)?" and "how should the iteration be evaluated?"

The convergence question is answered by the following theorem.

Table 5.1. *Iteration functions $f_{\ell m}$ from the Padé family* (5.27).

	$m=0$	$m=1$	$m=2$
$\ell=0$	x	$\dfrac{2x}{1+x^2}$	$\dfrac{8x}{3+6x^2-x^4}$
$\ell=1$	$\dfrac{x}{2}(3-x^2)$	$\dfrac{x(3+x^2)}{1+3x^2}$	$\dfrac{4x(1+x^2)}{1+6x^2+x^4}$
$\ell=2$	$\dfrac{x}{8}(15-10x^2+3x^4)$	$\dfrac{x}{4}\dfrac{(15+10x^2-x^4)}{1+5x^2}$	$\dfrac{x(5+10x^2+x^4)}{1+10x^2+5x^4}$

Theorem 5.8 (convergence of Padé iterations). *Let $A \in \mathbb{C}^{n\times n}$ have no pure imaginary eigenvalues. Consider the iteration* (5.28) *with $\ell + m > 0$ and any subordinate matrix norm.*

(a) *For $\ell \ge m-1$, if $\|I-A^2\| < 1$ then $X_k \to \operatorname{sign}(A)$ as $k\to\infty$ and $\|I-X_k^2\| < \|I-A^2\|^{(\ell+m+1)^k}$.*

(b) *For $\ell = m-1$ and $\ell = m$,*

$$(S-X_k)(S+X_k)^{-1} = \left[(S-A)(S+A)^{-1}\right]^{(\ell+m+1)^k}$$

and hence $X_k \to \operatorname{sign}(A)$ as $k\to\infty$.

Proof. See Kenney and Laub [343, 1991]. □

Theorem 5.8 shows that the iterations with $\ell = m-1$ and $\ell = m$ are globally convergent, while those with $\ell \ge m+1$ have local convergence, the convergence rate being $\ell+m+1$ in every case.

We now concentrate on the cases $\ell = m-1$ and $\ell = m$ which we call the *principal Padé iterations.* For these ℓ and m we define

$$g_r(x) \equiv g_{\ell+m+1}(x) = f_{\ell m}(x). \tag{5.29}$$

The g_r are the iteration functions from the Padé table taken in a zig-zag fashion from the main diagonal and first superdiagonal:

$$g_1(x) = x, \qquad g_2(x) = \frac{2x}{1+x^2}, \qquad g_3(x) = \frac{x(3+x^2)}{1+3x^2},$$
$$g_4(x) = \frac{4x(1+x^2)}{1+6x^2+x^4}, \quad g_5(x) = \frac{x(5+10x^2+x^4)}{1+10x^2+5x^4}, \quad g_6(x) = \frac{x(6+20x^2+6x^4)}{1+15x^2+15x^4+x^6}.$$

We know from Theorem 5.8 that the iteration $X_{k+1} = g_r(X_k)$ converges to $\operatorname{sign}(X_0)$ with order r whenever $\operatorname{sign}(X_0)$ is defined. These iterations share some interesting properties that are collected in the next theorem.

Theorem 5.9 (properties of principal Padé iterations). *The principal Padé iteration function g_r defined in* (5.29) *has the following properties.*

(a) $g_r(x) = \dfrac{(1+x)^r-(1-x)^r}{(1+x)^r+(1-x)^r}$. *In other words, $g_r(x) = p_r(x)/q_r(x)$, where $p_r(x)$ and $q_r(x)$ are, respectively, the odd and even parts of $(1+x)^r$.*

(b) $g_r(x) = \tanh(r \operatorname{arctanh}(x))$.

(c) $g_r(g_s(x)) = g_{rs}(x)$ (*the semigroup property*).

(d) g_r *has the partial fraction expansion*

$$g_r(x) = \frac{2}{r} \sum_{i=0}^{\lceil \frac{r-2}{2} \rceil}{}' \frac{x}{\sin^2\left(\frac{(2i+1)\pi}{2r}\right) + \cos^2\left(\frac{(2i+1)\pi}{2r}\right) x^2}, \tag{5.30}$$

where the prime on the summation symbol denotes that the last term in the sum is halved when r is odd.

Proof.

(a) See Kenney and Laub [343, 1991, Thm. 3.2].

(b) Recalling that $\tanh(x) = (e^x - e^{-x})/(e^x + e^{-x})$, it is easy to check that

$$\operatorname{arctanh}(x) = \frac{1}{2} \log\left(\frac{1+x}{1-x}\right).$$

Hence

$$r \operatorname{arctanh}(x) = \log\left(\frac{1+x}{1-x}\right)^{r/2}.$$

Taking the tanh of both sides gives

$$\tanh(r \operatorname{arctanh}(x)) = \frac{\left(\frac{1+x}{1-x}\right)^{r/2} - \left(\frac{1-x}{1+x}\right)^{r/2}}{\left(\frac{1+x}{1-x}\right)^{r/2} + \left(\frac{1-x}{1+x}\right)^{r/2}} = \frac{(1+x)^r - (1-x)^r}{(1+x)^r + (1-x)^r} = g_r(x).$$

(c) Using (b) we have

$$\begin{aligned} g_r(g_s(x)) &= \tanh(r \operatorname{arctanh}(\tanh(s \operatorname{arctanh}(x)))) = \tanh(rs \operatorname{arctanh}(x)) \\ &= g_{rs}(x). \end{aligned}$$

(d) The partial fraction expansion is obtained from a partial fraction expansion for the hyperbolic tangent; see Kenney and Laub [345, 1994, Thm. 3]. □

Some comments on the theorem are in order. The equality in (a) is a scalar equivalent of (b) in Theorem 5.8, and it provides an easy way to generate the g_r. Property (c) says that one rth order principal Padé iteration followed by one sth order iteration is equivalent to one rsth order iteration. Whether or not it is worth using higher order iterations therefore depends on the efficiency with which the different iterations can be evaluated. The properties in (b) and (c) are analogous to properties of the Chebyshev polynomials. Figure 5.1 confirms, for real x, that $g_r(x) = \tanh(r \operatorname{arctanh}(x))$ approximates $\operatorname{sign}(x)$ increasingly well near the origin as r increases.

Some more insight into the convergence, or nonconvergence, of the iteration $x_{k+1} = g_r(x_k)$ from (5.29) can be obtained by using Theorem 5.8 (b) to write, in polar form,

$$\rho_{k+1} e^{i\theta_{k+1}} := (s - x_{k+1})(s + x_{k+1})^{-1} := \left[(s - x_k)(s + x_k)^{-1}\right]^r = \left[\rho_k e^{i\theta_k}\right]^r,$$

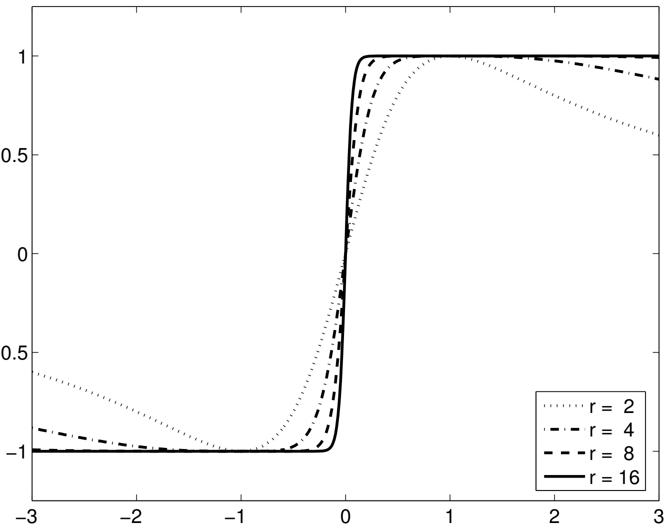

Figure 5.1. *The function* $g_r(x) = \tanh(r\,\mathrm{arctanh}(x))$ *for* $r = 2, 4, 8, 16$.

where $s = \mathrm{sign}(x_0)$. Hence

$$\rho_{k+1} = \rho_k^r, \qquad \theta_{k+1} = r\theta_k.$$

These relations illustrate the convergence of x_k to s for x_0 off the imaginary axis, since $\rho_0 < 1$. But they also reveal a chaotic aspect to the convergence through θ_k, which, in view of the periodicity of $e^{i\theta_k}$, can be written

$$\theta_{k+1} = r\theta_k \bmod 2\pi. \tag{5.31}$$

This recurrence can be described as a linear congruential random number generator [211, 2003, Sec. 1.2], [357, 1998, Sec. 3.2], though with a real, rather than integer, modulus. If x_0 is pure imaginary then the iteration does not converge: $\rho_k \equiv 1$, x_k remains pure imaginary for all k, and $(s - x_k)(s + x_k)^{-1}$ wanders chaotically around the circle of radius 1 centred at the origin; see also Problem 5.11.

We turn now to evaluation of the matrix iteration $X_{j+1} = g_r(X_j)$. As discussed in Section 4.4.2, several approaches are possible, based on different representations of the rational iteration function g_k. Evaluating $g_k(X_j)$ as the ratio of two polynomials may require more flops than via the partial fraction expansion (5.30). For example, evaluating g_3 from the formula $x(3 + x^2)/(1 + 3x^2)$ at an $n \times n$ matrix requires $6\frac{2}{3}n^3$ flops, whereas (5.30) can be written as

$$g_3(x) = \frac{1}{3}\left(x + \frac{8x}{1 + 3x^2}\right) \tag{5.32}$$

and evaluated in $4\frac{2}{3}n^3$ flops. An attractive feature of the partial fraction expansion (5.30) is that it comprises $\lceil \frac{r-2}{2} \rceil$ independent matrix inversions (or multiple right-hand side linear systems), which can be carried out in parallel.

5.5. Scaling the Newton Iteration

For scalar a, the Newton iteration (5.16) is

$$x_{k+1} = \frac{1}{2}(x_k + x_k^{-1}), \qquad x_0 = a, \tag{5.33}$$

which converges to $\mathrm{sign}(a) = \pm 1$ if a is not pure imaginary. This is precisely Newton's method for the square root of 1 and convergence is at a quadratic rate, as described by (5.17). Once the error is sufficiently small (in practice, less than, say, 0.5), successive errors decrease rapidly, each being approximately the square of the previous one (see (5.19)). However, initially convergence can be slow: if $|x_k| \gg 1$ then $x_{k+1} \approx x_k/2$ and the iteration is an expensive way to divide by 2! From (5.18) and (5.21) we also see that slow convergence will result when a is close to the imaginary axis. Therefore a way is needed of speeding up the initial phase of convergence in these unfavourable cases. For matrices, the same comments apply to the eigenvalues, because the Newton iteration (5.16) is effectively performing the scalar iteration (5.33) independently on each eigenvalue. However, the behaviour of the matrix iteration is not entirely determined by the eigenvalues: nonnormality of A can delay, though not prevent, convergence, as the following finite termination result shows.

Theorem 5.10 (Kenney and Laub). *For the Newton iteration* (5.16)*, if* X_k *has eigenvalues* ± 1 *for some* k *then* $X_{k+p} = \mathrm{sign}(A)$ *for* $2^p \ge m$*, where* m *is the size of the largest Jordan block of* X_k (*which is no larger than the size of the largest Jordan block of* A)*.*

Proof. Let X_k have the Jordan form $X_k = ZJ_kZ^{-1}$, where $J_k = D + N_k$, with $D = \mathrm{diag}(\pm 1) = \mathrm{sign}(J_k)$ and N_k strictly upper triangular. N_k has index of nilpotence m, that is, $N_k^m = 0$ but all lower powers are nonzero. We can restrict our attention to the convergence of the sequence beginning with J_k to $\mathrm{diag}(\pm 1)$, and so we can set $Z = I$. The next iterate, $X_{k+1} = D + N_{k+1}$, satisfies, in view of (5.19),

$$N_{k+1} = \frac{1}{2} X_k^{-1} N_k^2.$$

Since N_k has index of nilpotence m, N_{k+1} must have index of nilpotence $\lceil m/2 \rceil$. Applying this argument repeatedly shows that for $2^p \ge m$, N_{k+p} has index of nilpotence 1 and hence is zero, as required. That m is no larger than the order of the largest Jordan block of A follows from Theorem 1.36. □

An effective way to enhance the initial speed of convergence is to scale the iterates: prior to each iteration, X_k is replaced by $\mu_k X_k$, giving the scaled Newton iteration

Scaled Newton iteration:

$$X_{k+1} = \frac{1}{2}\big(\mu_k X_k + \mu_k^{-1} X_k^{-1}\big), \qquad X_0 = A. \tag{5.34}$$

As long as μ_k is real and positive, the sign of the iterates is preserved. Three main scalings have been proposed:

$$\text{determinantal scaling:} \quad \mu_k = |\det(X_k)|^{-1/n}, \tag{5.35}$$

$$\text{spectral scaling:} \quad \mu_k = \sqrt{\rho(X_k^{-1})/\rho(X_k)}, \tag{5.36}$$

$$\text{norm scaling:} \quad \mu_k = \sqrt{\|X_k^{-1}\|/\|X_k\|}. \tag{5.37}$$

For determinantal scaling, $|\det(\mu_k X_k)| = 1$, so that the geometric mean of the eigenvalues of $\mu_k X_k$ has magnitude 1. This scaling has the property that μ_k minimizes $d(\mu_k X_k)$, where

$$d(X) = \sum_{i=1}^{n} (\log |\lambda_i|)^2$$

and the are λ_i the eigenvalues of X. Hence determinantal scaling tends to bring the eigenvalues closer to the unit circle; see Problem 5.13.

When evaluating the determinantal scaling factor (5.35) some care is needed to avoid unnecessary overflow and underflow, especially when n is large. The quantity μ_k should be within the range of the floating point arithmetic, since its reciprocal has magnitude the geometric mean of the eigenvalues of X_k and hence lies between the moduli of the smallest and largest eigenvalues. But $\det(X_k)$ can underflow or overflow. Assuming that an LU factorization $PX_k = L_k U_k$ is computed, where U_k has diagonal elements u_{ii}, we can rewrite $\mu_k = |u_{11} \dots u_{nn}|^{-1/n}$ as $\mu_k = \exp((-1/n) \sum_{i=1}^{n} \log |u_{ii}|)$. The latter expression avoids underflow and overflow; however, cancellation in the summation can produce an inaccurate computed μ_k, so it may be desirable to use one of the summation methods from Higham [276, 2002, Chap. 4].

For spectral scaling, if $\lambda_n, \dots, \lambda_1$ are the eigenvalues of X_k ordered by increasing magnitude, then $\mu_k = |\lambda_1 \lambda_n|^{-1/2}$ and so $\mu_k X_k$ has eigenvalues of smallest and largest magnitude $|\mu_k \lambda_n| = |\lambda_n/\lambda_1|^{1/2}$ and $|\mu_k \lambda_1| = |\lambda_1/\lambda_n|^{1/2}$. If λ_1 and λ_n are real, then in the Cayley metric

$$\mathcal{C}(x, \operatorname{sign}(x)) := \begin{cases} |x-1|/|x+1|, & \operatorname{Re} x > 0, \\ |x+1|/|x-1|, & \operatorname{Re} x < 0, \end{cases}$$

$\mu_k \lambda_n$ is the same distance from $\operatorname{sign}(\lambda_n)$ as $\mu_k \lambda_1$ is from $\operatorname{sign}(\lambda_1)$, so in this case spectral scaling equalizes the extremal eigenvalue errors in the Cayley metric. The norm scaling (5.37) can be regarded as approximating the spectral scaling.

What can be said about the effectiveness of these scaling strategies? In general, all of them work well, but there are some specific advantages and disadvantages.

Spectral scaling is essentially optimal when all the eigenvalues of A are real; indeed it yields finite termination, as the following result shows.

Theorem 5.11 (Barraud). *Let the nonsingular matrix $A \in \mathbb{C}^{n \times n}$ have all real eigenvalues and let $S = \operatorname{sign}(A)$. Then, for the Newton iteration* (5.34) *with spectral scaling, $X_{d+p-1} = \operatorname{sign}(A)$, where d is the number of distinct eigenvalues of A and $2^p \geq m$, where m is the size of the largest Jordan block of A.*

Proof. We will need to use the following easily verified properties of the iteration function $f(x) = \frac{1}{2}(x + 1/x)$:

$$f(x) = f(1/x), \tag{5.38a}$$

$$0 \leq x_2 \leq x_1 \leq 1 \quad \text{or} \quad 1 \leq x_1 \leq x_2 \quad \Rightarrow \quad 1 \leq f(x_1) \leq f(x_2). \tag{5.38b}$$

Let the eigenvalues of $X_0 = A$, which we know to be real, be ordered $|\lambda_n| \leq \cdots \leq |\lambda_1|$. Then, from (5.36), $\mu_0 = |\lambda_n \lambda_1|^{-1/2}$, and the eigenvalues of $\mu_0 X_0$ have moduli lying between $|\mu_0 \lambda_n| = |\lambda_n/\lambda_1|^{1/2}$ and $|\mu_0 \lambda_1| = |\lambda_1/\lambda_n|^{1/2}$. These values are reciprocals, and hence by (5.38a), and since the eigenvalues are real, λ_n and λ_1 are mapped to values with the same modulus. By (5.38b) these values are the eigenvalues of

X_1 of largest modulus. Hence X_1 has eigenvalues $\lambda_i^{(1)}$ satisfying $|\lambda_n^{(1)}| \le \cdots \le |\lambda_2^{(1)}| = |\lambda_1^{(1)}|$. Each subsequent iteration increases by at least 1 the number of eigenvalues with maximal modulus until, after $d-1$ iterations, X_{d-1} has eigenvalues of constant modulus. Then $\mu_{d-1}X_{d-1}$ has converged eigenvalues ± 1 (as does X_d). By Theorem 5.10, at most a further p iterations after X_{d-1} are needed to dispose of the Jordan blocks (and during these iterations $\mu_k \equiv 1$, since the eigenvalues are fixed at ± 1). □

For 1×1 matrices spectral scaling and determinantal scaling are equivalent, and both give convergence in at most two iterations (see Problem 5.14). For 2×2 matrices spectral scaling and determinantal scaling are again equivalent, and Theorem 5.11 tells us that we have convergence in at most two iterations if the eigenvalues are real. However, slightly more is true: both scalings give convergence in at most two iterations for any *real* 2×2 matrix (see Problem 5.14).

Determinantal scaling can be ineffective when there is a small group of outlying eigenvalues and the rest are nearly converged. Suppose that A has an eigenvalue 10^q ($q \ge 1$) with the rest all ± 1. Then determinantal scaling gives $\mu_k = 10^{-q/n}$, whereas spectral scaling gives $\mu_k = 10^{-q/2}$; the former quantity is close to 1 and hence the determinantally scaled iteration will behave like the unscaled iteration. Spectral scaling can be ineffective when the eigenvalues of A cluster close to the imaginary axis (see the numerical examples in Section 5.8).

All three scaling schemes are inexpensive to implement. The determinant $\det(X_k)$ can be computed at negligible cost from the LU factorization that will be used to compute X_k^{-1}. The spectral scaling parameter can be cheaply estimated by applying the power method to X_k and its inverse, again exploiting the LU factorization in the latter case. Note, however, that for a real spectrum spectral scaling increases the number of eigenvalues with maximal modulus on each iteration, which makes reliable implementation of the power method more difficult. The norm scaling is trivial to compute for the Frobenius norm, and for the 2-norm can be estimated using the power method (Algorithm 3.19).

The motivation for scaling is to reduce the length of the initial phase during which the error is reduced below 1. Should we continue to scale throughout the whole iteration? All three scaling parameters (5.35)–(5.37) converge to 1 as $X_k \to S$, so scaling does not destroy the quadratic convergence. Nor does it bring any benefit, so it is sensible to set $\mu_k \equiv 1$ once the error is sufficiently less than 1.

5.6. Terminating the Iterations

Crucial to the success of any sign iteration is an inexpensive and effective way to decide when to terminate it. We begin with a lemma that provides some bounds that help guide the choice of stopping criterion in both relative error-based and residual-based tests.

Lemma 5.12 (Kenney, Laub, Pandey, and Papadopoulos). *Let $A \in \mathbb{C}^{n\times n}$ have no pure imaginary eigenvalues, let $S = \operatorname{sign}(A)$, and let $\|\cdot\|$ be any subordinate matrix norm. If $\|S(A-S)\| = \epsilon < 1$ then*

$$\left(\frac{1-\epsilon}{2+\epsilon}\right)\|A-A^{-1}\| \le \|A-S\| \le \left(\frac{1+\epsilon}{2-\epsilon}\right)\|A-A^{-1}\| \qquad (5.39)$$

and

$$\frac{\|A^2-I\|}{\|S\|(\|A\|+\|S\|)} \le \frac{\|A-S\|}{\|S\|} \le \|A^2-I\|. \tag{5.40}$$

The lower bound in (5.40) *always holds.*

Proof. Let $E = A - S$. Since $S^2 = I$, we have $A = S + E = (I + ES)S$. It is then straightforward to show that

$$E(2I + ES) = (A - A^{-1})(I + ES),$$

using the fact that A and S, and hence also E and S, commute. The upper bound in (5.39) is obtained by postmultiplying by $(2I + ES)^{-1}$ and taking norms, while postmultiplying by $(I + ES)^{-1}$ and taking norms gives the lower bound.

The lower bound in (5.40) is obtained by taking norms in $A^2-I = (A-S)(A+S)$. For the upper bound, we write the last equation as $A - S = (A^2 - I)(A + S)^{-1}$ and need to bound $\|(A + S)^{-1}\|$. Since $A + S = 2S(I + \frac{1}{2}S(A - S))$, we have

$$\|(A+S)^{-1}\| = \frac{1}{2}\|S^{-1}\big(I+\tfrac{1}{2}S(A-S)\big)^{-1}\| \le \frac{\frac{1}{2}\|S^{-1}\|}{1-\frac{1}{2}\epsilon} \le \|S\|. \qquad \square$$

Note that since the iterations of interest satisfy $\mathrm{sign}(X_k) = \mathrm{sign}(A)$, the bounds of Lemma 5.12 are applicable with A replaced by an iterate X_k.

We now describe some possible convergence criteria, using η to denote a convergence tolerance proportional to both the unit roundoff (or a larger value if full accuracy is not required) and a constant depending on the matrix dimension, n. A norm will denote any easily computable norm such as the 1-, ∞-, or Frobenius norms. We begin with the Newton iteration, describing a variety of existing criteria followed by a new one.

A natural stopping criterion, of negligible cost, is

$$\delta_{k+1} := \frac{\|X_{k+1}-X_k\|}{\|X_{k+1}\|} \le \eta. \tag{5.41}$$

As discussed in Section 4.9, this criterion is really bounding the error in X_k, rather than X_{k+1}, so it may stop one iteration too late. This drawback can be seen very clearly from (5.39): since $X_{k+1}-X_k = \frac{1}{2}(X_k^{-1}-X_k)$, (5.39) shows that $\|X_{k+1}-X_k\| \approx \|S - X_k\|$ is an increasingly good approximation as the iteration converges.

The test (5.41) could potentially never be satisfied in floating point arithmetic. The best bound for the error in the computed $\widehat{Z}_k = fl(X_k^{-1})$, which we assume to be obtained by Gaussian elimination with partial pivoting, is of the form [276, 2002, Sec. 14.3]

$$\frac{\|\widehat{Z}_k - X_k^{-1}\|}{\|X_k^{-1}\|} \le c_n u \kappa(X_k), \tag{5.42}$$

where c_n is a constant. Therefore for the computed sequence X_k, $\|X_{k+1} - X_k\| \approx \frac{1}{2}\|\widehat{Z}_k - X_k\|$ might be expected to be proportional to $\kappa(X_k)\|X_k\|u$, suggesting the test $\delta_{k+1} \le \kappa(X_k)\eta$. Close to convergence, $X_{k+1} \approx X_k \approx S = S^{-1}$ and so $\kappa(X_k) \approx \|X_{k+1}\|^2$. A test $\delta_{k+1} \le \|X_{k+1}\|^2\eta$ is also suggested by the asymptotic conditioning of the sign function, discussed at the end of Section 5.1. On the other hand, a test of the form $\delta_{k+1} \le \|X_{k+1}\|\eta$ is suggested by Byers, He, and Mehrmann [89, 1997], based on a perturbation bound for the sign function. To summarize, there are arguments for using the stopping criterion

$$\delta_{k+1} \le \|X_{k+1}\|^p \, \eta \tag{5.43}$$

for each of $p = 0$, 1, and 2.

A different approach is based on the bound (5.17): $\|X_{k+1} - S\| \le \frac{1}{2}\|X_k^{-1}\| \, \|X_k - S\|^2$. Since $\|X_{k+1} - X_k\| \approx \|S - X_k\|$ close to convergence, as noted above,

$$\|X_{k+1} - S\| \lesssim \frac{1}{2}\|X_k^{-1}\| \, \|X_{k+1} - X_k\|^2.$$

Hence we can expect $\|X_{k+1} - S\|/\|X_{k+1}\| \lesssim \eta$ if

$$\|X_{k+1} - X_k\| \le \left(2\eta \frac{\|X_{k+1}\|}{\|X_k^{-1}\|}\right)^{1/2}. \tag{5.44}$$

This is essentially the same test as (4.25), bearing in mind that in the latter bound $c = \|S^{-1}\|/2 \approx \|X_k^{-1}\|/2$. This bound should overcome the problem of (5.41) of stopping one iteration too late, but unlike (5.43) with $p = 1, 2$ it takes no explicit account of rounding error effects. A test of this form has been suggested by Benner and Quintana-Ortí [55, 1999]. The experiments in Section 5.8 give further insight.

For general sign iterations, intuitively appealing stopping criteria can be devised based on the fact that trace(sign(A)) is an integer, but these are of little practical use; see Problem 5.16.

The upper bound in (5.40) shows that $\|A - X_k\|/\|X_k\| \le \|X_k^2 - I\|$ and hence suggests stopping when

$$\|X_k^2 - I\| \le \eta. \tag{5.45}$$

This test is suitable for iterations that already form X_k^2, such as the Schulz iteration (5.22). Note, however, that the error in forming $fl(X_k^2 - I)$ is bounded at best by $c_n u\|X_k\|^2 \approx c_n u\|S\|^2$, so when $\|S\|$ is large it may not be possible to satisfy (5.45), and a more suitable test is then

$$\frac{\|X_k^2 - I\|}{\|X_k\|^2} \le \eta.$$

5.7. Numerical Stability of Sign Iterations

The question of the stability of sign iterations, where stability is defined in Definition 4.17, has a particularly nice answer for all the iterations of interest.

Theorem 5.13 (stability of sign iterations). *Let $S = \operatorname{sign}(A)$, where $A \in \mathbb{C}^{n\times n}$ has no pure imaginary eigenvalues. Let $X_{k+1} = g(X_k)$ be superlinearly convergent to $\operatorname{sign}(X_0)$ for all X_0 sufficiently close to S and assume that g is independent of X_0. Then the iteration is stable, and the Fréchet derivative of g at S is idempotent and is given by $L_g(S, E) = L(S, E) = \frac{1}{2}(E - SES)$, where $L(S)$ is the Fréchet derivative of the matrix sign function at S.*

Proof. Since the sign function is idempotent, stability, the idempotence of L_g, and the equality of $L_g(S)$ and $L(S)$, follow from Theorems 4.18 and 4.19. The formula for $L(S, E)$ is obtained by taking $N = I$ in Theorem 5.3. □

Theorem 5.13 says that the Fréchet derivative at S is the same for any superlinearly convergent sign iteration and that this Fréchet derivative is idempotent. Unbounded propagation of errors near the solution is therefore not possible for any such iteration. The constancy of the Fréchet derivative is not shared by iterations for all the functions in this book, as we will see in the next chapter.

Turning to limiting accuracy (see Definition 4.20), Theorem 5.13 yields $\|L_g(S,E)\| \le \frac{1}{2}(1+\|S\|^2)\|E\|$, so an estimate for the limiting accuracy of any superlinearly convergent sign iteration is $\|S\|^2u$. Hence if, for example, $\kappa(S) = \|S\|^2 \le u^{-1/2}$, then we can hope to compute the sign function to half precision.

If S commutes with E then $L_g(S,E) = 0$, which shows that such errors E are eliminated by the iteration to first order. To compare with what convergence considerations say about E, note first that in all the sign iterations considered here the matrix whose sign is being computed appears only as the starting matrix and not within the iteration. Hence if we start the iteration at $S+E$ then the iteration converges to $\operatorname{sign}(S+E)$, for sufficiently small $\|E\|$ (so that the sign exists and any convergence conditions are satisfied). Given that S has the form (5.1), any E commuting with S has the form $Z\operatorname{diag}(F_{11},F_{22})Z^{-1}$, so that $\operatorname{sign}(S+E) = Z\operatorname{sign}(\operatorname{diag}(-I_p+F_{11}, I_q+F_{22}))Z^{-1}$. Hence there is an ϵ such that for all $\|E\| \le \epsilon$, $\operatorname{sign}(S+E) = S$. Therefore, the Fréchet derivative analysis is consistent with the convergence analysis.

Of course, to obtain a complete picture, we also need to understand the effect of rounding errors on the iteration prior to convergence. This effect is surprisingly difficult to analyze, even though the iterative methods are built purely from matrix multiplication and inversion. The underlying behaviour is, however, easy to describe. Suppose, as discussed above, that we have an iteration for $\operatorname{sign}(A)$ that does not contain A, except as the starting matrix. Errors on the $(k-1)$st iteration can be accounted for by perturbing X_k to X_k+E_k. If there are no further errors then (regarding X_k+E_k as a new starting matrix) $\operatorname{sign}(X_k+E_k)$ will be computed. The error thus depends on the conditioning of X_k and the size of E_k. Since errors will in general occur on each iteration, the overall error will be a complicated function of $\kappa_{\rm sign}(X_k)$ and E_k for all k.

We now restrict our attention to the Newton iteration (5.16). First, we note that the iteration can be numerically unstable: the relative error is not always bounded by a modest multiple of the condition number $\kappa_{\rm sign}(A)$, as is easily shown by example (see the next section). Nevertheless, it generally performs better than might be expected, given that it inverts possibly ill conditioned matrices. We are not aware of any published rounding error analysis for the computation of $\operatorname{sign}(A)$ via the Newton iteration.

Error analyses aimed at the application of the matrix sign function to invariant subspace computation (Section 2.5) are given by Bai and Demmel [29, 1998] and Byers, He, and Mehrmann [89, 1997]. These analyses show that the matrix sign function may be more ill conditioned than the problem of evaluating the invariant subspaces corresponding to eigenvalues in the left half-plane and right half-plane. Nevertheless, they show that when Newton's method is used to evaluate the sign function the computed invariant subspaces are usually about as good as those computed by the QR algorithm. In other words, the potential instability rarely manifests itself. The analyses are complicated and we refer the reader to the two papers for details.

In cases where the matrix sign function approach to computing an invariant subspace suffers from instability, iterative refinement can be used to improve the com-

Table 5.2. *Number of iterations for scaled Newton iteration. The unnamed matrices are (quasi)-upper triangular with normal* $(0,1)$ *distributed elements in the upper triangle.*

	Scaling			
Matrix	none	determinantal	spectral	norm
Lotkin	25	9	8	9
Grcar	11	9	9	15
$A(j\colon j+1, j\colon j+1) = \left[\begin{smallmatrix} 1 & (j/n)1000 \\ -(j/n)1000 & 1 \end{smallmatrix}\right]$	24	16	19	19
$a_{jj} = 1 + 1000i(j-1)/(n-1)$	24	16	22	22
$a_{11} = 1000,\ a_{jj} \equiv 1,\ j \geq 2$	14	12	6	10
$a_{11} = 1 + 1000i,\ a_{jj} \equiv 1,\ j \geq 2$	24	22	8	19

puted subspace [29, 1998]. Iterative refinement can also be used when the sign function is used to solve algebraic Riccati equations (as described in Section 2.4) [88, 1987].

Finally, we note that all existing numerical stability analysis is for the *unscaled* Newton iteration. Our experience is that scaling tends to improve stability, not worsen it.

5.8. Numerical Experiments and Algorithm

We present some numerical experiments to illustrate the theory of the previous three sections and to give further insight into the choice of iteration, acceleration scheme, and stopping criterion. In all the tests, scaling was used as long as the relative change $\delta_k = \|X_k - X_{k-1}\|_\infty / \|X_k\|_\infty$ exceeded 10^{-2}; thereafter $\mu_k \equiv 1$ and, where relevant, μ_k is not shown in the tables.

First, we consider the effects of scaling. For a variety of matrices we ran the Newton iteration (5.34) with no scaling and with the scalings (5.35)–(5.37), with the 2-norm used for norm scaling. We recorded how many iterations are required to produce an error $\|S - X_k\|_\infty / \|S\|_\infty \leq 5 \times 10^{-14}$. The matrices are as follows:

1. The 8×8 Lotkin matrix, MATLAB's `gallery('lotkin',8)`: badly conditioned with many negative eigenvalues of small magnitude.
2. The 25×25 Grcar matrix, `gallery('grcar',25)`: a Toeplitz matrix with sensitive eigenvalues.
3. 25×25 (quasi-) upper triangular matrices with elements in the upper triangle (outside the diagonal blocks) from the normal (0,1) distribution.

Table 5.2 reports the results. The Lotkin matrix is a typical example of how scaling can greatly reduce the number of iterations. The Grcar example shows how norm scaling can perform poorly (indeed being worse than no scaling). The third matrix (real) and fourth matrix (complex) have eigenvalues on a line with real part 1 and imaginary parts between 0 and 1000. Here, spectral scaling and norm scaling are both poor. The fifth and sixth matrices, again real and complex, respectively, have eigenvalues all equal to 1 except for one large outlier, and they are bad cases for determinantal scaling.

Table 5.3 illustrates the convergence results in Theorems 5.10 and 5.11 by showing the behaviour of the Newton iteration with spectral scaling for $J(2) \in \mathbb{R}^{16\times16}$, which

Table 5.3. *Newton iteration with spectral scaling for Jordan block $J(2) \in \mathbb{R}^{16\times 16}$.*

k	$\dfrac{\lVert S - X_k\rVert_\infty}{\lVert S\rVert_\infty}$	δ_k	$\dfrac{\lVert X_k^2 - I\rVert_\infty}{\lVert X_k\rVert_\infty^2}$	μ_k	(5.41)	(5.44)
1	2.5e-1	1.8e+0	3.6e-1	5.0e-1		
2	2.5e-2	2.2e-1	4.8e-2	1.0e0		
3	3.0e-4	2.5e-2	6.0e-4	1.0e0		
4	0	3.0e-4	0	1.0e0		
5	0	0	0		$\surd$	$\surd$

Table 5.4. *Newton iteration with determinantal scaling for random $A \in \mathbb{R}^{16\times 16}$ with $\kappa_2(A) = 10^{10}$; $\kappa_{\text{sign}}(A) = 3 \times 10^8$, $\lVert S\rVert_F = 16$.*

k	$\dfrac{\lVert S - X_k\rVert_\infty}{\lVert S\rVert_\infty}$	δ_k	$\dfrac{\lVert X_k^2 - I\rVert_\infty}{\lVert X_k\rVert_\infty^2}$	μ_k	(5.41)	(5.44)
1	4.3e3	1.0e0	1.1e-1	1.0e5		
2	1.5e1	2.8e2	1.3e-1	6.8e-3		
3	1.9e0	6.3e0	5.9e-2	1.4e-1		
4	2.1e-1	1.7e0	2.1e-2	6.1e-1		
5	6.4e-2	2.3e-1	4.3e-3	9.5e-1		
6	2.0e-3	6.2e-2	1.6e-4	9.8e-1		
7	4.1e-6	2.0e-3	3.3e-7	1.0e0		
8	2.1e-9	4.1e-6	8.9e-13			
9	2.1e-9	1.1e-11	3.2e-17			$\surd$
10	2.1e-9	1.5e-15	3.5e-17		$\surd$	$\surd$

is a Jordan block with eigenvalue 2. Here and below the last two columns of the table indicate with a tick iterations on which the convergence conditions (5.41) and (5.44) are satisfied for the ∞-norm, with $\eta = n^{1/2}u$. In Theorem 5.11, $d = 1$ and $p = 4$, and indeed $X_{d+p-1} = X_4 = \operatorname{sign}(J(2))$. At the start of the first iteration, $\mu_0 X_0$ has eigenvalues 1, and the remaining four iterations remove the nonnormal part; it is easy to see that determinantal scaling gives exactly the same results.

Table 5.4 reports 12 iterations for a random $A \in \mathbb{R}^{16\times 16}$ with $\kappa_2(A) = 10^{10}$ generated in MATLAB by `gallery('randsvd',16,1e10,3)`. Determinantal scaling was used. Note that the relative residual decreases significantly after the error has stagnated. The limiting accuracy of $\lVert S\rVert_2^2 u$ is clearly not relevant here, as the iterates do not approach S sufficiently closely.

Both these examples confirm that the relative change δ_{k+1} is a good estimate of the relative error in X_k (compare the numbers in the third column with those immediately to the northwest) until roundoff starts to dominate, but thereafter the relative error and relative change can behave quite differently.

Finally, Table 5.5 gives examples with large $\lVert S\rVert$. The matrix is of the form $A = QTQ^T$, where Q is a random orthogonal matrix and $T \in \mathbb{R}^{16\times 16}$ is generated as an upper triangular matrix with normal (0,1) distributed elements and t_{ii} is replaced by $d|t_{ii}|$ for $i = 1\colon 8$ and by $-d|t_{ii}|$ for $i = 9\colon 16$. As d is decreased the eigenvalues of A approach the origin (and hence the imaginary axis). Determinantal scaling was used and we terminated the iteration when the relative error stopped decreasing sig-

Table 5.5. *Newton iteration with determinantal scaling for random $A \in \mathbb{R}^{16\times 16}$ with real eigenvalues parametrized by d.*

d	no. iterations	$\min_k \dfrac{\|S - X_k\|_\infty}{\|S\|_\infty}$	$\|A\|_2$	$\kappa_2(A)$	$\|S\|_2$	$\kappa_{\mathrm{sign}}(A)$
1	6	2.7e-13	6.7	4.1e3	1.3e2	4.7e3
3/4	6	4.1e-10	6.5	5.7e5	5.4e3	6.5e5
1/2	6	2.6e-6	6.2	2.6e8	3.9e5	6.5e7
1/3	3	7.8e-1	6.4	2.6e15	7.5e11	3.9e7

nificantly. This example shows that the Newton iteration can behave in a numerically unstable way: the relative error can greatly exceed $\kappa_{\mathrm{sign}}(A)u$. Note that the limiting accuracy $\|S\|_2^2 u$ provides a good estimate of the relative error for the first three values of d.

Our experience indicates that (5.44) is the most reliable termination criterion, though on badly behaved matrices such as those in Table 5.5 no one test can be relied upon to terminate at the "right moment", if at all.

Based on this and other evidence we suggest the following algorithm based on the scaled Newton iteration (5.34).

Algorithm 5.14 (Newton algorithm for matrix sign function). Given a nonsingular $A \in \mathbb{C}^{n\times n}$ with no pure imaginary eigenvalues this algorithm computes $X = \mathrm{sign}(A)$ using the scaled Newton iteration. Two tolerances are used: a tolerance tol_cgce for testing convergence and a tolerance tol_scale for deciding when to switch to the unscaled iteration.

```
1  $X_0 = A$; scale = true
2  for $k = 1\colon\infty$
3      $Y_k = X_k^{-1}$
4      if scale
5          Set $\mu_k$ to one of the scale factors (5.35)–(5.37).
6      else
7          $\mu_k = 1$
8      end
9      $X_{k+1} = \frac{1}{2}(\mu_k X_k + \mu_k^{-1} Y_k)$
10     $\delta_{k+1} = \|X_{k+1} - X_k\|_F / \|X_{k+1}\|_F$
11     if scale = true and $\delta_{k+1} \le$ tol_scale, scale = false, end
12     if $\|X_{k+1} - X_k\|_F \le (\text{tol\_cgce}\|X_{k+1}\|/\|Y_k\|)^{1/2}$ or
       ($\delta_{k+1} > \delta_k/2$ and scale = false)
13         goto line 16
14     end
15 end
16 $X = X_{k+1}$
```

Cost: $2kn^3$ flops, where k iterations are used.

The algorithm uses the unscaled Newton iteration once the relative change in the iterates is less than tol_scale. A value of tol_scale safely less than 1 is intended and the motivation is to avoid the (nonoptimal) scaling parameters interfering with the quadratic convergence once the convergence has set in. The convergence test is (5.44) combined with the requirement to stop if, in the final convergence phase, δ_k has not decreased by at least a factor 2 during the previous iteration (which is a sign that roundoff errors are starting to dominate).

We have left the choice of scale factor at line 5 open, as the best choice will depend on the class of problems considered.

5.9. Best L_∞ Approximation

Most applications of the matrix sign function involve nonnormal matrices of small to medium size. An exception is the application in lattice quantum chromodynamics (QCD) described in Section 2.7, where the action on a vector of the sign of a large, sparse, Hermitian matrix is required. For Hermitian A, approximating $\operatorname{sign}(A)$ reduces to approximating $\operatorname{sign}(x)$ at the eigenvalues of A, which is a scalar problem on the real axis. The full range of scalar approximation techniques and results can therefore be brought into play. In particular, we can use best L_∞ rational approximations. For the sign function and the interval $[-\delta_{\max}, -\delta_{\min}] \cup [\delta_{\min}, \delta_{\max}]$ an explicit formula for the best L_∞ approximation is known. It follows from a corresponding result for the inverse square root. The result is phrased in terms of elliptic functions. The Jacobi elliptic function $\operatorname{sn}(w;\kappa) = x$ is defined implicitly by the elliptic integral

$$w = \int_0^x \frac{1}{\sqrt{(1-t^2)(1-\kappa^2t^2)}}\,dt$$

and the complete elliptic integral (for the modulus κ) is defined by

$$K = \int_0^1 \frac{1}{\sqrt{(1-t^2)(1-\kappa^2t^2)}}\,dt.$$

Theorem 5.15 (Zolotarev, 1877).

(a) *The best L_∞ approximation $\widetilde{r}$ from $\mathcal{R}_{m-1,m}$ to $x^{-1/2}$ on the interval* $[1, (\delta_{\max}/\delta_{\min})^2]$ *is*

$$\widetilde{r}(x) = D\frac{\prod_{j=1}^{m-1}(x+c_{2j})}{\prod_{j=1}^{m}(x+c_{2j-1})},$$

where

$$c_j = \frac{\operatorname{sn}^2(jK/(2m);\kappa)}{1-\operatorname{sn}^2(jK/(2m);\kappa)},$$

$\kappa = (1-(\delta_{\min}/\delta_{\max})^2)^{1/2}$, and K is the complete elliptic integral for the modulus κ. The constant D is determined by the condition

$$\max_{x\in[1,(\delta_{\min}/\delta_{\max})^2)]}(1-\sqrt{x}\,\widetilde{r}(x)) = -\min_{x\in[1,(\delta_{\min}/\delta_{\max})^2)]}(1-\sqrt{x}\,\widetilde{r}(x)),$$

and the extrema occur at $x_j = \operatorname{dn}^{-2}(jK/(2m))$, $j = 0\colon 2m$, where $\operatorname{dn}^2(w;\kappa) = 1 - \kappa^2\operatorname{sn}^2(w;\kappa)$.

(b) *The best L_∞ approximation r from $\mathcal{R}_{2m-1,2m}$ to* $\operatorname{sign}(x)$ *on the interval* $[-\delta_{\max}, -\delta_{\min}] \cup [\delta_{\min}, \delta_{\max}]$ *is $r(x) = (x/\delta_{\min})\widetilde{r}((x/\delta_{\min})^2)$, where $\widetilde{r}$ is defined in* (a). □

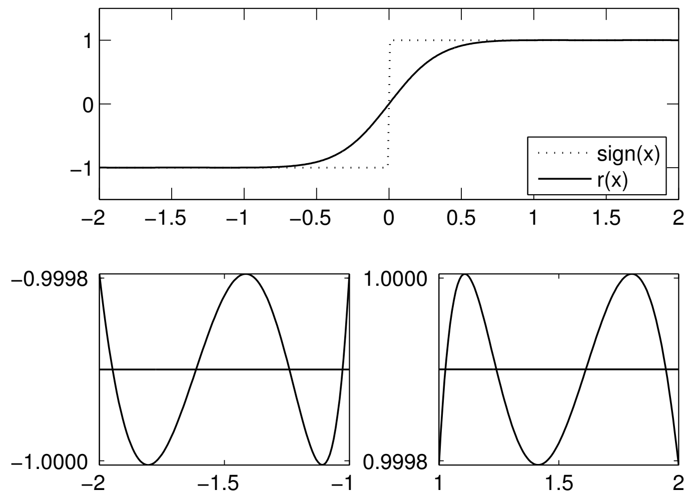

Figure 5.2. *Best L_∞ approximation $r(x)$ to* $\mathrm{sign}(x)$ *from $\mathcal{R}_{3,4}$ on* $[-2,-1]\cup[1,2]$. *The lower two plots show $r(x)$ in particular regions of the overall plot above.*

Figure 5.2 plots the best L_∞ approximation to $\mathrm{sign}(x)$ from $\mathcal{R}_{3,4}$ on $[-2,-1]\cup[1,2]$, and displays the characteristic equioscillation property of the error, which has maximum magnitude about 10^{-4}. In the QCD application $\delta_{\min}$ and $\delta_{\max}$ are chosen so that the spectrum of the matrix is enclosed and r is used in partial fraction form.

5.10. Notes and References

The matrix sign function was introduced by Roberts [496] in 1971 as a tool for model reduction and for solving Lyapunov and algebraic Riccati equations. He defined the sign function as a Cauchy integral and obtained the integral (5.3). Roberts also proposed the Newton iteration (5.16) for computing $\mathrm{sign}(A)$ and proposed scaling the iteration, though his scale parameters are not as effective as the ones described here.

Interest in the sign function grew steadily in the 1970s and 1980s, initially among engineers and later among numerical analysts. Kenney and Laub give a thorough survey of the matrix sign function and its history in [347, 1995].

The attractions of the concise representation $\mathrm{sign}(A) = A(A^2)^{-1/2}$ in (5.2) were pointed out by Higham [273, 1994], though the formula can be found in earlier work of Tsai, Shieh, and Yates [576, 1988].

Theorem 5.2 is from Higham, Mackey, Mackey, and Tisseur [283, 2005].

Theorems 5.3 and 5.7 are due to Kenney and Laub [342, 1991]. The expression (5.10) and upper bound (5.11) for the matrix sign function condition number are from Higham [273, 1994]. Theorem 5.4 is a refined version of a result of Kenney and Laub [342, 1991]. Another source of perturbation results for the matrix sign function is Sun [548, 1997].

The Schur method, Algorithm 5.5, is implemented in function `signm` of the Matrix Computation Toolbox [264] (on which `signm` in the Matrix Function Toolbox is based) but appears here in print for the first time.

For more on the recursions related to (5.26), and related references, see Chapter 8.

It is natural to ask how sharp the sufficient condition for convergence $\|I-A^2\| < 1$ in Theorem 5.8 (a) is for $\ell > m$ and what can be said about convergence for $\ell < m-1$. These questions are answered experimentally by Kenney and Laub [343, 1991], who give plots showing the boundaries of the regions of convergence of the scalar iterations in $\mathbb{C}$.

The principal Padé iterations for the sign function were first derived by Howland [302, 1983], though for even k his iteration functions are the inverses of those given here. Iannazzo [307, 2007] points out that these iterations can be obtained from the general König family (which goes back to Schröder [509, 1870], [510, 1992]) applied to the equation $x^2-1=0$. Parts (b)–(d) of Theorem 5.9 are from Kenney and Laub [345, 1994]. Pandey, Kenney, and Laub originally obtained the partial fraction expansion (5.30), for even k only, by applying Gaussian quadrature to an integral expression for $h(\xi)$ in (5.26) [457, 1990]. The analysis leading to (5.31) is from Kenney and Laub [345, 1994].

Theorem 5.10 is due to Kenney and Laub [344, 1992], and the triangular matrices in Table 5.2 are taken from the same paper.

Theorem 5.11 is due to Barraud [44, 1979, Sec. 4], but, perhaps because his paper is written in French, his result went unnoticed until it was presented by Kenney and Laub [344, 1992, Thm. 3.4].

Lemma 5.12 collects results from Kenney, Laub, and Papadopoulos [350, 1993] and Pandey, Kenney, and Laub [457, 1990].

The spectral scaling (5.36) and norm scaling (5.37) were first suggested by Barraud [44, 1979], while determinantal scaling (5.35) is due to Byers [88, 1987].

Kenney and Laub [344, 1992] derive a "semioptimal" scaling for the Newton iteration that requires estimates of the dominant eigenvalue (not just its modulus, i.e., the spectral radius) of X_k and of X_k^{-1}. Numerical experiments show this scaling to be generally at least as good as the other scalings we have described. Semioptimal scaling does not seem to have become popular, probably because it is more delicate to implement than the other scalings and the other scalings typically perform about as well in practice.

Theorem 5.13 on the stability of sign iterations is new. Indeed we are not aware of any previous analysis of the stability of sign iterations.

Our presentation of Zolotarev's Theorem 5.15 is based on that in van den Eshof, Frommer, Lippert, Schilling, and Van der Vorst [585, 2002] and van den Eshof [586, 2003]. In the numerical analysis literature this result seems to have been first pointed out by Kenney and Laub [347, 1995, Sec. III]. Theorem 5.15 can also be found in Achieser [1, 1956, Sec. E.27], Kennedy [338, 2004], [339, 2005], and Petrushev and Popov [470, 1987, Sec. 4.3].

A "generalized Newton sign iteration" proposed by Gardiner and Laub [205, 1986] has the form

$$X_{k+1} = \frac{1}{2}(X_k + BX_k^{-1}B), \qquad X_0 = A.$$

If B is nonsingular this is essentially the standard Newton iteration applied to $B^{-1}A$ and it converges to $B\,\mathrm{sign}(B^{-1}A)$. For singular B, convergence may or may not occur and can be at a linear rate; see Bai, Demmel, and Gu [31, 1997] and Sun and Quintana-Ortí [550, 2002]. This iteration is useful for computing invariant subspaces of matrix pencils $A - \lambda B$ (generalizing the approach in Section 2.5) and for solving generalized algebraic Riccati equations.

Problems

5.1. Show that $\operatorname{sign}(A) = A$ for any involutory matrix.

5.2. How are $\operatorname{sign}(A)$ and $\operatorname{sign}(A^{-1})$ related?

5.3. Derive the integral formula (5.3) from (5.2) by using the Cauchy integral formula (1.12).

5.4. Show that $\operatorname{sign}(A) = (2/\pi)\lim_{t\to\infty}\tan^{-1}(tA)$.

5.5. Can

$$A = \begin{bmatrix} -1 & 1 & 1/2 \\ 0 & 1 & -1 \\ 0 & 0 & 1 \end{bmatrix}$$

be the sign of some matrix?

5.6. Show that the geometric mean $A \# B$ of two Hermitian positive definite matrices A and B satisfies

$$\begin{bmatrix} 0 & A \# B \\ (A \# B)^{-1} & 0 \end{bmatrix} = \operatorname{sign}\left(\begin{bmatrix} 0 & B \\ A^{-1} & 0 \end{bmatrix}\right).$$

5.7. (Kenney and Laub [342, 1991]) Verify that for $A \in \mathbb{R}^{2\times 2}$ the matrix sign decomposition (5.5) is given as follows. If $\det(A) > 0$ and $\operatorname{trace}(A) \neq 0$ then $S = \operatorname{sign}(\operatorname{trace}(A))I$ and $N = \operatorname{sign}(\operatorname{trace}(A))A$; if $\det(A) < 0$ then

$$S = \mu\bigl(A - \det(A)A^{-1}\bigr), \qquad N = \mu\bigl(A^2 - \det(A)I\bigr),$$

where

$$\mu = \bigl(-\det(A - \det(A)A^{-1})\bigr)^{-1/2};$$

otherwise S is undefined.

5.8. Show that the Newton iteration (5.16) for the matrix sign function can be derived by applying Newton's method to the equation $X^2 = I$.

5.9. By expanding the expression $\operatorname{sign}(S + E) = (S + E)((S + E)^2)^{-1/2}$ from (5.2), show directly that the Fréchet derivative of the matrix sign function at $S = \operatorname{sign}(S)$ is given by $L(S, E) = \frac{1}{2}(E - SES)$.

5.10. Consider the scalar Newton sign iteration $x_{k+1} = \frac{1}{2}(x_k + x_k^{-1})$. Show that if $x_0 = \coth\theta_0$ then $x_k = \coth 2^k\theta_0$. Deduce a convergence result.

5.11. (Schroeder [511, 1991]) Investigate the behaviour of the Newton iteration (5.16) for scalar, pure imaginary x_0. Hint: let $x_0 = ir_0 \equiv -i\cot(\pi\theta_0)$ and work in θ coordinates.

5.12. Halley's iteration for solving $f(x) = 0$ is [201, 1985]

$$x_{k+1} = x_k - \frac{f_k/f_k'}{1 - \frac{1}{2}f_k f_k''/(f_k')^2},$$

where f_k, f_k', and f_k'' denote the values of f and its first two derivatives at x_k. Show that applying Halley's iteration to $f(x) = x^2 - 1$ yields the iteration function $f_{1,1}$ in Table 5.1.

5.13. (Byers [88, 1987]) Show that determinantal scaling $\mu = |\det(X)|^{-1/n}$ minimizes $d(\mu X)$, where

$$d(X) = \sum_{i=1}^{n} (\log |\lambda_i|)^2$$

and the λ_i are the eigenvalues of X. Show also that $d(X) = 0$ if and only if the spectrum of X lies on the unit circle and that $d(X)$ is an increasing function of $|1 - |\lambda_i||$ for each eigenvalue λ_i.

5.14. Consider the Newton iteration (5.34), with determinantal scaling (5.35) and spectral scaling (5.36). Show that with both scalings the iteration converges in at most two iterations (a) for scalars and (b) for any real 2×2 matrix.

5.15. (Higham, Mackey, Mackey, and Tisseur [283, 2005]) Suppose that $\mathrm{sign}(A) = I$ and $A^2 = I + E$, where $\|E\| < 1$, for some consistent norm. Show that

$$\|A - I\| \le \frac{\|E\|}{1 + \sqrt{1 - \|E\|}} < \|E\|.$$

How does this bound compare with the upper bound in (5.40)?

5.16. Discuss the pros and cons of terminating an iteration $X_{k+1} = g(X_k)$ for the matrix sign function with one of the tests

$$|\operatorname{trace}(X_k^2) - n| \le \eta, \tag{5.46}$$

$$|\operatorname{trace}(X_k) - \operatorname{round}(\operatorname{trace}(X_k))| \le \eta, \tag{5.47}$$

where $\mathrm{round}(x)$ denotes the nearest integer to x.

5.17. (Byers [88, 1987]) The matrix

$$W = \begin{bmatrix} A^* & G \\ F & -A \end{bmatrix}, \qquad F = F^*, \quad G = G^*,$$

arising in (2.14) in connection with the Riccati equation is Hamiltonian, that is, it satisfies the condition that JW is Hermitian, where $J = \left[\begin{smallmatrix} 0 & I_n \\ -I_n & 0 \end{smallmatrix}\right]$. Show that the Newton iteration for $\mathrm{sign}(W)$ can be written in such a way that only Hermitian matrices need to be inverted. The significance of this fact is that standard algorithms or software for Hermitian matrices can then be used, which halves the storage and computational costs compared with treating W as a general matrix.

> *The sign function of a square matrix can be defined in terms of a contour integral or as the result of an iterated map $Z_{r+1} = \frac{1}{2}(Z_r + Z_r^{-1})$. Application of this function enables a matrix to be decomposed into two components whose spectra lie on opposite sides of the imaginary axis.*
>
> — J. D. ROBERTS, *Linear Model Reduction and Solution of the Algebraic Riccati Equation by Use of the Sign Function* (1980)

> *The matrix sign function method is an elegant and, when combined with defect correction, effective numerical method for the algebraic Riccati equation.*
>
> — VOLKER MEHRMANN, *The Autonomous Linear Quadratic Control Problem: Theory and Numerical Solution* (1991)