Electronic Transactions on Numerical Analysis.
Volume 39, pp. 414-436, 2012.

ISSN 1068-9613.

ETNA
Kent State University
http://etna.math.kent.edu

COMPUTATION OF THE MATRIX PTH ROOT AND ITS FRÉCHET DERIVATIVE BY INTEGRALS*

JOÃO R. CARDOSO†

Abstract. We present new integral representations for the matrix pth root and its Fréchet derivative and then investigate the computation of these functions by numerical quadrature. Three different quadrature rules are considered: composite trapezoidal, Gauss-Legendre and adaptive Simpson. The problem of computing the matrix pth root times a vector without the explicit evaluation of the pth root is also analyzed and bounds for the norm of the matrix pth root and its Fréchet derivative are derived.

Key words. matrix pth root, Fréchet derivative, quadrature, composite trapezoidal rule, Gauss-Legendre rule, adaptive Simpson rule

AMS subject classifications. 65F60, 65D30

1. Introduction. Let $p \geq 2$ be a positive integer. Given a matrix $A \in \mathbb{R}^{n\times n}$ with eigenvalues not belonging to the closed negative real axis, there exists a unique real matrix X such that $X^p = A$, whose eigenvalues lie on the sector of the complex plane defined by $-\pi/p < \arg(z) < \pi/p$, where $\arg(z)$ denotes the argument of the complex number z. This unique matrix X is called the *principal* pth root of A and will be denoted by $A^{1/p}$. For background on matrix pth roots and general matrix functions, see [20].

The sensitivity of a matrix function to small perturbations in the data (at first order) is measured by a condition number based on the norm of the Fréchet derivative. Let $A, E \in \mathbb{R}^{n\times n}$. The Fréchet derivative of a matrix function f at A in the direction of E is a linear operator $L_f(A)$ that maps each "direction matrix" E to $L_f(A,E)$ such that

$$f(A+E) - f(A) - L_f(A,E) = o(\|E\|).$$

The Fréchet derivative of the matrix function f may not exist at A, but if it does it is unique and $L_f(A,E)$ coincides with the directional (or Gâteaux) derivative of f at A in the direction E. Any consistent matrix norm $\|.\|$ on $\mathbb{R}^{n\times n}$ induces the operator norm

$$\|L_f(A)\| := \max_{\|E\|=1} \|L_f(A,E)\|,$$

which allows one to define the condition number of f at A

$$\kappa_f(A) := \frac{\|L_f(A)\|\,\|A\|}{\|f(A)\|}\,.$$

Here one uses the same notation to denote both the matrix norm and the induced operator norm. Once an approximation to the matrix $L_f(A,E)$ is known, some algorithms are available to estimate $\|L_f(A)\|$ and the condition number $\kappa_f(A)$. A well known example is the power method on the Fréchet derivative. For more details about Fréchet derivatives of matrix functions, see [20, ch. 3] and the references therein.

One of the goals of this work is to investigate numerical quadrature for the computation of the Fréchet derivative of the matrix pth root, $L_{x^{1/p}}(A,E)$. Given $A, E \in \mathbb{R}^{n\times n}$, with A

*Received January 31, 2012. Accepted for publication September 18, 2012. Published online on November 19, 2012. Recommended by A. Frommer.
†Coimbra Institute of Engineering, Rua Pedro Nunes, 3030-199 Coimbra – Portugal, and Institute of Systems and Robotics, University of Coimbra, Pólo II, 3030-290 Coimbra, Portugal (jocar@isec.pt).

ETNA
Kent State University
http://etna.math.kent.edu

having no eigenvalue on the closed negative real axis, the Fréchet derivative of the matrix pth root exists and $L_{x^{1/p}}(A, E)$ is the unique solution of the generalized Sylvester equation

$$\sum_{j=0}^{p-1} \left(A^{1/p}\right)^{p-1-j} X \left(A^{1/p}\right)^{j} = E$$

(see [20, Theorem 3.8] and [25, Sec. 2.5]). It can be proved that its Frobenius norm can be written as

$$\|L_{x^{1/p}}(A)\|_F = \left\| \left(\sum_{j=0}^{p-1} \left[\left(A^{1/p}\right)^T \right]^j \otimes \left(A^{1/p}\right)^{p-1-j} \right)^{-1} \right\|_2$$

(see Problem 7.4 and its solution in [20]). Here $\|.\|_F$ and $\|.\|_2$ denote, respectively, the Frobenius and the 2-norm, and $\otimes$ denotes the Kronecker product. An effective method for computing the Fréchet derivative of the matrix pth root was recently proposed in [6].

A well-known integral representation of the principal matrix pth root of A is given by ([17], [20, p. 174])

$$A^{1/p} = \frac{p \sin(\pi/p)}{\pi} A \int_0^{\infty} (t^p I + A)^{-1}\, dt. \tag{1.1}$$

This integral representation over the non negative real line will be the basis of our work. From (1.1) we derive new integral representations for both the matrix pth root and its Fréchet derivative and then investigate quadrature for these integrals. Three different types of quadrature are considered: composite trapezoidal, Gauss-Legendre and adaptive Simpson. Our numerical experiments show that while Gauss-Legendre quadrature is a good choice for computing the matrix pth root $A^{1/p}$, it is the composite trapezoidal rule that presents the best performance for the Fréchet derivative.

Another topic of investigation in this paper is the computation of $A^{1/p}b$, where b is a vector. Whereas the methods for computing $A^{1/p}$ involve in general $O(n^3)$ arithmetic operations, the computation of $T^{1/p}b$ by quadrature, with T triangular, quasi-triangular or upper Hessenberg, can be performed with just $O(n^2)$ operations. Triangular and Hessenberg forms of A can be found by the Schur and Hessenberg decompositions, respectively. As observed previously in [9, 16], this reinforces the role of integral representations and quadrature in the matrix functions computation problem. We recall that integral representations for matrix functions have been known for a long time, but just recently they have been used for practical computations.

It is worth noting that many technical problems arising in areas such as control, geography, finance and healthcare involve the computation of the matrix pth root. For papers including applications see the references in [21]. We add three recent papers containing applications: [7], [24] and [26].

This paper is organized as follows. In Section 2 we show that the integral in (1.1) can be written as a sum of an integral over a finite range plus an integral over an infinite range with norm bounded by a constant. We also show that new integral representations for both the matrix pth root and its Fréchet derivative over a finite range can be derived from (1.1), by choosing an appropriate change of variable. Since all the integral representations involve a resolvent function, some estimates for the norm of this function are revisited in Section 3 and used to obtain bounds for the norm of the matrix pth root and its Fréchet derivative. The difficulty of finding sharp and practical error estimates is discussed. Composite trapezoidal, Gauss-Legendre and adaptive Simpson rules are implemented in Sections 5, 6, and 7 for the evaluation of the matrix pth root, the matrix pth root times a vector and the Fréchet derivative, respectively. Numerical experiments are carried out to illustrate and understand the behavior of these three types of quadrature. Finally, in Section 8, we draw some conclusions.

Notation: Unless otherwise stated, throughout the text $p \geq 2$ will denote a positive integer, A a real matrix with no eigenvalue on the closed negative real axis and $\|.\|$ a consistent matrix or operator norm; $\|.\|_F$ and $\|.\|_2$ stand for the Frobenius norm and the 2-norm, respectively.

2. Integral representations. We start by proving that the integral in (1.1) can be split into two integrals, where the norm of the second one, over an infinite range, can be bounded by a constant. This result provides an important contribution to understand the behavior of the matrix pth root integral representation (1.1).

THEOREM 2.1. *Let A have no eigenvalues on the closed negative real axis. If $r \geq (2\|A\|)^{1/p}$, then*

$$A^{1/p} = \frac{p\,\sin(\pi/p)}{\pi} A \left(\int_0^r (t^p I + A)^{-1}\, dt + \int_r^\infty (t^p I + A)^{-1}\, dt \right), \tag{2.1}$$

where

$$\left\| \int_r^\infty (t^p I + A)^{-1}\, dt \right\| \leq \frac{2\, r^{1-p}}{p-1}\,.$$

Proof. Assume that $t \in [r, \infty[$, with $r \geq (2\|A\|)^{1/p}$. This ensures that $\|A/t^p\| < 1$ and we can write

$$(t^p I + A)^{-1} = (t^p)^{-1} \left(I + \frac{A}{t^p} \right)^{-1} = \frac{1}{t^p} \sum_{k=0}^{\infty} (-1)^k \left(\frac{A}{t^p} \right)^k.$$

Integrating over the range $[r, \infty[$, some calculation shows that

$$\int_r^\infty (t^p I + A)^{-1}\, dt = r^{1-p} \sum_{k=0}^{\infty} \frac{(-1)^k}{p(k+1)-1} \left(\frac{A}{r^p} \right)^k.$$

The condition $r \geq (2\|A\|)^{1/p}$ means that $\|A/r^p\| \leq 1/2$. Since $p \geq 2$

$$\begin{aligned} \left\| \sum_{k=0}^{\infty} \frac{(-1)^k}{p(k+1)-1} \left(\frac{A}{r^p} \right)^k \right\| &\leq \sum_{k=0}^{\infty} \frac{1}{p(k+1)-1} \left\| \frac{A}{r^p} \right\|^k \\ &\leq \frac{1}{p-1} \sum_{k=0}^{\infty} \left(\frac{1}{2} \right)^k \\ &\leq \frac{2}{p-1}, \end{aligned}$$

and then the result follows. □

Theorem 2.1 gives a bound for the truncation error arising when we replace the infinite interval in (1.1) by a finite one. To illustrate this, consider for instance a matrix A with norm $\|A\| = 10^2$ and $p = 7$. For $r = 10 > (2\|A\|)^{1/7}$,

$$\left\| A^{1/7} - \frac{7\,\sin(\pi/7)}{\pi} A \int_0^{10} (t^7 I + A)^{-1}\, dt \right\| \leq 3.2 \times 10^{-5}.$$

Increasing r to 50, one has

$$\left\| A^{1/7} - \frac{7\,\sin(\pi/7)}{\pi} A \int_0^{50} (t^7 I + A)^{-1}\, dt \right\| \le 2.1 \times 10^{-9}.$$

This is of particular interest if one wants to approximate (1.1) by quadrature. Assume that

$$X = A^{1/p} = \frac{p\,\sin(\pi/p)}{\pi} A(X_1 + X_2),$$

where

$$X_1 = \int_0^r (t^p I + A)^{-1}\, dt \quad \text{and} \quad X_2 = \int_r^\infty (t^p I + A)^{-1}\, dt,$$

and a quadrature rule is applied to X_1 yielding the approximation $\tilde{X}_1$ with an absolute error $\|X_1 - \tilde{X}_1\| \le \epsilon$. If $\tilde{X} = \frac{p\,\sin(\pi/p)}{\pi} A\tilde{X}_1$ denotes the corresponding approximation to $A^{1/p}$ and r is such that $2\,r^{1-p}/(p-1) \le \epsilon$, then

$$\begin{aligned} \|A^{1/p} - \tilde{X}\| &\le \frac{p\,\sin(\pi/p)}{\pi} \|A\| \left(\|X_1 - \tilde{X}_1\| + \|X_2\| \right) \\ &\le \frac{p\,\sin(\pi/p)}{\pi} \|A\| \left(\epsilon + \frac{2\,r^{1-p}}{p-1} \right) \\ &\le \frac{2p\,\sin(\pi/p)}{\pi} \|A\| \epsilon. \end{aligned} \tag{2.2}$$

Let $E \in \mathbb{R}^{n\times n}$ be a matrix such that $A+E$ has no eigenvalues on the closed negative real axis. Using (1.1) together with some algebra, an integral representation for the perturbation of the matrix pth root follows,

$$(A+E)^{1/p} - A^{1/p} = \frac{p\,\sin(\pi/p)}{\pi} \int_0^\infty t^p (t^p I + A + E)^{-1} E (t^p I + A)^{-1}\, dt. \tag{2.3}$$

Considering only first-order perturbation arguments, an integral representation for the Fréchet derivative arises,

$$L_{x^{1/p}}(A, E) = \frac{p\,\sin(\pi/p)}{\pi} \int_0^\infty t^p (t^p I + A)^{-1} E (t^p I + A)^{-1}\, dt. \tag{2.4}$$

This formula can also be obtained from [3, Eq. (8)] by making the substitution $t = x^p$.

The analogue of Theorem 2.1 for the Fréchet derivative is stated below.

THEOREM 2.2. *Let $A, E \in \mathbb{R}^{n\times n}$. Assume in addition that A has no eigenvalue on the closed negative real axis and denote $g(t) = t^p(t^pI + A)^{-1}E(t^pI + A)^{-1}$. If $r \ge (2\|A\|)^{1/p}$, then*

$$L_{x^{1/p}}(A, E) = \frac{p\,\sin(\pi/p)}{\pi} \left(\int_0^r g(t)\, dt + \int_r^\infty g(t)\, dt \right), \tag{2.5}$$

where

$$\left\| \int_r^\infty g(t)\, dt \right\| \le \frac{2\,r^{1-p}}{p-1} \|E\|.$$

ETNA
Kent State University
http://etna.math.kent.edu

Proof. Proceed similarly to the proof of Theorem 2.1. □

Another manner of dealing with the integral over an infinite interval (1.1) is the reduction to an integral over a finite range by appropriately changing variables. One possibility is to consider the Cayley transform $t = (1+x)/(1-x)$. Some calculation enables one to conclude that

$$A^{1/p} = \frac{2p\,\sin(\pi/p)}{\pi} A \int_{-1}^{1} (1-x)^{p-2}\left[(1+x)^p I + (1-x)^p A\right]^{-1}\, dx. \tag{2.6}$$

Since we are assuming that A has no eigenvalues on the closed negative real axis, the integrand in (2.6) has no singularities in the interval $[-1,\,1]$. Moreover, the function is continuous on that interval. Other changes of variable also result in finite intervals. For instance, $t = x/(1-x)$ leads to

$$A^{1/p} = \frac{p\,\sin(\pi/p)}{\pi} A \int_{0}^{1} (1-x)^{p-2}\left[x^p I + (1-x)^p A\right]^{-1}\, dx,$$

and $t = \tan\theta$ to

$$A^{1/p} = \frac{p\,\sin(\pi/p)}{\pi} A \int_{0}^{\pi/2} (\cos\theta)^{p-2}\left[(\sin\theta)^p\, I + (\cos\theta)^p\, A\right]^{-1}\, d\theta.$$

Assume that r satisfies the assumptions of Theorem 2.1. The substitution $t = (1+x)/(1-x)$ in both the integrals on the right-hand side of (2.1) leads to a splitting of the integral (2.6), allowing us to write

$$\begin{aligned} A^{1/p} = \frac{2p\,\sin(\pi/p)}{\pi} A \Bigg(&\int_{-1}^{\frac{r-1}{r+1}} (1-x)^{p-2}\left[(1+x)^p I + (1-x)^p A\right]^{-1}\, dx \\ &+ \int_{\frac{r-1}{r+1}}^{1} (1-x)^{p-2}\left[(1+x)^p I + (1-x)^p A\right]^{-1}\, dx \Bigg), \end{aligned} \tag{2.7}$$

with

$$\left\| \int_{\frac{r-1}{r+1}}^{1} (1-x)^{p-2}\left[(1+x)^p I + (1-x)^p A\right]^{-1}\, dx \right\| \leq \frac{2\, r^{1-p}}{p-1}\,.$$

A change of variables can also turn the improper integral (2.4) into a proper integral. For instance, with $t = (1+x)/(1-x)$, the integral (2.4) can be transformed to

$$L_{x^{1/p}}(A,E) = \frac{2p\,\sin(\pi/p)}{\pi} \int_{-1}^{1} (1+x)^p (1-x)^{p-2}\left[h(x)\right]^{-1} E\left[h(x)\right]^{-1}\, dx, \tag{2.8}$$

where $h(x) = (1+x)^p I + (1-x)^p A$.

Alternative representations for the Fréchet derivative can be derived by performing the variable transformations $t = x/(1-x)$ and $t = \tan\theta$ in (2.4).

3. Bounds for $\|\mathbf{A^{1/p}}\|$ and $\|\mathbf{L_{x^{1/p}}(A)}\|$. Several bounds available in the literature for general matrix functions (see, for instance, [20, p. 102] and the references therein) can be adapted to the particular case of the matrix pth root. However, some of them seem to be of little interest for practical use, because they may not be sharp and it is not clear how to evaluate them. Our goal in this section is to derive new bounds for the matrix pth root and its

Fréchet derivative by means of the integral representations addressed in the previous section, and investigate under which conditions they may have interest from a practical point of view. We shall note that the problem of bounding $\|A^{1/p}\|$ and $\|L_{x^{1/p}}(A)\|$ reduces to bounding the resolvent functions that are involved in the integral representations. Bounds for the resolvent can be found for instance in [13] and [27], where we can observe that unless severe restrictions are imposed on the matrix A, finding a satisfactory bound valid for all A with no eigenvalues on the closed negative real axis seems to be out of reach.

Consider the resolvent involved in (1.1),

$$f(t) = (t^p I + A)^{-1}, \tag{3.1}$$

with $t \in [0, \infty[$. The value of the norm of the resolvent f depends in particular on how close the eigenvalues of A are to the closed negative real axis. To illustrate this, let us consider $p = 7$ and the matrix

$$A = \begin{bmatrix} e^a & 0 & 0 \\ 0 & \cos(\theta) & -\sin(\theta) \\ 0 & \sin(\theta) & \cos(\theta) \end{bmatrix}$$

with eigenvalues e^a and $\cos(\theta) \pm i\sin(\theta)$. Figure 3.1 displays the values of the norm of the resolvent $\|(t^p I + A)^{-1}\|$ against t for two different pairs of values: $a = -2,\ \theta = 3\pi/4$ and $a = -5,\ \theta = 3.13$. The peak above $t = 1$ is typical and becomes higher as θ approaches π (that is, as the two conjugate eigenvalues approach -1). This predicts some difficulties in bounding the corresponding resolvent.

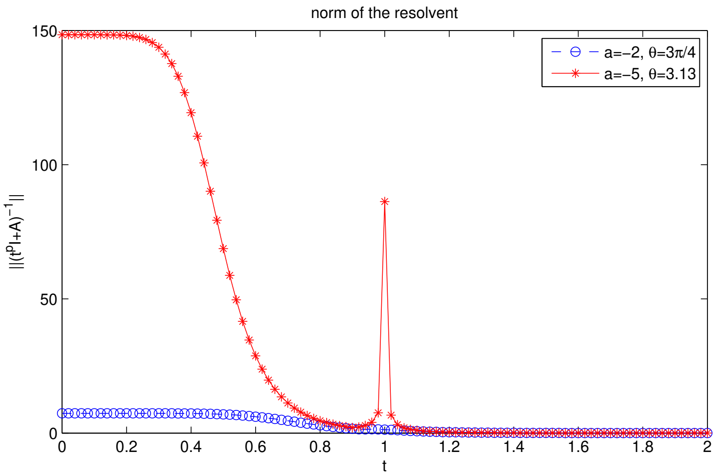

FIG. 3.1. *Norm of the resolvent* $\|(t^p I + A)^{-1}\|$ *for* $t \in [0, 2]$*, with* $p = 7$*, for* $a = -2,\ \theta = 3\pi/4$ *and* $a = -5,\ \theta = 3.13$.

This phenomenon is well understood in light of the pseudospectra theory. Recall that for a given matrix A and $\epsilon > 0$, the ϵ-*pseudospectrum* $\sigma_\epsilon(A)$ of A is the set of $z \in \mathbb{C}$ such that $\|(zI - A)^{-1}\| > \epsilon^{-1}$ (see [27]), that is, $\sigma_\epsilon(A)$ is the open subset of the complex plane bounded by the ϵ^{-1} level curve of the norm of the resolvent. For the matrix A defined above, Figure 3.2 shows the boundaries of $\sigma_\epsilon(A)$ for some values of ϵ between 10^{-4} and 10, from inner to outer. Eigenvalues are marked by a cross. The left-hand side plot corresponds to the values $a = -2,\ \theta = 3\pi/4$ and the right-hand side plot to $a = -5,\ \theta = 3.13$. Since

ETNA
Kent State University
http://etna.math.kent.edu

$-t^p \in]-\infty, 0]$, the norm of the resolvent (3.1) attains large values whenever the contours cross the closed negative real axis very closely to an eigenvalue of A.

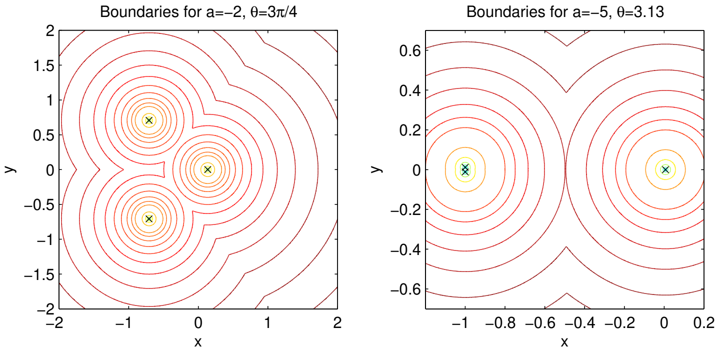

FIG. 3.2. *Boundaries of the ϵ-pseudospectrum of A with $a = -2$, $\theta = 3\pi/4$ (left) and $a = -5$, $\theta = 3.13$ (right) for some $\epsilon \in [10^{-4}, 10]$.*

Let $d(z, \sigma(A))$ denote the distance between the complex number z and the spectrum of A, $\sigma(A) = \{\lambda_1, \ldots, \lambda_n\}$, that is,

$$d(z, \sigma(A)) = \min_{\lambda \in \sigma(A)} |z - \lambda|.$$

Assuming that $z \notin \sigma(A)$, the following error estimate for the resolvent is available in [13, p. 12]:

$$\|(zI - A)^{-1}\|_2 \leq \sum_{k=0}^{n} \frac{(\gamma(A))^k}{\sqrt{k!}\,[d(z, \sigma(A))]^{k+1}}, \tag{3.2}$$

where

$$\gamma(A) = \left(\|A\|_F^2 - \sum_{k=1}^{n} |\lambda_k|^2 \right)^{1/2}$$

can be interpreted as a quantity measuring the departure of A from normality. If in particular A is normal, then $\gamma(A) = 0$, and therefore (3.2) simplifies to

$$\|(zI - A)^{-1}\|_2 = \frac{1}{d(z, \sigma(A))}.$$

Another simplification of (3.2) occurs when A is diagonalizable. Indeed, assuming that $A = SDS^{-1}$, with S nonsingular and D diagonal,

$$\|(zI - A)^{-1}\|_2 \leq \frac{\kappa(S)}{d(z, \sigma(A))},$$

where $\kappa(S) = \|S\|_2 \|S^{-1}\|_2$ stands for the condition number of S.

To compute an estimate for the norm of the resolvent using (3.2), it is helpful to write $d(z, \sigma(A))$ as an elementary function of z or to find a lower bound depending on z. This seems to be difficult for a general matrix A with no eigenvalues on the closed negative real

ETNA
Kent State University
http://etna.math.kent.edu

axis. However, assuming that all the eigenvalues of A lie on the open right-half plane, for the particular resolvent (3.1), the inequality

$$d(-t^p, \sigma(A)) \geq \beta(A) + t^p \tag{3.3}$$

holds for all $t \in [0, \infty[$, where $\beta(A) := \min\{\mathrm{Re}(\lambda) : \ \lambda \in \sigma(A)\}$. In particular, if the spectrum of A is real positive, (3.3) becomes an equality. The following result is a consequence of Theorem 2.1 and the discussion above.

THEOREM 3.1. *Let $A, E \in \mathbb{R}^{n\times n}$, with A having eigenvalues on the open right-half plane, and let $\beta(A) := \min\{\mathrm{Re}(\lambda) : \ \lambda \in \sigma(A)\}$. Assume that $r \geq (2\|A\|_2)^{1/p}$.*

(i) Putting $\gamma(A) = \left(\|A\|_F^2 - \sum_{k=1}^n |\lambda_k|^2\right)^{1/2}$, we have

$$\|A^{1/p}\|_2 \leq \frac{p\,\sin(\pi/p)}{\pi}\|A\|_2 \left(\int_0^r \sum_{k=0}^n \frac{[\gamma(A)]^k}{\sqrt{k!}\,(\beta(A)+t^p)^{k+1}}\,dt + \frac{2\,r^{1-p}}{p-1}\right)$$

and

$$\|L_{x^{1/p}}(A)\|_2 \leq \frac{p\,\sin(\pi/p)}{\pi}\left(\int_0^r t^p \left(\sum_{k=0}^n \frac{[\gamma(A)]^k}{\sqrt{k!}\,(\beta(A)+t^p)^{k+1}}\right)^2 dt + \frac{2\,r^{1-p}}{p-1}\right).$$

(ii) If A is diagonalizable, with $A = SDS^{-1}$, then

$$\|A^{1/p}\|_2 \leq \frac{p\,\sin(\pi/p)}{\pi}\|A\|_2 \left(\kappa(S)\int_0^r \frac{1}{\beta(A)+t^p}\,dt + \frac{2\,r^{1-p}}{p-1}\right)$$

and

$$\|L_{x^{1/p}}(A)\|_2 \leq \frac{p\,\sin(\pi/p)}{\pi}\left([\kappa(S)]^2\int_0^r \frac{t^p}{(\beta(A)+t^p)^2}\,dt + \frac{2\,r^{1-p}}{p-1}\right),$$

where $\kappa(S)$ is the condition number of S with respect to the 2-norm.

More bounds are given in the next theorem.

THEOREM 3.2. *Let $A \in \mathbb{R}^{n\times n}$ satisfy the condition $\|I - A\| = \omega < 1$, and assume that $r \geq (2\|A\|)^{1/p}$. Then*

$$\|A^{1/p}\| \leq \frac{p\,\sin(\pi/p)}{\pi}\|A\|\left(\int_0^r \frac{1}{1-\omega+t^p}\,dt + \frac{2\,r^{1-p}}{p-1}\right) \tag{3.4}$$

and

$$\|L_{x^{1/p}}(A)\| \leq \frac{p\,\sin(\pi/p)}{\pi}\left(\int_0^r t^p\left(\frac{1}{1-\omega+t^p}\right)^2 dt + \frac{2\,r^{1-p}}{p-1}\right). \tag{3.5}$$

Proof. Since $t \in [0,\,\infty[$ and $\|I - A\| = \omega < 1$, we can write

$$\begin{aligned}(t^pI + A)^{-1} &= ((t^p+1)I - (I-A))^{-1}\\ &= \frac{1}{t^p+1}\left(I - \frac{I-A}{t^p+1}\right)^{-1}\end{aligned}$$

ETNA
Kent State University
http://etna.math.kent.edu

and then

$$\begin{aligned}\|(t^pI+A)^{-1}\| &\le \left(\frac{1}{t^p+1}\right)\left(\frac{1}{1-\frac{\omega}{t^p+1}}\right)\\ &= \frac{1}{1-\omega+t^p}.\end{aligned}\tag{3.6}$$

By Theorem 2.1, (3.4) follows. For any $E \in \mathbb{R}^{n\times n}$, Theorem 2.2 guarantees that

$$\|L_{x^{1/p}}(A,E)\| \le \frac{p\,\sin(\pi/p)}{\pi}\left(\int_0^r t^p\|E\|\,\left\|(t^pI+A)^{-1}\right\|^2\,dt+\|E\|\frac{2\,r^{1-p}}{p-1}\right).$$

Hence,

$$\max_{\|E\|=1}\|L_{x^{1/p}}(A,E)\| \le \max_{\|E\|=1}\left\{\frac{p\,\sin(\pi/p)}{\pi}\left(\int_0^r t^p\|E\|\,\left\|(t^pI+A)^{-1}\right\|^2\,dt\right.\right.$$
$$\left.\left.+\|E\|\frac{2\,r^{1-p}}{p-1}\right)\right\},$$

and, for the induced operator norm, we have the inequality

$$\|L_{x^{1/p}}(A)\| \le \frac{p\,\sin(\pi/p)}{\pi}\left(\int_0^r t^p\,\left\|(t^pI+A)^{-1}\right\|^2\,dt+\frac{2\,r^{1-p}}{p-1}\right).\tag{3.7}$$

Now (3.5) follows from (3.6) and (3.7). □

Note that all the integrals appearing in the bounds of Theorems 3.1 and 3.2 are scalar and thus can be estimated by scalar quadrature.

4. Matrix integrals.

Given a matrix valued function depending on a parameter

$$f : t \in [a,b] \longrightarrow f(t) \in \mathbb{R}^{n\times n}$$

satisfying some requirements related with integrability and differentiability, the integral $\int_a^b f(t)\,dt$ and the derivative $f'(t)$ are defined componentwise [14, Sec. 11.2.6]. With some precautions, scalar quadrature can be extended to matrix integrals. The following theorem plays an important role in the study of matrix integrals. It gives a bound for the truncation error arising in the approximation of a matrix integral by quadrature.

THEOREM 4.1. [23] *Let $[a,b]$ be a finite interval. Assume that $c > 0$, $t_i \in [a,b]$ and $w_i \in \mathbb{R}$, $i = 1,2,\ldots,m$, be such that for any scalar function g that is $k+1$ times differentiable on $[a,b]$,*

$$\left|\int_a^b g(t)\,dt - \sum_{i=1}^m w_i g(t_i)\right| \le c\,\max_{\xi\in[a,b]}|g^{(k+1)}(\xi)|.$$

Let $f : [a,b] \longrightarrow \mathbb{R}^{n\times n}$ be such that $f^{(k+1)}(t)$ exists for all $t \in [a,b]$. Then

$$\left\|\int_a^b f(t)\,dt - \sum_{i=1}^m w_i f(t_i)\right\| \le c\,\max_{s\in[a,b]}\|f^{(k+1)}(s)\|.$$

ETNA
Kent State University
http://etna.math.kent.edu

At first glance, Theorem 4.1 may lead us to think that the error formula of a given scalar quadrature

$$\int_a^b g(t)\,dt - \sum_{i=1}^m w_i g(t_i) = c\, g^{(k+1)}(\xi),$$

for some $\xi \in [a,b]$, can be extended to matrix quadrature. Unfortunately, this is not true in general because a choice of a single $\xi \in [a,b]$ such that

$$\int_a^b f(t)\,dt - \sum_{i=1}^m w_i f(t_i) = c\, f^{(k+1)}(\xi),$$

for a matrix valued function $f : [a,b] \longrightarrow \mathbb{R}^{n\times n}$, may not be possible. For a simple counter–example, consider $f(t)$ as being an 2×2 diagonal matrix with different entries.

One of the aims of this paper is to investigate numerical quadrature for computing $A^{1/p}$, $A^{1/p}b$ and $L_{x^{1/p}}(A,E)$. Many numerical methods to approximate integrals are available (see for instance [8] and [10]), but we restrict our study to three popular methods: composite trapezoidal, Gauss-Legendre and adaptive Simpson rules.

Given $f : [a,b] \longrightarrow \mathbb{R}^{n\times n}$ having derivatives of second order for $t \in [a,b]$, the composite trapezoidal rule allows one to write

$$\int_a^b f(t)\,dt = \mathcal{T}(h) + \epsilon_T,$$

where

$$\mathcal{T}(h) = \frac{h}{2}(f(t_0) + f(t_m)) + h \sum_{k=1}^{m-1} f(t_k) \tag{4.1}$$

and ϵ_T denotes the truncation error. Recall that $t_0 = a, t_1, \ldots, t_m = b$ are equally spaced points partitioning the interval $[a,b]$ and $h = t_k - t_{k-1}$.

By Theorem 4.1 the composite trapezoidal truncation error can be bounded by

$$\|\epsilon_T\| \le \frac{b-a}{12} h^2 \max_{s\in[a,b]} \|f''(s)\|. \tag{4.2}$$

This error formula raises the question of how to find a bound for f'' on $[a,b]$. This is a major difficulty in the case of the integrals representing $A^{1/p}$ and $L_{x^{1/p}}(A,E)$ because the integrands involve resolvents; see the discussion on bounding resolvents in the previous section. Nevertheless, for a matrix A sufficiently close to the identity such that $\|I - A\| < 1$ finding a bound for the second derivative of f is possible, as we will see later in (5.1). The need of this restriction on A is also reported in [9] for the matrix logarithm.

Another technique to estimate the trapezoidal truncation error is based on Richardson extrapolation. For a sufficiently small h, the composite trapezoidal rule satisfies [8, pp. 10, 529]

$$\left\| \int_a^b f(t)\,dt - \mathcal{T}\left(\frac{h}{2}\right) \right\| \approx \frac{1}{3} \left\| \mathcal{T}(h) - \mathcal{T}\left(\frac{h}{2}\right) \right\|, \tag{4.3}$$

with $\mathcal{T}(h)$ defined by (4.1).

ETNA
Kent State University
http://etna.math.kent.edu

The relation (4.3) is very useful in practical computations because it avoids the use of derivatives. Recall that a similar relation holds for the composite Simpson rule, which is the basis of the adaptive Simpson quadrature.

It is worth noting that when the number of subintervals m is doubled the function evaluations in $\mathcal{T}(h)$ can be reused for $\mathcal{T}(h/2)$. Since the computation of a matrix function $f(t) \in \mathbb{R}^{n\times n}$ involves in general $O(n^3)$ arithmetic operations, this represents an important advantage of the trapezoidal rule for matrix integrals.

The m-point Gauss-Legendre quadrature rule is a widely used method for numerical evaluation of integrals,

$$\int_{-1}^{1} f(t)\, dt = \sum_{i=1}^{m} w_i f(t_i) + \epsilon_{GL}, \tag{4.4}$$

with ϵ_{GL} representing the truncation error. The w_i's are called the weights and the t_i's are the nodes [10]. For several values of m, the weights and the nodes can be found in the literature and several routines are available for their computation [1]. Attending to the formula for the scalar truncation error (see, for instance, [10, (2.7.11)]) and to Theorem 4.1, the truncation error for matrix quadrature can be bounded by

$$\|\epsilon_{GL}\| \leq \frac{2^{2m+1}(m!)^4}{(2m+1)((2m)!)^3} \max_{s\in[a,b]} \|f^{(2m)}(s)\|. \tag{4.5}$$

For integrals over $[a, b]$, the change of variable

$$t = \frac{1}{2}((b-a)x + (a+b))$$

maps the interval $[a, b]$ onto the standard interval $[-1, 1]$. The Gauss-Legendre rule is very popular in the scalar case, which is due in part to its optimality properties. Nevertheless it has the drawback of not allowing the reuse of the function evaluations when passing from m to $2m$. One possible way to overcome this is to consider Gauss-Kronrod rules (see [8, sec. 5.3.3] and the references therein), which are constructed from Gaussian rules. The extension of Gauss-Kronrod rules to matrix functions is not addressed here but it seems to be a very interesting topic for future research. The truncation error estimate (4.5) may be useless if the expression of the nth derivative of f is unknown or complicated. An alternative is to use an estimate similar to (4.3).

Let $\mathcal{G}(m) := \sum_{i=1}^{m} w_i f(t_i)$ be the m-point Gauss-Legendre quadrature and let $X := \int_{-1}^{1} f(t)\, dt$. By (4.5) it can be shown that $\|\mathcal{G}(m) - X\|$ tends to zero whenever $m \to \infty$. Assume that m is sufficiently large so that $\|\mathcal{G}(m) - X\| = \epsilon$ and $\|\mathcal{G}(2m) - X\| = c\epsilon$, with $0 < c \leq 0.5$. If $\|\mathcal{G}(m) - \mathcal{G}(2m)\| \leq \tilde{\epsilon}$, then

$$\|\mathcal{G}(m) - X\| \leq \|\mathcal{G}(m) - \mathcal{G}(2m)\| + \|\mathcal{G}(2m) - X\|,$$

that is, $\epsilon \leq \tilde{\epsilon} + c\epsilon$, or equivalently, $\epsilon \leq \frac{1}{1-c}\tilde{\epsilon}$. Hence

$$\|\mathcal{G}(2m) - X\| \leq \|\mathcal{G}(m) - \mathcal{G}(2m)\|. \tag{4.6}$$

The third method that we are concerned with is the adaptive Simpson quadrature [8, 10, 12]. In the scalar case, it involves extrapolation techniques and is particularly recommended for integrals with functions that strongly vary in different parts of the interval $[a, b]$. The MATLAB routine `quad` implements the algorithm of Gander and Gautschi [12]. Here we will use an adaptation of this algorithm for matrix integrals.

ETNA
Kent State University
http://etna.math.kent.edu

5. Computing $\mathbf{A}^{1/p}$ by quadrature. The composite trapezoidal rule applied to the integral (1.1) produces an approximation to the matrix pth root affected by the error ϵ_T, whose norm can be estimated by (4.2). This estimate involves second order derivatives of the integrand function $f(t) = (t^p I + A)^{-1}$, which can be given by the expression

$$f''(t) = pt^{p-2}[f(t)]^2(-(p-1)I + 2pt^p f(t)).$$

Under the assumption $\|I - A\| = \omega < 1$, (3.6) allows one to obtain the bounds

$$\|f(t)\| \leq \frac{1}{1-\omega}$$

and

$$\|t^p f(t)\| \leq \frac{r^p}{r^p + 1 - \omega},$$

that are valid for all $t \in [0, r]$, with $r > 0$. Therefore,

$$\|f''(t)\| \leq pr^{p-2} \left(\frac{1}{1-\omega}\right)^2 \left(p + 1 + 2p\frac{r^p}{r^p + 1 - \omega}\right), \tag{5.1}$$

for all $t \in [0, r]$. It turns out that this bound is not of much interest from a practical point of view. Indeed, some tests we have carried out showed that an estimate of the truncation error based on (5.1) may be very conservative and finding the number of subintervals in the trapezoidal rule by means of this bound may predict a larger m than one really needs. Moreover, it requires the strong restriction $\|I - A\| < 1$.

The same problem occurs with Gauss-Legendre rules, because the estimate (4.5) involves nth order derivatives of $f(t)$. With the assumption $\|I - A\| < 1$, a bound for the norm of the truncation error ϵ_{GL} (see (4.4)) may be obtained. Nevertheless, our experience with the bound (5.1) predicts a deterioration when the order of the derivatives increases.

By virtue of these difficulties in bounding the truncation error of quadrature and attending to (2.2), it may not be easy to find a minimal r in Theorem 2.1 that guarantees a prescribed accuracy. We have to deal with two sources of errors: the error arising from discarding the integral over the range $[r, \infty[$ and the quadrature truncation error. Moreover, some numerical experiments carried out with the integral (1.1) have shown that the number of function evaluations required in quadrature may be prohibitive. Thus, for practical purposes, it is preferable to work with the integral representation (2.6) instead of (1.1).

Two algorithms for the computation of the matrix pth root by quadrature applied to the integral (2.6) are proposed below. The first uses the composite trapezoidal rule and the second the Gauss-Legendre rule. To avoid the computation of the resolvent of matrices with eigenvalues nearby the closed negative real axis, the initial matrix A is preconditioned by the computation of one matrix square root [5, 19], that shifts all the eigenvalues to the open right half plane. This is possible because

$$A^{1/p} = \left[\left(A^{1/2^k}\right)^{1/p}\right]^{2^k}, \tag{5.2}$$

for all $k \in \mathbb{N}$. We recall that matrix square roots have been used successfully in the computation of the matrix logarithm [22] and the matrix pth root [21] in combination with Padé approximation. A prior Schur decomposition of $A = QTQ^T$ will also be computed. This costs about $25n^3$ (see [14, Algorithm 7.5.2]), but attending to the fact that many function evaluations have to be computed, this will contribute to reduce the computational cost. If T is triangular, evaluating the integrand in (2.6),

$$f(x) = (1-x)^{p-2}\left[(1+x)^p I + (1-x)^p T\right]^{-1},$$

by Gaussian elimination with partial pivoting requires about $n^3/3$ arithmetic operations . The number of subintervals in the composite trapezoidal rule will be estimated by (4.3).

According to (4.6), a possibility for estimating the number of nodes and weights in Gauss-Legendre rules is by requiring that $\|\mathcal{G}(2m) - \mathcal{G}(m)\|$ satisfies a prescribed tolerance, where $\mathcal{G}(m) := \sum_{i=1}^{m} w_i f(t_i)$. Unfortunately, it is not clear how to find a sufficiently large m to guarantee that (4.6) holds. An alternative is to require instead that the norm of the residual

$$\|\bar{X}^p - A\|, \tag{5.3}$$

be smaller than a given tolerance, where $\bar{X} := \frac{p\,\sin(\pi/p)}{\pi} A\,\mathcal{G}(m)$. If this tolerance is not met, m should be increased to, say, $2m$. Assuming that $X = A^{1/p}$ is the exact pth root of A, the residual (5.3) can be viewed as the backward error of $\bar{X}$, that is, it can be interpreted as a perturbation in A. Indeed, if F is a matrix such that $\bar{X} = (A+F)^{1/p}$, one has $F = \bar{X}^p - A$. A similar strategy is suggested in [4, Algorithm 2.1]. The Frobenius norm will be used throughout our experiments.

ALGORITHM 5.1. *Let $A \in \mathbb{R}^{n\times n}$ have no eigenvalues on the closed negative real axis, let $p \geq 2$ be an integer, m a positive integer and* `tol` *a given tolerance. This algorithm approximates $A^{1/p}$ by the composite trapezoidal rule for the integral (2.6).*

1. *Find the real Schur decomposition $A = QTQ^T$, where Q is orthogonal and T is quasi upper triangular;*
2. *Compute one square root of T; let $T_2 := T^{1/2}$;*
3. *Set $h = 2/m$ and $x_k = -1 + kh$, $k = 0, 1, \ldots, m$;*
4. *Compute $\mathcal{T}(h) := \frac{h}{2}(f(x_0) + f(x_m)) + h\sum_{k=1}^{m-1} f(x_k)$, where*
$$f(x) = (1-x)^{p-2}\left[(1+x)^p I + (1-x)^p T_2\right]^{-1},$$
and $\mathcal{T}(h/2)$;
5. *Double the number of subintervals $m \leftarrow 2m$ until $(1/3)\|\mathcal{T}(h) - \mathcal{T}(h/2)\|_F \leq$* `tol`*;*
6. *$A^{1/p} \approx \left(\frac{2p\,\sin(\pi/p)}{\pi}\right)^2 Q\,(T_2\,\mathcal{T}(h/2))^2\,Q^T$.*

Cost. $(29 + \frac{m}{3})n^3$.

The cost of Algorithm 5.1 can be interpreted as follows: $25n^3$ for the real Schur decomposition, $n^3/3$ for the computation of one matrix square root of a block triangular matrix in Step 2, $mn^3/3$ for m functions evaluations (note that m refers to the final number of subintervals) and $3n^3 + 2n^3/3$ to compute the approximation for the pth root in Step 6.

ALGORITHM 5.2. *Let $A \in \mathbb{R}^{n\times n}$ have no eigenvalues on the closed negative real axis, let $p \geq 2$ be an integer, m a positive integer and* `tol` *a given tolerance. This algorithm approximates $A^{1/p}$ by the Gauss-Legendre rule for the integral (2.6).*

1. *Find the real Schur decomposition of $A = QTQ^T$, where Q is orthogonal and T is quasi upper triangular;*
2. *Compute one square root of T, $T_2 := T^{1/2}$;*

ETNA
Kent State University
http://etna.math.kent.edu

3. *Compute* $\mathcal{G}(m) := \sum_{k=1}^{m} w_k f(x_k)$, *where*

$$f(x) = (1-x)^{p-2}\left[(1+x)^p I + (1-x)^p T_2\right]^{-1}$$

and w_k, x_k *are, respectively, the weights and nodes of the* m*-point Gauss-Legendre rule;*

4. *Denoting* $\tilde{X} := \frac{2p\,\sin(\pi/p)}{\pi} T_2\,\mathcal{G}(m)$*, double the number of weights (and nodes)* $m \leftarrow 2m$ *until* $\|\tilde{X}^p - T\|_F \leq \texttt{tol}$*;*
5. $A^{1/p} \approx \left(\frac{2p\,\sin(\pi/p)}{\pi}\right)^2 Q(T_2\,\mathcal{G}(m))^2\,Q^T$.

Cost. $28n^3 + (m_f+1)\frac{n^3}{3} + m_r$, where m_f is the total number of function evaluations and m_r is the total cost of the operations involved in Step 4.

In contrast to the trapezoidal rule, each time the number of weights (and nodes) is doubled the previously computed function evaluations cannot be reused. This represents an additional cost in Gauss-Legendre rules in comparison with trapezoidal rule. If m_0 is the initial number of nodes taken in Algorithm 5.2 and assuming that this number is doubled q times, then the total number of function evaluations is $m_f = (1+2+\ldots+2^q)m_0$. The computation of each $\tilde{X}$ in Step 4 involves about $n^3/3$ operations and the computation of the norm of the residual $\|\tilde{X}^p - A\|_F$ about $2n^3\lfloor\log_2 p\rfloor/3$, where $\lfloor a\rfloor$ denotes the floor of a, by the binary powering algorithm. So the total cost of Step 4 is $m_r = (q/3)(n^3 + 2n^3\lfloor\log_2 p\rfloor)$ operations. We are assuming that the nodes and weights are known.

The adaptive Simpson quadrature is another method that will be considered in our numerical examples. An algorithm for this successful method for approximating scalar integrals is proposed in [12] and is implemented in the MATLAB routine `quad`. Since it does not work with matrix integrals, we have carried out minor modifications and adapted it to matrices. The resulting algorithm includes a prior Schur decomposition and outputs the number of function evaluations m_f. The total cost is $28n^3 + (m_f+1)n^3/3$ arithmetic operations.

The three algorithms mentioned above, Algorithm 5.1, Algorithm 5.2 and the modified adaptive Simpson, were implemented in MATLAB with unit roundoff $u \approx 1.1 \times 10^{-16}$. The following twelve matrices were used in our tests:

$A_1 = 3 * \texttt{eye}(10) + \texttt{gallery}('\texttt{rando}', 10);$	$\kappa(A_1) = 14.6115;$
$A_2 = \texttt{gallery}('\texttt{lehmer}', 8);$	$\kappa(A_2) = 78.1523;$
$A_3 = 6 * \texttt{eye}(15) + \texttt{randn}(15);$	$\kappa(A_3) = 27.0730;$
$A_4 = 6 * \texttt{eye}(15) + \texttt{randn}(15);$	$\kappa(A_4) = 20.9047;$
$A_5 = \texttt{expm}(\texttt{rand}(10));$	$\kappa(A_5) = 357.8323;$
$A_6 = \texttt{expm}(\texttt{rand}(10));$	$\kappa(A_6) = 583.5014;$
$A_7 = \texttt{rand}(10)\hat{}2;$	$\kappa(A_7) = 1.6565 \times 10^4;$
$A_8 = \texttt{rand}(10)\hat{}2;$	$\kappa(A_8) = 2.0091 \times 10^4;$
$A_9 = \texttt{gallery}('\texttt{frank}', 8);$	$\kappa(A_9) = 3.0320 \times 10^5;$
$A_{10} = \texttt{pascal}(8);$	$\kappa(A_{10}) = 2.0667 \times 10^7;$
$A_{11} = \texttt{expm}(\texttt{randn}(10));$	$\kappa(A_{11}) = 1.3780 \times 10^6;$
$A_{12} = \texttt{randn}(10)\hat{}2;$	$\kappa(A_{12}) = 8.6168 \times 10^5.$

In the first experiment we assumed that $p = 7$, $\texttt{tol} = 10^{-5}$ and that the initial m in Algorithms 5.1 and 5.2 is $m = 20$. To decide about the quality of the computed result we evaluated the relative residual [15]

$$\rho_A(\bar{X}) := \frac{\|\bar{X}^p - A\|_F}{\|A\|_F \left\| \sum_{i=0}^{p-1} \left(\bar{X}^{p-1-i}\right)^T \otimes \bar{X}^i \right\|_F}, \tag{5.4}$$

where $\bar{X}$ is the computed pth root. The results are shown in Figure 5.1. The picture on the left-hand side plots the number of function evaluations involved in each computation and the picture on the right-hand side plots the values of the relative residual (5.4) associated with $\bar{X} \approx A^{1/7}$ by the three algorithms. We abbreviate `Trap` for Algorithm 5.1, `GL` for Algorithm 5.2 and `AS` for the modified adaptive Simpson.

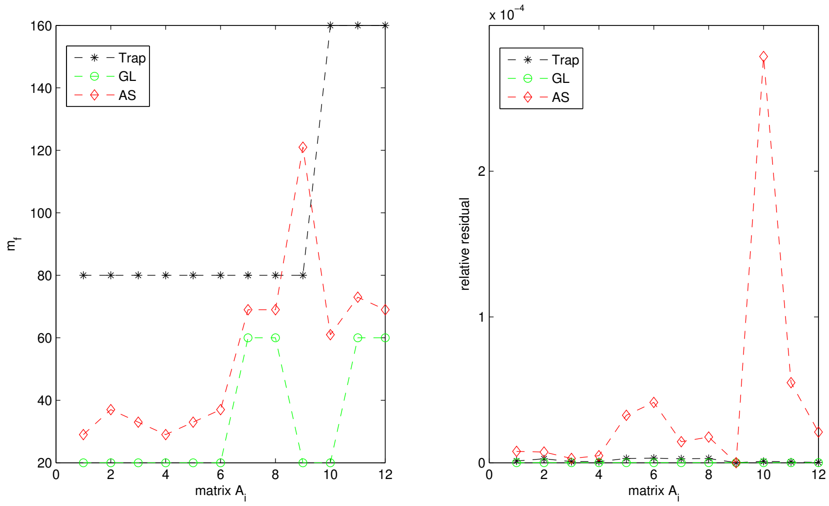

FIG. 5.1. *Number of function evaluations (left) and relative residual (right) for Algorithm 5.1 (*`Trap`*), Algorithm 5.2 (*`GL`*) and the modified adaptive Simpson (*`AS`*) with $p = 7$, $tol = 10^{-5}$, $m = 20$ (the initial value of m).*

Figure 5.1 shows that the Gauss-Legendre rule (Algorithm 5.2) performs considerably better than the other rules, both in number of function evaluations and relative residual. The adaptive Simpson quadrature does not perform as well as expected. Although it requires in general fewer function evaluations than the trapezoidal rule, the value of the relative residual is larger and strongly varies for the same tolerance.

The three algorithms were also tested for the smaller tolerance $\texttt{tol} = 10^{-14}$. We noticed that the computational effort increased considerably, with hundreds or even thousands of function evaluations involved. Then we combined the three algorithms with the square rooting and squaring technique which exploits (5.2). With the exception of the Gauss-Legendre quadrature, no significant reduction of the number of function evaluations has occurred. Although we have not found any connection between Padé approximation and Gauss-Legendre quadrature for the integral (2.6), this quadrature seems to work very well when combined with a prior computation of a certain number k of square roots ensuring the condition

$$\|I - A^{1/2^k}\| < 1.$$

This is similar to what happens with the matrix logarithm, for which Gauss-Legendre quadrature and diagonal Padé approximation are equivalent; see [11].

ETNA
Kent State University
http://etna.math.kent.edu

The results for the combination of Algorithm 5.2 with the square rooting and squaring technique are displayed in Figure 5.2. They show that Gauss-Legendre quadrature applied to the integral (2.6) can be seen as a promising method for the matrix pth root computation, despite being a bit more expensive than other methods such as the Schur-Newton [15] and the Schur-Padé [21] methods. This is more evident if we increase p. Figure 5.3 shows the behavior of Gauss-Legendre quadrature for $p = 3, 29, 53$, showing that the cost blows up with p. Gauss-Legendre quadrature is a topic that needs further research. In particular, one needs to know sharp error estimates, which are important to find, for instance, the optimal number of square roots required before applying the Gauss-Legendre quadrature.

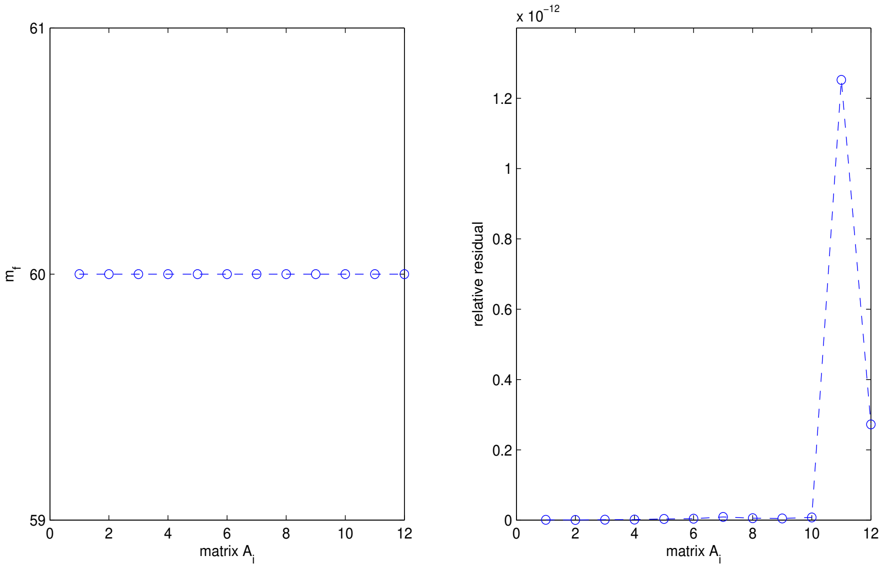

FIG. 5.2. *Number of function evaluations (left) and relative residual (right) for Algorithm 5.2 combined with the square rooting and squaring technique for $p = 7$, $tol = 10^{-14}$, $m = 20$.*

6. Computing $\mathbf{A}^{1/p}b$. Assuming that a given matrix function $f(A)$ allows an integral representation, quadrature provides an interesting method for computing the vector $f(A)b$, where $b \in \mathbb{R}^n$, without the explicit computation of $f(A)$ [9]. This method becomes more effective when combined with an initial reduction of A to a simpler form, such as Hessenberg or Schur forms. In this section we investigate the specific case of computing $A^{1/p}b$ by a quadrature rule applied to (2.6).

Let $A = QTQ^T$, with Q orthogonal and T quasi triangular, be the real Schur decomposition of A. For $f(x) = (1-x)^{p-2}\left[(1+x)^p I + (1-x)^p T\right]^{-1}$, we have

$$\int_{-1}^{1} f(x)\, dx \approx \sum_{k=1}^{m} w_k f(x_k),$$

where the values of w_k and x_k ($k = 1, \ldots, m$) depend on the chosen quadrature. Hence

$$Q\left(\int_{-1}^{1} f(x)\, dx\right) Q^T b \approx Q \sum_{k=1}^{m} w_k (1-x_k)^{p-2} y_k,$$

ETNA
Kent State University
http://etna.math.kent.edu

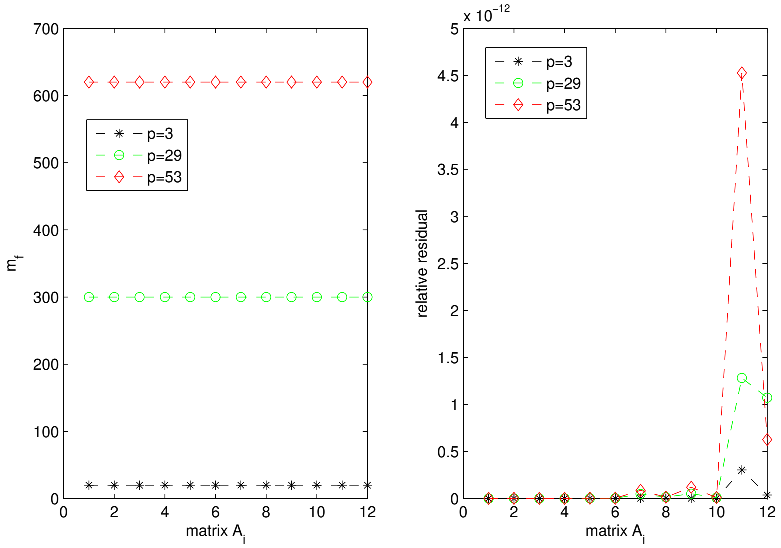

FIG. 5.3. *Number of function evaluations (left) and relative residual (right) for Algorithm 5.2 combined with the square rooting and squaring technique for three values of p: $p = 3, 29, 53$, and tol $= 10^{-10}$, $m = 20$.*

where each vector y_k is the $n \times 1$ vector solution of the $n \times n$ quasi triangular system of linear equations

$$[(1 + x_k)^p I + (1 - x_k)^p T]\, y_k = Q^T b. \tag{6.1}$$

Each linear system of this type can be solved in n^2 arithmetic operations which means that any function evaluation in the quadrature can be carried out in $O(n^2)$ arithmetic operations instead of $O(n^3)$ involved in the function evaluations for the matrix pth root quadrature; see Section 5. Thus the total cost for the matrix pth root times a vector using

$$A^{1/p} b \approx \frac{2p\,\sin(\pi/p)}{\pi}\, Q\, T \sum_{k=1}^{m} w_k (1 - x_k)^{p-2} y_k,$$

where y_k is given by (6.1), is about $26n^3 + 2mn^2$.

Composite trapezoidal, Gauss-Legendre and adaptive Simpson rules have been implemented in MATLAB. We have computed

$$A_i^{1/p} b_i, \quad i = 1, \ldots, 12$$

for the same matrices $A_1, \ldots, A_{12}$ tested in Section 5 and $\mathtt{b_i} = \mathtt{randn(lenght(A_i), 1)}$. To avoid too many function evaluations, the three rules were combined with the relation

$$A^{1/p} b = \left(A^{1/2}\right)^{1/p} \left(A^{1/2}\right)^{1/p} b,$$

which involves the prior computation of one square root of A_i. The main reason is to avoid the resolvent of a matrix with eigenvalues nearby the closed negative real axis. The three rules have to be applied twice: first to compute $\tilde{b} = \left(A^{1/2}\right)^{1/p} b$ and then to compute

ETNA
Kent State University
http://etna.math.kent.edu

$A^{1/p}b = \left(A^{1/2}\right)^{1/p}\tilde{b}$. At first glance it seems that this increases the total number of function evaluations but we shall note that, at least in our tests, the number of function evaluations required for computing $\tilde{b} = \left(A^{1/2}\right)^{1/p} b$ is in general less than half of the number required for $A^{1/p}b$. The results are shown in Figure 6.1 for $p = 7$, $\texttt{tol} = 10^{-5}$ and $m = 20$.

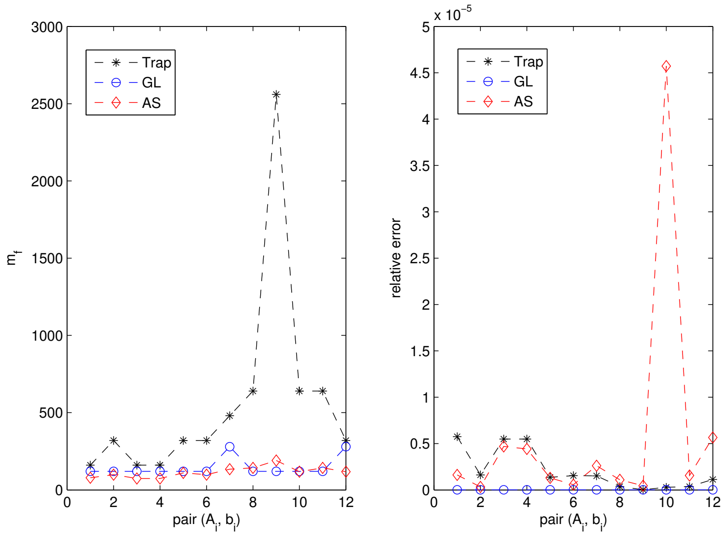

FIG. 6.1. *Number of function evaluations (left) and relative error (right) of composite trapezoidal, Gauss-Legendre and adaptive Simpson rules for approximating the vector $A_i^{1/p}b_i$, $i = 1,\ldots,12$, combined with the computation of one square root; $p = 7$, $\texttt{tol} = 10^{-5}$, $m = 20$.*

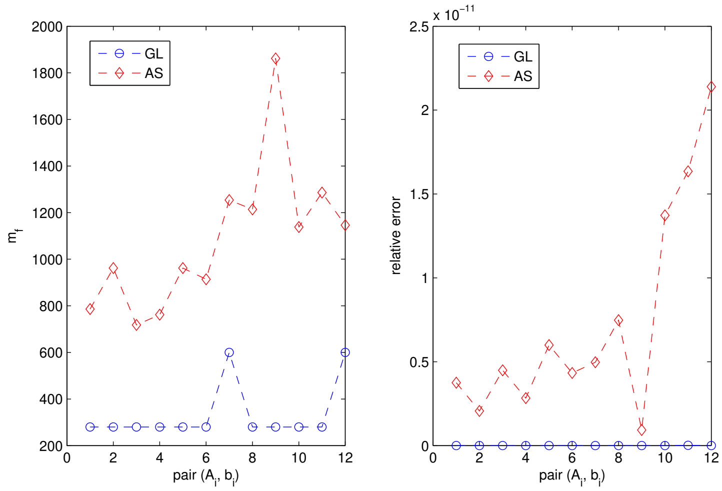

FIG. 6.2. *Number of function evaluations (left) and relative error (right) of Gauss-Legendre and Adaptive Simpson rules for approximating the vector $A_i^{1/p}b_i$, $i = 1,\ldots,12$, combined with the computation of one square root with $p = 7$, tol $= 10^{-10}$, $m = 20$.*

ETNA
Kent State University
http://etna.math.kent.edu

To estimate the relative error

$$\frac{\|\bar{b} - A^{1/p}b\|_F}{\|A^{1/p}b\|_F},$$

with $\bar{b}$ being the computed approximation for $A^{1/p}b$, we have assumed that the "exact" vector $A^{1/p}b$ is the result of multiplying the computed $A^{1/p}$ (with relative residual less or equal than the unit roundoff) by the vector b.

Since the number of function evaluations required by the composite trapezoidal rule is the largest, this rule has been excluded in the next experiment, where the tolerance has been reduced to $\mathtt{tol} = 10^{-10}$. The results are depicted in Figure 6.2, which evidences once more the good performance of the Gauss-Legendre rule.

7. Computing the Fréchet derivative. Let $A, E \in \mathbb{R}^{n\times n}$, with A having no eigenvalue on the closed negative real axis. Denoting $\alpha(x) := (1+x)^p(1-x)^{p-2}$ and $h(x) := (1+x)^p I + (1-x)^p A$, the integrand in (2.8) can be written as

$$g(x) := \alpha(x)\,[h(x)]^{-1}\,E\,[h(x)]^{-1}.$$

For any $x \in [-1, 1]$, computing $g(x)$ is equivalent to solving two coupled matrix equations,

$$h(x)X = \alpha(x)E, \quad Y h(x) = X,$$

where X and Y represent the matrices to be determined. If A is triangular, each function evaluation $g(x)$ costs about $2n^3$ operations.

To investigate quadrature for the integral (2.8), we proceed as in Section 5. We consider three algorithms for evaluating the Fréchet derivative $L_{x^{1/p}}(A, E)$: Algorithm 7.1, which is based on the composite trapezoidal rule, Algorithm 7.2, involving Gauss-Legendre quadrature, and a modification of the Gander and Gautschi's adaptive Simpson method. All the algorithms involve a prior Schur decomposition of A to reduce the cost and the computation of one matrix square root to avoid the evaluation of the resolvent of a matrix with eigenvalues nearby the closed negative real axis. It is possible to combine the algorithms with both the Schur decomposition and matrix square roots by virtue of the following two properties of the Fréchet derivative,

$$L_{x^{1/p}}(A, E) = Q\, L_{x^{1/p}}(T, Q^T E Q)\, Q^T, \tag{7.1}$$

where $A = QTQ^T$, with Q orthogonal and T quasi triangular, is the Schur decomposition of A (see [20, Problem 3.2]) and

$$L_{x^{1/p}}(A, E) = L_{x^2}\left(A^{1/p}, L_{x^{1/p}}\left(A^{1/2}, L_{x^{1/2}}(A, E)\right)\right). \tag{7.2}$$

The identity (7.2) follows immediately from the application of the chain rule [20, Th. 3.4] to the identity

$$A^{1/p} = \left(\left(A^{1/2}\right)^{1/p}\right)^2.$$

We shall recall that $L_1 := L_{x^{1/2}}(A, E)$ is the unique matrix that satisfies the Sylvester equation $A^{1/2}\, L_1 + L_1\, A^{1/2} = E$ and that $L_{x^2}(A, E) = AE + EA$; see [20, ch. 6]. One of the most popular methods for solving the Sylvester equation is due to Bartels and Stewart [2]. MATLAB codes for this method are available in the Matrix Function Toolbox [18]. If T is triangular, finding X such that $TX + XT = E$ requires about $2n^3$ arithmetic operations.

ETNA
Kent State University
http://etna.math.kent.edu

ALGORITHM 7.1. *Let $A, E \in \mathbb{R}^{n\times n}$, with A having no eigenvalue on the closed negative real axis, let $p \geq 2$ be an integer, m a positive integer and* `tol` *a given tolerance. This algorithm approximates $L_{x^{1/p}}(A, E)$ by the composite trapezoidal rule for the integral (2.8).*

1. *Find the real Schur decomposition of $A = QTQ^T$, where Q is orthogonal and T is quasi upper triangular;*
2. *Compute one square root of T, $T_2 := T^{1/2}$;*
3. *Evaluate $E_1 := Q^T EQ$;*
4. *Find L_1 in the Sylvester equation $T_2 L_1 + L_1 T_2 = E_1$;*
5. *Set $h = 2/m$ and $x_k = -1 + kh$, $k = 0, 1, \ldots, m$;*
6. *Compute $\mathcal{T}(h) = \frac{h}{2}(f(x_0) + f(x_m)) + h\sum_{k=1}^{m-1} f(x_k)$, where*
$$f(x) = (1+x)^p(1-x)^{p-2}\left[(1+x)^pI + (1-x)^pT_2\right]^{-1} L_1 \left[(1+x)^pI + (1-x)^pT_2\right]^{-1},$$
and $\mathcal{T}(h/2)$;
7. *Double the number of subintervals $m \leftarrow 2m$ until $(1/3)\|\mathcal{T}(h)-\mathcal{T}(h/2)\|_F <$* `tol`*;*
8. *$L_2 := \frac{2p\,\sin(\pi/p)}{\pi}\mathcal{T}(h/2)$,*
9. *$L_3 := T_2^{1/p}L_2 + L_2T_2^{1/p}$;*
10. *$L_{x^{1/p}}(A, E) \approx QL_3Q^T$.*

Cost. $(36 + 2m + \frac{1}{3})n^3$.

ALGORITHM 7.2. *Let $A \in \mathbb{R}^{n\times n}$ have no eigenvalues on the closed negative real axis, let $p \geq 2$ be an integer, m a positive integer and* `tol` *a given tolerance. This algorithm approximates $L_{x^{1/p}}(A, E)$ by the Gauss-Legendre quadrature for the integral (2.8).*

1. *Find the real Schur decomposition of $A = QTQ^T$, where Q is orthogonal and T is quasi upper triangular;*
2. *Compute one square root of T, $T_2 := T^{1/2}$;*
3. *Evaluate $E_1 := Q^T EQ$;*
4. *Find L_1 in the Sylvester equation $T_2 L_1 + L_1 T_2 = E_1$;*
5. *Compute $\mathcal{G}(m) = \sum_{k=1}^{m} w_k f(x_k)$, where*
$$f(x) = (1+x)^p(1-x)^{p-2}\left[(1+x)^pI + (1-x)^pT_2\right]^{-1} L_1 \left[(1+x)^pI + (1-x)^pT_2\right]^{-1},$$
and w_k, x_k are, respectively, the weights and nodes of the m-point Gauss-Legendre quadrature;
6. *Double the number of weights (nodes) until $\|\mathcal{G}(2m) - \mathcal{G}(m)\|_F <$* `tol` *(see (4.6));*
7. *$L_2 := \frac{2p\,\sin(\pi/p)}{\pi}\mathcal{G}(2m)$,*
8. *$L_3 := T_2^{1/p}L_2 + L_2T_2^{1/p}$;*
9. *$L_{x^{1/p}}(A, E) \approx QL_3Q^T$.*

Cost. $(36 + 2m_f + \frac{1}{3})n^3$, where m_f is the total number of function evaluations.

Algorithm 7.1, Algorithm 7.2 and a modified version of the adaptive Simpson quadrature were implemented in MATLAB. The modified adaptive Simpson was also combined with the Schur decomposition and the computation of one square root. The results for the Fréchet derivatives

$$L_{x^{1/p}}(A_i, E_i), \quad i = 1, 2, \ldots, 12,$$

where the A_i's are the matrices of Section 5 and `Ei = randn(length(Ai))`, are displayed in Figure 7.1. The relative residual is the same that has been used in [6, Eq. (5.1)]:

$$\rho(A, E) = \frac{\|M\,\mathrm{vec}(\tilde{L}) - \mathrm{vec}(E)\|_F}{\|M\|_F\|\,\mathrm{vec}(\tilde{L})\|_F}\,, \tag{7.3}$$

where $\tilde{L} \approx L_{x^{1/p}}(A, E)$ and $M := \sum_{j=0}^{p-1} \left[\left(A^{1/p}\right)^T \right]^j \otimes \left(A^{1/p}\right)^{p-1-j}$. We can observe that in our tests the modified adaptive Simpson has a poor performance. It requires the largest number of function evaluations and has the highest relative residual. Surprisingly, it is the trapezoidal rule that gives the best results.

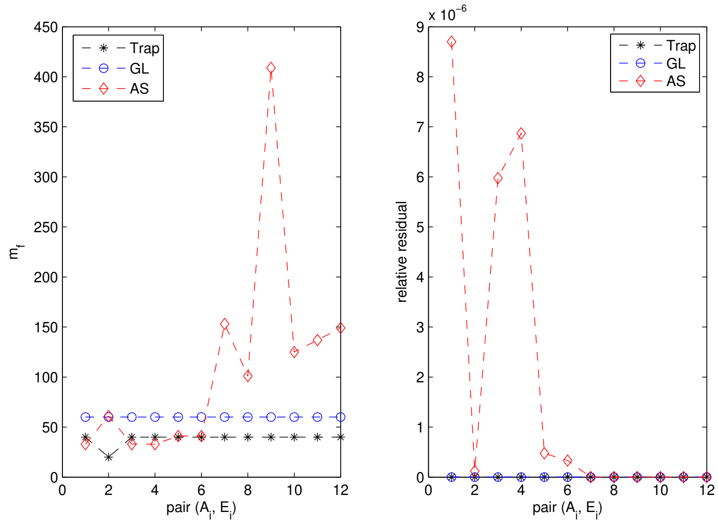

FIG. 7.1. *Number of function evaluations (left) and relative residual (right) for Algorithm 7.1 (`Trap`), Algorithm 7.2 (`GL`) and the modified adaptive Simpson (`AS`) with $p = 7$, `tol` $= 10^{-5}$ and $m = 20$.*

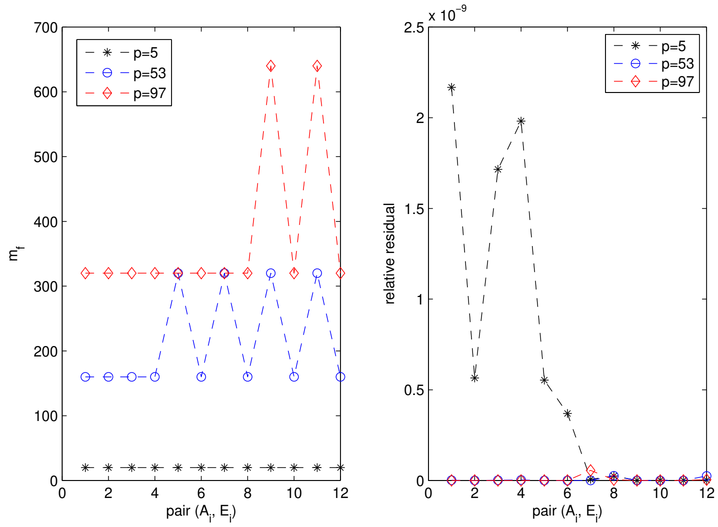

FIG. 7.2. *Number of function evaluations (left) and relative residual (right) for Algorithm 7.1 with $p = 3,\ 53,\ 97$, `tol` $= 10^{-5}$, $m = 20$.*

The formulae for the total cost of the algorithms do not depend directly on p. But our experience with the computation of the pth root in Section 5 tell us that the number of function evaluations is likely to increase. This is clear in Figure 7.2, where the Fréchet derivative $L_{x^{1/p}}(A_i, E_i)$, $i = 1, \ldots, 12$, is evaluated by the trapezoidal rule for three different values of p.

8. Conclusions. In this work we have derived new integral representations for the matrix pth root and its Fréchet derivative. Such integral representations have been used to bound those functions and to develop algorithms for their computation. Three numerical integration methods have been considered: composite trapezoidal rule, Gaussian-Legendre rule and adaptive Simpson quadrature. Our experiments have shown in particular that the combination of Gaussian quadrature with matrix square roots and squaring can be seen as an effective method for the computation of the matrix pth root, whereas the composite trapezoidal rule has revealed to be a good choice for the Fréchet derivative, at least in computations that do not require high accuracy. The approximation of the matrix pth root times a vector by quadrature has been also addressed. The Gauss-Legendre rule has proved to be once more the right choice to work out that approximation. However, the Gauss-Legendre rule for the matrix pth root has not been completely understood yet, mainly because practical error estimates for the truncation error are lacking. This is a topic that needs further research.

REFERENCES

[1] M. ABRAMOWITZ AND I. A. STEGUN, EDS., *Handbook of Mathematical Functions*, Dover, Mineola, NY, 1965.

[2] R. H. BARTELS AND G. W. STEWART, *Algorithm 432: Solution of the matrix equation AX+XB=C*, Comm. ACM, 15 (1972), pp. 820–826.

[3] R. BHATIA AND M. UCHIYAMA, *The operator equation* $\sum_{i=0}^{n} A^{n-i} X B^i = Y$, Expo. Math., 27 (2009), pp. 251–255.

[4] D. A. BINI, N. J. HIGHAM, AND B. MEINI, *Algorithms for the matrix pth root*, Numer. Algorithms, 39 (2005), pp. 349–378.

[5] Å. BJÖRCK AND S. HAMMARLING, *A Schur method for the square root of a matrix*, Linear Algebra Appl., 52/53 (1983), pp. 127–140.

[6] J. R. CARDOSO, *Evaluating the Fréchet derivative of the matrix pth root*, Electron. Trans. Numer. Anal., 38 (2011), pp. 202–217.
http://etna.math.kent.edu/vol.38.2011/pp202-217.dir

[7] E. CINQUEMANI, A. MILIAS-ARGEITIS, S. SUMMERS, AND J. LYGEROS, *Local identifications of piecewise deterministic models of genetics networks*, in Hybrid Systems: Computation and Control, R. Majumdar and P. Tabuada, eds., Lecture Notes in Comput. Sci., 5469, Springer, New York, 2009, pp. 105–119.

[8] G. DAHLQUIST AND Å. BJÖRCK, *Numerical Methods in Scientific Computing*, Vol. I, SIAM, Philadelphia, 2008.

[9] P. I. DAVIES AND N. J. HIGHAM, *Computing f(A)b for matrix functions f*, in QCD and Numerical Analysis III, A. Boriçi, A. Frommer, B. Joó, A. Kennedy, and B. Pendleton, eds., Lect. Notes Comput. Sci. Eng., 47, Springer, Berlin, 2005, pp. 15–24.

[10] P. J. DAVIS AND P. RABINOWITZ, *Methods of Numerical Integration*, Second ed., Academic Press, London, 1984.

[11] L. DIECI, B. MORINI, AND A. PAPINI, *Computational techniques for real logarithms of matrices*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 570–593.

[12] W. GANDER AND W. GAUTSCHI, *Adaptive quadrature – revisited*, BIT, 40 (2000), pp. 84–101.

[13] M. I. GIL', *Operator Functions and Localization of Spectra*, Lectures Notes in Math., 1830, Springer, Berlin, 2003.

[14] G. GOLUB AND C. VAN LOAN, *Matrix Computations*, Third ed., Johns Hopkins University Press, Baltimore and London, 1996.

[15] C.-H. GUO AND N. J. HIGHAM, *A Schur-Newton method for the matrix pth root and its inverse*, SIAM J. Matrix Anal. Appl., 28 (2006), pp. 788–804.

[16] N. HALE, N. J. HIGHAM, AND L. N. TREFETHEN, *Computing* A^α, $\log(A)$, *and related matrix functions by countour integrals*, SIAM J. Numer. Anal., 46 (2008), pp. 2505–2523.

[17] M. A. HASAN, J. A. HASAN, AND L. SCHARENROICH, *New integral representations and algorithms for*

computing nth roots and the matrix sector function of nonsingular complex matrices, in Proc. of the 39th IEEE Conf. on Decision and Control, Sydney, Australia (2000), IEEE Control Systems Society, Piscataway, NJ, 2000, pp. 4247–4252.

[18] N. J. HIGHAM, *The Matrix Function Toolbox*, `http://www.ma.man.ac.uk/~higham/mftoolbox`.

[19] N. J. HIGHAM, *Computing real square roots of a real matrix,* Linear Algebra Appl., 88/89 (1987), pp. 405–430.

[20] N. J. HIGHAM, *Functions of Matrices: Theory and Computation*, SIAM, Philadelphia, PA, 2008.

[21] N. J. HIGHAM AND L. LIN, *A Schur–Padé algorithm for fractional powers of a matrix*, SIAM J. Matrix Anal. and Appl., 32 (2011), pp. 1056–1078.

[22] C. KENNEY AND A. J. LAUB, *Condition estimates for matrix functions*, SIAM J. Matrix Anal. Appl., 10 (1989), pp. 191–209.

[23] R. MATHIAS, *Approximation of matrix-valued functions*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 1061–1063.

[24] J. L. SILVÁN-CARDENAS, *Sub-pixel Remote Sensing for Mapping and Modelling Invasive Tamarix: A Case Study in West Texas,* 1993−2005, Theses and Dissertations-Geography, Paper 27, Texas State University, San Marcos, 2009.

[25] M. I. SMITH, *Numerical Computation of Matrix Functions*, PhD thesis, University of Manchester, Manchester, UK, 2002.

[26] Y.-W. TAI, P. TAN, AND M. S. BROWN, *Richardson-Lucy deblurring for scenes under a projective motion path*, IEEE Trans. on Pattern Anal. and Machine Intelligence, 33 (2011), pp. 1603–1618.

[27] L. N. TREFETHEN AND M. EMBREE, *Spectra and Pseudospectra–The Behaviour of Nonnormal Matrices and Operators*, Princeton University Press, Princeton, NJ, 2005.