Linear Algebra and its Applications 504 (2016) 354–371

Contents lists available at ScienceDirect

Linear Algebra and its Applications

www.elsevier.com/locate/laa

Taylor's theorem for matrix functions with applications to condition number estimation ☆

Edvin Deadman, Samuel D. Relton *

School of Mathematics, The University of Manchester, Manchester, M13 9PL, UK

ARTICLE INFO

Article history:
Received 27 April 2015
Accepted 6 April 2016
Available online 18 April 2016
Submitted by V. Mehrmann

MSC:
15A12
15A16

Keywords:
Matrix function
Taylor polynomial
Remainder
Condition number
Pseudospectrum
Fréchet derivative

ABSTRACT

We derive an explicit formula for the remainder term of a Taylor polynomial of a matrix function. This formula generalizes a known result for the remainder of the Taylor polynomial for an analytic function of a complex scalar. We investigate some consequences of this result, which culminate in new upper bounds for the level-1 and level-2 condition numbers of a matrix function in terms of the pseudospectrum of the matrix. Numerical experiments show that, although the bounds can be pessimistic, they can be computed much faster than the standard methods. This makes the upper bounds ideal for a quick estimation of the condition number whilst a more accurate (and expensive) method can be used if further accuracy is required. They are also easily applicable to more complicated matrix functions for which no specialized condition number estimators are currently available.

☆ This work was supported by European Research Council Advanced Grant MATFUN (267526).
* Corresponding author.
E-mail addresses: edvin.deadman@manchester.ac.uk (E. Deadman), samuel.relton@manchester.ac.uk (S.D. Relton).

http://dx.doi.org/10.1016/j.laa.2016.04.010

1. Introduction

Taylor's theorem is a standard result in elementary calculus (see e.g. [17]). If $f : \mathbb{R} \to \mathbb{R}$ is k times continuously differentiable at $a \in \mathbb{R}$, then the theorem states that there exists $R_k : \mathbb{R} \to \mathbb{R}$ such that

$$f(x) = \sum_{j=0}^{k} \frac{f^{(j)}(a)}{j!}(x-a)^j + R_k(x)$$

and $R_k(x) = o(|x-a|^k)$ as $x \to a$. Depending on any additional assumptions on f, various precise formulae for the remainder term $R_k(x)$ are available. For example, if f is $k+1$ times continuously differentiable on the closed interval between a and x, then

$$R_k(x) = \frac{f^{(k+1)}(c)}{(k+1)!}(x-a)^{k+1} \tag{1}$$

for some c between a and x. This is known as the Lagrange form of the remainder. Alternative expressions, such as the Cauchy form or the integral form for the remainder are well known [17].

Taylor's theorem generalizes to analytic functions in the complex plane: the remainder must now be expressed in terms of a contour integral. If $f(z)$ is complex analytic in an open subset $\mathcal{D} \subset \mathbb{C}$ of the complex plane, the kth-degree Taylor polynomial of f at $a \in \mathcal{D}$ satisfies

$$f(z) = \sum_{j=0}^{k} \frac{f^{(k)}(a)}{k!}(z-a)^j + R_k(z),$$

where

$$R_k(z) = \frac{(z-a)^{k+1}}{2\pi i} \int_{\Gamma} \frac{f(w)dw}{(w-a)^{k+1}(w-z)}, \tag{2}$$

and Γ is a circle, centred at a, such that $\Gamma \subset \mathcal{D}$. See [1, Chap. 5, Sec. 1.2] for a proof of this result.

The first goal of this paper is to generalize (2) to matrices, thereby providing an explicit expression for the remainder term for the kth-degree Taylor polynomial of a matrix function. Note that it will not be possible to obtain an expression similar to (1) because its derivation relies on the mean value theorem which does not have an exact analogue for matrix-valued functions. Our second goal is to investigate applications of this result in bounding the derivatives and condition numbers of matrix functions via pseudospectra.

Convergence results for Taylor polynomials of matrix functions have been known since the work of Hensel [8], Turnbull [20], and Weyr [21] (see [11, Thm. 4.7] for a more recent

exposition). Mathias [15] also obtains a normwise truncation error bound for matrix function Taylor polynomials which form part of the Schur–Parlett algorithm [4]. There are also a number of remainder theorems within the operator theory literature which can be applied to matrix functions. However, to our knowledge, this paper represents the first time an explicit remainder term (as opposed to a bound) has been specifically obtained for the Taylor polynomial of a matrix function.

The remaining sections of this paper are organized as follows. In section 2 we state and prove the remainder term for the kth-degree Taylor polynomial of a matrix function. In section 3 we investigate some applications of this result by bounding the first order remainder term using pseudospectral techniques and relating it to the condition number of $f(A)$. In section 4 we extend these results to the level-2 condition number of a matrix function, introduced in [13]. In section 5 we examine the behaviour of the pseudospectral bounds on some test problems and show that they can be computed efficiently. Finally in section 6 we present our conclusions and discuss some potential extensions of this work.

2. Remainder term for Taylor polynomials

The Taylor series theorems found in Higham's monograph [11] primarily involve expanding $f(A)$ about a multiple of the identity matrix I:

$$f(A) = \sum_{j=0}^{\infty} \frac{f^{(j)}(\alpha)}{j!}(A - \alpha I)^j.$$

Our starting point is the more general Taylor series expansion in terms of Fréchet derivatives, obtained by Al-Mohy and Higham [2, Thm. 1]. Suppose that f has a power series expansion $\sum_{j=0}^{\infty} a_j x^j$ with radius of convergence $r > 0$ centered at the origin. The interior of the circle $|x| < r$ defines a simply connected open set $\mathcal{D}$. Then, given $A, E \in \mathbb{C}^{n\times n}$ with $\Lambda(A)$, $\Lambda(A+E) \subset \mathcal{D}$ (where $\Lambda(X)$ denotes the spectrum of the matrix X), Al-Mohy and Higham proved that

$$f(A + E) = \sum_{j=0}^{\infty} \frac{1}{j!} D_f^{[j]}(A, E), \tag{3}$$

where

$$D_f^{[j]}(A, E) = \left.\frac{d^j}{dt^j}\right|_{t=0} f(A + tE). \tag{4}$$

They called the $D_f^{[j]}(A, E)$ terms Fréchet derivatives. More precisely, the terms $D_f^{[j]}(A, E)$ are a special case of the jth order Fréchet derivatives described by Higham and Relton [13], in which the perturbations in the j directions are all E. The first of these

terms, $D_f^{[1]}(A,E)$, coincides with the "standard" Fréchet derivative $L_f(A,E)$. Additionally, if A and E commute then we have $D_f^{[j]}(A,E) = E^j f^{(j)}(A)$, where $f^{(j)}$ denotes the jth derivative of the scalar function $f(x)$.

Before writing down the remainder term obtained by truncating the Taylor series in (3), we first recall the standard result that, for any invertible A and B,

$$A^{-1} - B^{-1} = A^{-1}(B-A)B^{-1}. \tag{5}$$

We will also need the following lemma.

Lemma 2.1. *Let $X(t) = A - tB$, where t is a scalar. Then*

$$\frac{d^j}{dt^j}\bigg|_{t=0} X(t)^{-1} = j!\, A^{-1}(BA^{-1})^j.$$

Proof. Note that

$$\frac{d}{dt}X^{-1} = -X^{-1}X'X^{-1},$$

where X' denotes the derivative of X, and that, since higher derivatives of X vanish,

$$\frac{d^j}{dt^j}X^{-1} = (-1)^j j!\, X^{-1}(X'X^{-1})^j.$$

The result then follows by substituting $X = A - tB$ and setting $t = 0$. □

Furthermore, we note that by the Cauchy–Hadamard theorem any power series in the complex plane converging to a function f must converge on a circular domain with radius of convergence r (which can be infinite). In the following results, for the purpose of maximizing generality, we say that f has a power series expansion which converges on a simply connected open set $\mathcal{D}$. Clearly $\mathcal{D}$ must be a subset of this circular domain, but need not be circular itself. The reason for this distinction is that the ϵ-pseudospectra of A, used in section 3, give rise to sets of differing shape.

We now state and prove the main result of this paper, which gives an explicit form of the remainder term when truncating (3).

Theorem 2.2. *Let f have a power series expansion about the origin with radius of convergence r and let $\mathcal{D} \subset \mathbb{C}$ be a simply connected open set within the circle of radius r centered at 0. Let A, $E \in \mathbb{C}^{n\times n}$ be such that $\Lambda(A)$, $\Lambda(A+E) \subset \mathcal{D}$. Then for any $k \in \mathbb{N}$*

$$f(A+E) = T_k(A,E) + R_k(A,E),$$

where

$$T_k(A,E) = \sum_{j=0}^{k} \frac{1}{j!} D_f^{[j]}(A,E), \tag{6}$$

$$R_k(A,E) = \frac{1}{2\pi i} \int_\Gamma f(z)(zI - A - E)^{-1}[E(zI-A)^{-1}]^{k+1} dz, \tag{7}$$

and Γ *is a closed contour in* $\mathcal{D}$ *enclosing* $\Lambda(A)$ *and* $\Lambda(A+E)$.

Proof. The result is proved by induction on k. For the case $k = 0$ we have $f(A+E) = f(A) + R_0(A,E)$. Then

$$\begin{aligned} R_0(A,E) &= f(A+E) - f(A) \\ &= \frac{1}{2\pi i} \int_\Gamma f(z)[(zI-A-E)^{-1} - (zI-A)^{-1}] dz, \end{aligned}$$

using the Cauchy integral definition of a matrix function. It follows from (5) that

$$R_0(A,E) = \frac{1}{2\pi i} \int_\Gamma f(z)(zI-A-E)^{-1} E (zI-A)^{-1} dz.$$

For the inductive step, we assume that $f(A+E) = T_k(A,E) + R_k(A,E)$. The remainder for the $(k+1)$st degree Taylor polynomial is given by

$$\begin{aligned} R_{k+1}(A,E) &= f(A+E) - T_{k+1}(A,E) \\ &= f(A+E) - T_k(A,E) - \frac{1}{(k+1)!} D_f^{[k+1]}(A,E) \\ &= R_k(A,E) - \frac{1}{(k+1)!} \left. \frac{d^{k+1}}{dt^{k+1}} \right|_{t=0} f(A+tE). \end{aligned}$$

Substituting the inductive hypothesis for $R_k(A,E)$ and the Cauchy integral form for $f(A+tE)$, assuming that t is sufficiently small, gives

$$\begin{aligned} R_{k+1}(A,E) = \frac{1}{2\pi i} \int_\Gamma & f(z)(zI-A-E)^{-1}[E(zI-A)^{-1}]^{k+1} dz \\ & - \frac{1}{2\pi i (k+1)!} \frac{d^{k+1}}{dt^{k+1}} \int_\Gamma f(z)(zI-A-tE)^{-1} dz. \end{aligned}$$

By the continuity of f we can apply the Leibniz integral rule to differentiate the integrand in the second term and simplify it using Lemma 2.1. We obtain

$$\begin{aligned} R_{k+1}(A,E) &= \frac{1}{2\pi i}\int_\Gamma f(z)\big[(zI-A-E)^{-1}[E(zI-A)^{-1}]^{k+1} \\ &\quad -(zI-A)^{-1}[E(zI-A)^{-1}]^{k+1}\big]dz \\ &= \frac{1}{2\pi i}\int_\Gamma f(z)(zI-A-E)^{-1}[E(zI-A)^{-1}]^{k+2}dz, \end{aligned}$$

where (5) has been used once more. This completes the proof. □

We end this section by briefly describing how Theorem 2.2 also allows us to obtain a remainder term for Padé approximants (this was first done in the scalar case by Elliot [5]).

Suppose that we approximate $f(z)$ using a rational function $p_m(z)/q_n(z)$, where $p_m(z)$ and $q_n(z)$ are polynomials of degree m and n respectively. The Padé approximant is the unique choice (up to scalar multiples) of $p_m(z)$ and $q_n(z)$ such that $f(z)-p_m(z)/q_n(z) = O(z^{m+n+1})$. Therefore, using the same rational function to approximate the corresponding matrix function, we have $q_n(X)f(X)-p_m(X) = O(\|X\|^{m+n+1})$. We introduce the truncation error term $S_{m,n}(X)$ to the Padé approximant such that

$$f(X) = q_n(X)^{-1}p_m(X) + S_{m,n}(X).$$

Then, by rearranging the above,

$$q_n(X)f(X) = p_m(X) + q_n(X)S_{m,n}(X).$$

The term $q_n(X)S_{m,n}(X)$ is then the remainder term if we consider $p_m(X)$ to be a power series expansion of $q_n(X)f(X)$ when we set $A = 0$ and $E = X$. The remainder has degree at least $m+n$ and so, by applying (7) with $k = m+n$, we obtain

$$S_{m,n}(X) = \frac{q_n(X)^{-1}X^{m+n+1}}{2\pi i}\int_\Gamma \frac{q_n(z)f(z)(zI-X)^{-1}}{z^{m+n+1}}\,dz,$$

where the closed contour Γ encloses $\Lambda(X)$ and the origin.

3. Application to condition numbers and pseudospectra

In this section we use Theorem 2.2 to study the behaviour of the condition number of a matrix function, which measures the sensitivity of $f(A)$ to small perturbations in A. The results in this section are applicable for any induced matrix norm. Our approach requires borrowing a number of techniques from the analysis of pseudospectra. Recall that the ϵ-pseudospectrum of a matrix X is the set

$$\Lambda_\epsilon(X) = \left\{z \in \mathbb{C} : \|(zI-X)^{-1}\| \geq \epsilon^{-1}\right\}. \tag{8}$$

To begin, the following lemma provides some pseudospectral bounds on the size of the remainder terms.

Lemma 3.1. *Let f and $\mathcal{D}$ satisfy the criteria of Theorem 2.2. Furthermore let $\epsilon > 0$ be such that $\Lambda_\epsilon(A) \subset \mathcal{D}$ and $\Lambda_\epsilon(A+E) \subset \mathcal{D}$, and take $\tilde{\Gamma}_\epsilon \subset \mathcal{D}$ to be a closed contour that encloses both $\Lambda_\epsilon(A)$ and $\Lambda_\epsilon(A+E)$. Then the remainder term $R_k(A,E)$ is bounded by*

$$\|R_k(A,E)\| \leq \frac{\|E\|^{k+1}\tilde{L}_\epsilon}{2\pi\epsilon^{k+2}} \max_{z\in\tilde{\Gamma}_\epsilon} |f(z)|, \tag{9}$$

where $\tilde{L}_\epsilon$ is the length of $\tilde{\Gamma}_\epsilon$. In particular, when a circular contour centered at 0 is used,

$$\|R_k(A,E)\| \leq \frac{\|E\|^{k+1}\tilde{\rho}_\epsilon}{\epsilon^{k+2}} \max_{\theta\in[0,2\pi]} |f(\tilde{\rho}_\epsilon e^{i\theta})|, \tag{10}$$

where $\tilde{\rho}_\epsilon = \max\{|z| : z \in \Lambda_\epsilon(A+E) \cap \Lambda_\epsilon(A)\}$ is the radius of the circle.

(Note that tildes on $\tilde{L}_\epsilon$, $\tilde{\Gamma}_\epsilon$, and $\tilde{\rho}_\epsilon$ are used because, for this result only, the contour needs to enclose $\Lambda_\epsilon(A+E)$ in addition to $\Lambda_\epsilon(A)$. For subsequent results, the contour need only enclose $\Lambda_\epsilon(A)$ and the tildes are dropped.)

Proof. The proof is analogous to that of the bound

$$\|f(A)\| \leq \frac{\tilde{L}_\epsilon}{2\pi\epsilon} \max_{z\in\tilde{\Gamma}_\epsilon} |f(z)|,$$

obtained by Trefethen and Embree [19, Ch. 14]. We bound the norm of $R_k(A,E)$ by noting that

$$\|R_k(A,E)\| \leq \frac{\|E\|^{k+1}}{2\pi} \int_{\tilde{\Gamma}_\epsilon} |f(z)| \|(zI-A-E)^{-1}\| \|(zI-A)^{-1}\|^{k+1}.$$

On $\tilde{\Gamma}_\epsilon$ we have $\|(zI-A-E)^{-1}\| \leq \epsilon^{-1}$ and $\|(zI-A)^{-1}\| \leq \epsilon^{-1}$. The first part of the lemma follows immediately. For the second part, take $\tilde{\Gamma}_\epsilon$ to be a circle with center 0 and radius $\tilde{\rho}_\epsilon = \max\{|z| : z \in \Lambda_\epsilon(A+E) \cap \Lambda_\epsilon(A)\}$. □

We can also use this result to bound the absolute condition number of a matrix function. Recall that the absolute condition number measures the first order sensitivity of $f(A)$ to small perturbations in A and is given by [11, Chap. 3]

$$\begin{aligned} \mathrm{cond}_{\mathrm{abs}}(f,A) &:= \lim_{\tau\to 0} \sup_{\|E\|\leq\tau} \frac{\|f(A+E)-f(A)\|}{\tau} \\ &= \max_{\|E\|\leq 1} \|L_f(A,E)\|. \end{aligned} \tag{11}$$

Lemma 3.1 provides us with the following bound on the absolute condition number.

Corollary 3.2. *Let f and $\mathcal{D}$ satisfy the criteria of Theorem 2.2. Let $\epsilon > 0$ be such that $\Lambda_\epsilon(A) \subset \mathcal{D}$, and let $\Gamma_\epsilon \subset \mathcal{D}$ be a closed contour of length L_ϵ that encloses the ϵ-pseudospectrum. Then*

$$\mathrm{cond}_{\mathrm{abs}}(f, A) \leq \frac{L_\epsilon}{2\pi\epsilon^2} \max_{z\in\Gamma_\epsilon} |f(z)|. \tag{12}$$

In particular, when a circular contour centered at 0 is used,

$$\mathrm{cond}_{\mathrm{abs}}(f, A) \leq \frac{\rho_\epsilon}{\epsilon^2} \max_{\theta\in[0,2\pi]} |f(\rho_\epsilon e^{i\theta})|, \tag{13}$$

where $\rho_\epsilon = \max\{|z| : z \in \Lambda_\epsilon(A)\}$ is the pseudospectral radius of A.

Proof. Set $k = 0$ in (9). Consider $\|E\| = \alpha < \epsilon$ so that, by an equivalent definition of the ϵ-pseudospectrum, we have $\Lambda(A + E) \subset \Lambda_\epsilon(A)$. Then, since $R_0(A, E) = L_f(A, E) + o(\|E\|)$, we have

$$\|L_f(A, E) + o(\alpha)\| \leq \frac{\alpha L_\epsilon}{2\pi\epsilon^2} \max_{z\in\Gamma_\epsilon} |f(z)|.$$

We divide by α and take the supremum over all E such that $\|E\| \leq \alpha$ to obtain

$$\sup_{\|E\|\leq\alpha} \|L_f(A, E/\alpha) + o(\alpha)/\alpha\| \leq \frac{L_\epsilon}{2\pi\epsilon^2} \max_{z\in\Gamma_\epsilon} |f(z)|.$$

The proof of (12) is completed by taking the limit $\alpha \to 0$ and recalling that the absolute condition number of a matrix function is given by the operator norm of the Fréchet derivative (11).

The proof of (13) is essentially the same, except that (10) is taken as the starting point rather than (9).

Note that an alternative proof of the corollary can be obtained by starting with the integral representation of the Fréchet derivative

$$L_f(A, E) = \frac{1}{2\pi i} \int_{\Gamma_\epsilon} f(z)(zI - A)^{-1} E (zI - A)^{-1} dz,$$

and bounding it above using the techniques from the proof of Lemma 3.1. □

Assuming that these bounds can be computed efficiently they are of considerable interest as most existing results regarding the estimation of the condition number provide only lower bounds [11, Chap. 3]. Indeed, this is particularly interesting when combined with a bound on the size of the ϵ-pseudospectrum given by the following result.

Lemma 3.3 *(Reddy, Schmid, and Henningson). Let $W(A)$ be the numerical range of A and Δ_δ be a closed disk of radius δ. Then for all $\epsilon > 0$*

$$\Lambda_\epsilon(A) \subset W(A) + \Delta_\epsilon,$$

where set addition is defined componentwise; that is $S_1+S_2 = \{s_1+s_2 : s_1 \in S_1,\ s_2 \in S_2\}$.

Proof. See Reddy, Schmid, and Henningson [16, Thm. 2.1]. □

Since the numerical radius, $r(A) := \sup_{z\in W(A)} |z|$, is equal to $\|A\|_2$ we know that the ϵ-pseudospectral radius is no larger than $\|A\|_2 + \epsilon$. Thus we obtain the following corollary.

Corollary 3.4. *Let f, $\mathcal{D}$, and $\epsilon > 0$ satisfy the criteria of Corollary 3.2 and suppose that $\|A\|_2 + \epsilon < r$, the radius of convergence for the power series expansion of f. Then*

$$\mathrm{cond}_{\mathrm{abs}}(f, A) \leq \frac{\|A\|_2 + \epsilon}{\epsilon^2} \max_{|z|=\|A\|_2+\epsilon} |f(z)|. \tag{14}$$

Proof. The circle of radius $\|A\|_2 + \epsilon$ around the origin encloses $\Lambda_\epsilon(A)$ and is of length $2\pi(\|A\|_2 + \epsilon)$. Using this contour in (13) gives the desired result. □

One potential application of this result is in the design and analysis of algorithms for computing matrix functions. Many such algorithms work by rescaling A to be of small norm, applying the function to this scaled matrix (via a Padé approximant or Taylor series), and then undoing the effect of the scaling. This corollary may allow us to better understand the numerical effect of applying the matrix function to the scaled matrix, since such analysis is typically done only in exact arithmetic.

We end this section by briefly mentioning a related theorem due to Lui [14, Thm. 3.1], concerning the relationship between the pseudospectra of A and $f(A)$. The theorem is restated here in our notation. Recall that $R_k(A, E)$ was defined in Theorem 2.2 and that $R_0(A, E) = L_f(A, E) + o(\|E\|)$.

Lemma 3.5 *(Lui). Let ϵ, f, and Γ_ϵ satisfy the conditions of Corollary 3.2. Furthermore let $f(\Lambda_\epsilon(A)) = \{f(z) : z \in \Lambda_\epsilon(A)\}$ and $M = \max_{\|E\|\leq\epsilon} \|R_0(A, E)\|$. Then $f(\Lambda_\epsilon(A)) \subset \Lambda_M(f(A))$.*

Proof. If z is an eigenvalue of $A + E$ with $\|E\| \leq \epsilon$ (so that $z \in \Lambda_\epsilon(A)$), then $f(z)$ is an eigenvalue of $f(A + E) = f(A) + R_0(A, E)$ and $\|R_0(A, E)\| \leq M$. □

This result shows that, to first order in ϵ, the ϵ-pseudospectrum of A is related to the δ-pseudospectrum of $f(A)$ via $f(\Lambda_\epsilon(A)) \subset \Lambda_\delta(f(A))$, where $\delta = \mathrm{cond}_{\mathrm{abs}}(f, A)\epsilon$.

4. Application to higher order condition numbers

Higham and Relton [13] introduce the level-q condition number for matrix functions, which is defined recursively by

$$\operatorname{cond}_{\mathrm{abs}}^{(q)}(f,A) := \lim_{\alpha\to 0} \sup_{\|Z\|\leq\alpha} \frac{|\operatorname{cond}_{\mathrm{abs}}^{(q-1)}(f,A+Z) - \operatorname{cond}_{\mathrm{abs}}^{(q-1)}(f,A)|}{\alpha}, \tag{15}$$

where $\operatorname{cond}_{\mathrm{abs}}^{(1)}(f,A) := \operatorname{cond}_{\mathrm{abs}}(f,A)$. In section 3 we focused on the first order remainder term, $R_0(A,E)$, and results concerning the condition number $\operatorname{cond}_{\mathrm{abs}}(f,A)$ but – by choosing $k>0$ in Lemma 3.1 – we can attempt to extend results such as Corollary 3.2 to these higher order condition numbers.

Before proceeding, we must first investigate the relationship between the $D_f^{[j]}(A,E)$ defined in (4) and higher order Fréchet derivatives. Recall that $D_f^{[j]}(A,E)$ is a special case of the jth order Fréchet derivative in which the perturbation in each direction is E. In [13] a definition of the jth order Fréchet derivative, assuming it is continuous in A, is given in terms of the mixed partial derivative:

$$L_f^{(j)}(A,E_1,\dots,E_j) = \frac{\partial}{\partial s_1}\cdots\frac{\partial}{\partial s_j}\bigg|_{(s_1,\dots,s_j)=0} f(A+s_1E_1+\cdots+s_jE_j). \tag{16}$$

The following theorem expresses this jth order Fréchet derivative in terms of a contour integral.

Theorem 4.1. *Let f be j times Fréchet differentiable such that the jth Fréchet derivative is continuous at A, and let Γ be a closed contour enclosing $\Lambda(A)$ such that f is analytic inside and on Γ. Then, the jth order Fréchet derivative of a matrix function $f(A)$ in the directions $E_1,\dots,E_j$ is given by*

$$L_f^{(j)}(A,E_1,\dots,E_j) = \frac{1}{2\pi i}\int_\Gamma f(z)(zI-A)^{-1}\sum_{\sigma\in\mathcal{S}_j}\prod_{i=1}^{k} E_{\sigma(i)}(zI-A)^{-1}dz, \tag{17}$$

where $\mathcal{S}_j$ is the set of permutations of $\{1,2,\dots,k\}$. In particular the derivative $D_f^{[j]}(A,E)$ is given by

$$D_f^{[j]}(A,E) = \frac{j!}{2\pi i}\int_\Gamma f(z)(zI-A)^{-1}[E(zI-A)^{-1}]^{j+1}dz. \tag{18}$$

Proof. For any choice of s_i (in some neighbourhood of 0) and E_i, we can write $f(A+s_1E_1+\cdots+s_jE_j)$ as a Cauchy integral by using the standard Cauchy integral definition of a matrix function and choosing a contour $\tilde{\Gamma}$ that encloses Γ and some neighbourhood of $\Lambda(A)$. Then (16) becomes

$$L_f^{(j)}(A, E_1, \ldots, E_j) = \\ \frac{\partial}{\partial s_1} \cdots \frac{\partial}{\partial s_j}\bigg|_{(s_1,\ldots,s_j)=0} \int_{\tilde{\Gamma}} f(z)(zI - (A + s_1E_1 + \cdots + s_jE_j))^{-1} dz.$$

Using the Leibniz integral rule, the differential operator

$$\frac{\partial}{\partial s_1} \cdots \frac{\partial}{\partial s_j}\bigg|_{(s_1,\ldots,s_j)=0}$$

can be brought inside the integral sign. The required integrand is then obtained by using the identity

$$\frac{d}{dx}U^{-1} = -U^{-1}\frac{dU}{dx}U^{-1}.$$

The result (17) follows by then restricting the contour to the closed curve Γ containing $\Lambda(A)$. The second part of the theorem, (18), follows by setting $E_1 = \cdots = E_j$. □

Theorem 4.1 shows that, to first order, the kth remainder term in the Taylor series is simply the $(k+1)$st derivative, as we might expect. Specifically, comparing (18) with (7) we find

$$R_k(A, E) = \frac{1}{(k+1)!}D_f^{[k+1]}(A, E) + o(\|E\|^{k+2}).$$

In addition, Theorem 4.1 allows us to prove the following theorem, which uses the pseudospectrum of A to bound the norm of the jth order Fréchet derivative.

Theorem 4.2. *Let f satisfy the criteria of Theorem 4.1 and let Γ_ϵ be a closed contour enclosing $\Lambda_\epsilon(A)$ such that f is analytic inside and on Γ_ϵ. Then the jth order Fréchet derivative can be bounded by*

$$\|L_f^{(j)}(A, E_1, \ldots, E_j)\| \leq \frac{j!\,L_\epsilon}{2\pi\epsilon^{j+1}}\left(\max_{z\in\Gamma_\epsilon}|f(z)|\right)\prod_{i=1}^{j}\|E_i\|, \tag{19}$$

where L_ϵ is the length of Γ_ϵ.

Proof. In (17) use the contour Γ_ϵ, take norms, and note that $\|(zI - A)^{-1}\| \leq \epsilon^{-1}$ on Γ_ϵ. □

It would be desirable to obtain a bound on the level-q condition number, by first bounding it in terms of the norm of the qth Fréchet derivative and then applying Theorem 4.2. However, in the general case such bounds prove to be far too weak to be of

any interest. Instead we restrict ourselves to the case $q = 2$ and the level-2 condition number.

Lemma 4.3. *Let f satisfy the criteria of Theorem 4.1 and let Γ_ϵ a closed contour enclosing $\Lambda_\epsilon(A)$ such that f is analytic inside and on Γ. The level-2 condition number is bounded by*

$$\operatorname{cond}^{(2)}_{\text{abs}}(f, A) \leq \frac{L_\epsilon}{\pi\epsilon^3} \max_{z\in\Gamma_\epsilon} |f(z)|.$$

When a circular contour centered at 0 is used,

$$\operatorname{cond}^{(2)}_{\text{abs}}(f, A) \leq \frac{2\rho_\epsilon}{\epsilon^3} \max_{\theta\in[0,2\pi]} |f(\rho_\epsilon e^{i\theta})|,$$

where ρ_ϵ, the pseudospectral radius, is the radius of the circle.

Proof. Higham and Relton [13, Sec. 5] give an upper bound for the level-2 absolute condition number in terms of the norm of the second Fréchet derivative

$$\operatorname{cond}^{(2)}_{\text{abs}}(f, A) \leq \max_{\|E_1\|=1} \max_{\|E_2\|=1} \|L^{(2)}_f(A, E_1, E_2)\|. \tag{20}$$

Substituting the bound from (19) into (20) gives the required results. □

5. Numerical experiments

In this section we show how our pseudospectral bounds on the condition number, (12) and (13), can be used to estimate the condition number of matrix functions in practice. We also find that they are cheaper than alternative approaches and, therefore, one might use the pseudospectral bound as a quick estimate of the condition number. If this estimate is unsatisfactorily large we can use existing methods to estimate it more accurately. The term "unsatisfactorily large" can be made precise in the following manner: many applications only require the first few digits of the result to be correct so that a relative error of, for example, 1e-4 is perfectly acceptable. When using a backward stable algorithm the relative error is approximately bounded above by the condition number multiplied by the unit roundoff ($u = 2^{-53}$ in IEEE double precision arithmetic).

Throughout this section, to compute our bound on the condition number, we will be using (12)

$$\operatorname{cond}_{\text{abs}}(f, A) \leq \frac{L_\epsilon}{2\pi\epsilon^2} \max_{z\in\Gamma_\epsilon} |f(z)|,$$

where Γ_ϵ is a closed contour of length L_ϵ that encloses the pseudospectrum of A and lies within the region where f has a convergent power series. Recall also that the relative condition number, $\operatorname{cond}_{\text{rel}}(f, A)$, is given by

$$\mathrm{cond}_{\mathrm{rel}}(f,A) = \mathrm{cond}_{\mathrm{abs}}(f,A)\frac{\|A\|}{\|f(A)\|}.$$

Combining these two results allows us to bound the relative condition number from above. This bound will be cheap to compute provided that the cost of computing L_ϵ and $\max_{z\in\Gamma_\epsilon}|f(z)|$ is sufficiently small.

In order to use this bound in practice we must choose which matrix norm to consider, the value of ϵ, and the contour Γ_ϵ. We will use the Frobenius norm since, in this norm, there is an explicit formula for the condition number which can be computed using [11, Alg. 3.17]. However, the pseudospectrum is not defined in the Frobenius norm, since it requires the use of an induced norm. To resolve this, one can easily show that the absolute condition number in the Frobenius norm is bounded above by $\sqrt{n}$ times the condition number in the 2-norm, where n is the size of the matrix. Hence we have

$$\mathrm{cond}_{\mathrm{rel}}(f,A,\|\cdot\|_F) \leq \sqrt{n}\,\mathrm{cond}_{\mathrm{abs}}(f,A,\|\cdot\|_2)\frac{\|A\|_F}{\|f(A)\|_F}.$$

The right-hand side of this equation is what we will compute, where the $\mathrm{cond}_{\mathrm{abs}}(\cdot)$ term is bounded above by (12).

It remains to choose ϵ and Γ_ϵ. Looking at (12) we see that, heuristically, in order to minimize the upper bound we would like ϵ to be reasonably far from 0. Some of our test functions will have power series that are convergent in a circle of radius 1 around the point $z = 0$; for these cases we choose Γ_ϵ to be a circle centered at 0 with radius 0.99 and find the largest ϵ such that the ϵ-pseudospectral radius lies inside this circle. This is computed using the nonlinear optimization routine `fminbnd` in MATLAB. When our function has a power series with an infinite radius of convergence we choose $\epsilon = 1$ and take Γ_ϵ to be a circle centered at the mean of the eigenvalues of A ($\gamma = \frac{1}{n}\sum\lambda_i$) with radius equal to the ϵ-pseudospectral radius of $A-\gamma I$. Finally, to find $\max|f(z)|$ on these contours, we again use the nonlinear optimization routine `fminbnd` in MATLAB. We use `psapsr` by Guglielmi and Overton [7] to compute the ϵ-pseudospectral radii throughout.

We will compare our pseudospectral method described above (hereafter referred to as `condpseudo`) in the Frobenius norm against two alternative methods for computing the condition number: `funm_condest_fro` from the Matrix Function Toolbox [10] and an "exact" method detailed by Higham [11, Alg. 3.17], which we refer to as `condold` and `condexact`, respectively.

The method `condold` uses finite difference approximations to the derivatives of the matrix function and has $O(n^3)$ cost. Meanwhile `condexact` expresses the condition number as the 2-norm of a matrix $K_f(A) \in \mathbb{C}^{n^2\times n^2}$, called the Kronecker form of the matrix function, which must be computed explicitly with cost $O(n^5)$. Therefore `condexact` is impractical for all but the smallest problems.

Our first experiment compares `condpseudo` and `condold` to `condexact` in terms of accuracy and reliability on a range of matrix functions. The purpose of this experiment is to confirm that `condpseudo` does indeed return an upper bound on the condition number and that this upper bound is not much larger than the exact value.

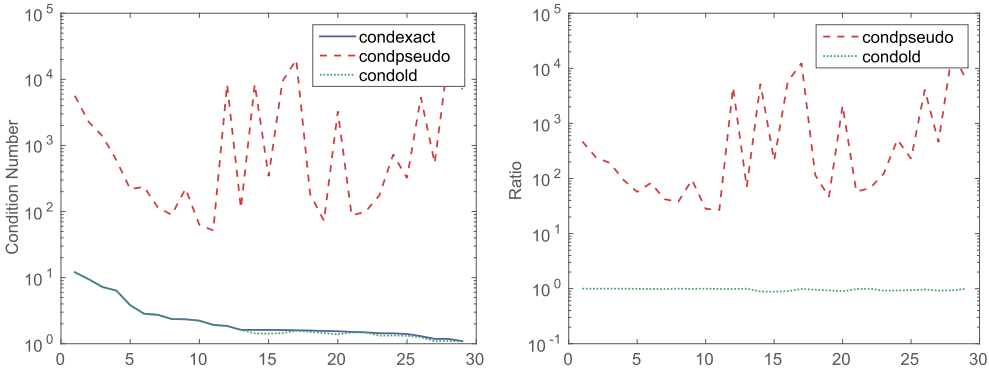

Fig. 1. Condition number estimates/bounds for the matrix function corresponding to $f(x) = \log(1 + x)$ in the Frobenius norm over 29 test matrices. We have `condold` and `condexact` overlapping almost entirely. *Left:* The condition number estimates/bounds. *Right:* The ratios of `condpseudo` and `condold` to `condexact`.

We compare the three different algorithms on four matrix functions corresponding to the scalar functions $\log(1+x)$, $(1+x)^{1/15}$, $\exp(x)$, and $\cos(x)$. The first two of these have a power series representation which is convergent for $|x| < 1$, whilst the latter have globally convergent power series. The matrix functions are computed using `logm` and `expm` in MATLAB, along with `cosm` from [3], and `powerm_fre_new` by Higham and Lin [12].

For each function we use 29 test matrices (of size $n = 10$) from the Matrix Computation Toolbox [9] and plot both the computed condition numbers and the ratio of `condpseudo` and `condold` to `condexact`. For the first two functions, where we need all eigenvalues to lie within the region of convergence, we transform each matrix to have eigenvalues centered at 0 with $\|A\|_2 = 1$ so that all eigenvalues lie within the unit disk.

In Fig. 1 we see the condition number as computed by the three methods, for each of the 29 test matrices, using the function $f(x) = \log(1 + x)$. The results are ordered by decreasing condition number as computed by `condexact`. We can immediately see that `condpseudo` is indeed an upper bound and is usually 2–4 orders of magnitude larger than the exact condition number. Meanwhile `condold` is generally a very good estimate of the condition number.

Next in Fig. 2 we compare the condition numbers when using the function $f(x) = (1 + x)^{1/15}$. In this case we see very similar behaviour to the previous function: `condold` and `condexact` are almost identical whilst `condpseudo` provides an upper bound that is generally 2–4 orders of magnitude larger than the true condition number.

Fig. 3 shows the results when using $f(x) = \exp(x)$. In this case `condpseudo` performs slightly better than previously being only 1–3 orders of magnitude above `condexact` on most test problems.

Finally, Fig. 4 displays the results for $f(x) = \cos(x)$. In this case, as for the exponential we see that `condpseudo` is a reliable upper bound, generally being 1–3 orders of magnitude above `condexact` except for one case on the left-hand side in which it is more than 10 orders of magnitude larger. This is due to the matrix in question having eigen-

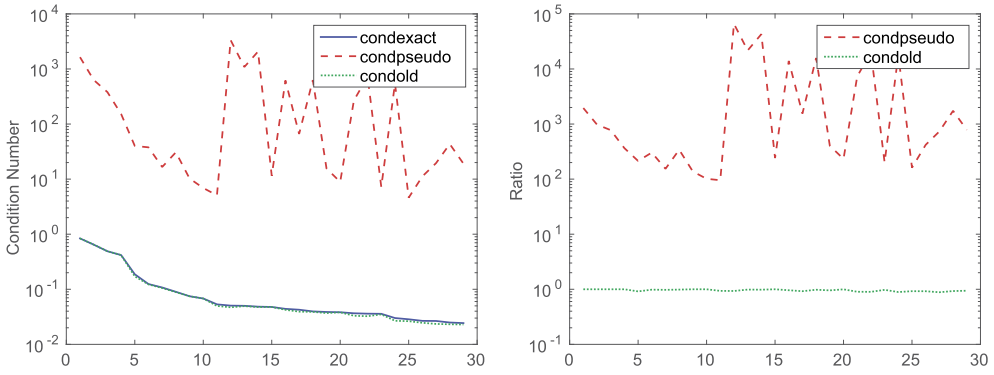

Fig. 2. Condition number estimates/bounds for the matrix function corresponding to $f(x) = (1+x)^{1/15}$ in the Frobenius norm over 29 test matrices. We have `condold` and `condexact` overlapping almost entirely. *Left:* The condition number estimates/bounds. *Right:* The ratios of `condpseudo` and `condold` to `condexact`.

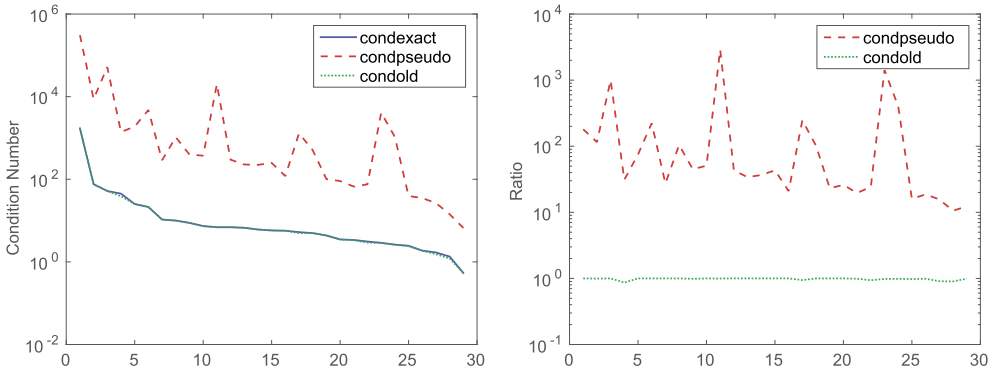

Fig. 3. Condition number estimates/bounds for the matrix function corresponding to $f(x) = \exp(x)$ in the Frobenius norm over 29 test matrices. We have `condold` and `condexact` overlapping almost entirely. *Left:* The condition number estimates/bounds. *Right:* The ratios of `condpseudo` and `condold` to `condexact`.

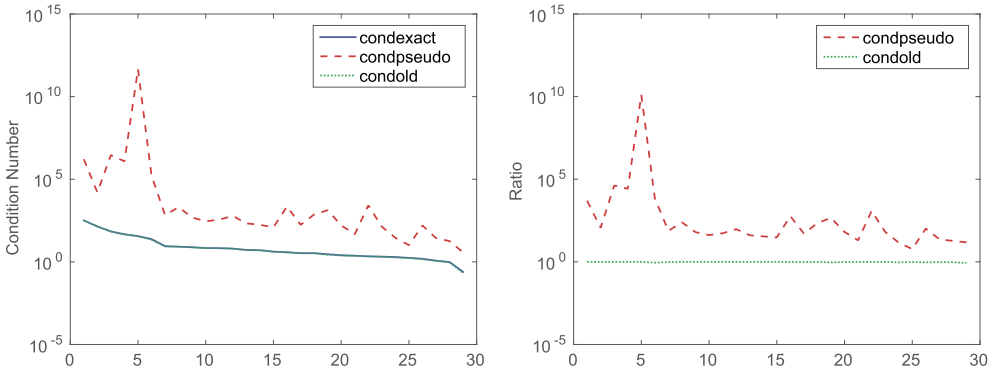

Fig. 4. Condition number estimates/bounds for the matrix function corresponding to $f(x) = \cos(x)$ in the Frobenius norm over 29 test matrices. We have `condold` and `condexact` overlapping almost entirely. *Left:* The condition number estimates/bounds. *Right:* The ratios of `condpseudo` and `condold` to `condexact`.

values extending far into the complex plane: as the cosine function grows exponentially in the direction of the imaginary axis $|\cos(z)|$ is extremely large on the chosen contour.

Each of these four cases shows that `condpseudo` provides a reliable upper bound on the condition number and is generally just a few orders of magnitude above the true value. We also note that, since `condpseudo` needs only the scalar function $f(x)$ and does not need to compute the derivatives of a matrix function, via finite differences or otherwise, it can easily be applied to very complicated matrix functions such as $\cos(\sqrt{A})$ (for which no specially designed algorithms exist) with no modification. In this particular case we would apply our algorithm to the function

$$f(A) = \sum_{k=0}^{\infty} \frac{(-1)^k A^k}{(2k!)},$$

which is analytic and equivalent to $\cos(\sqrt{A})$ away from the origin, regardless of which branch of the square root function is selected. Matrix functions such as this can arise in finite element semidiscretization of the wave equation. For example, the second order differential equation

$$y''(t) + Ay(t) = g(t), \quad y(0) = y_0, \quad y'(0) = y_0',$$

has the solution

$$y(t) = \cos(\sqrt{A}t)y_0 + (\sqrt{A})^{-1}\sin(\sqrt{A}t)y_0' + \int_0^t (\sqrt{A})^{-1}\sin\big(\sqrt{A}(t-s)\big)g(s)\,ds,$$

where $\sqrt{A}$ denotes any square root of A [6, p. 124], [18]; see also [11, Prob. 4.1] for the case $g(t) = 0$.

Our next experiment compares the speed of estimating the condition number as the size of the matrix grows. Here we focus on the function $f(x) = (1+x)^t$ for $t = 1/5, 1/15, 1/52$ and for n between 10 and 1000. For each value of n we take A to be a matrix with elements normally distributed with zero mean and unit variance, scaled to have unit norm. Since `condexact` is an $O(n^5)$ algorithm it becomes increasing impractical as n grows: instead we will compare `condpseudo` against `condold` and a different algorithm, `condhili`. This latter algorithm, designed by Higham and Lin [12], estimates the condition number of matrix powers in a similar manner to `condold` but actually computes the derivatives of the matrix function, as opposed to using finite difference approximations. The algorithm is designed to estimate the condition number in the 1-norm but has similar computational complexity to `condold` which works in the Frobenius norm. This experiment was run on a laptop with an Intel dual-core i7 processor using MATLAB R2014b.

Fig. 5 shows the results of this experiment. On the left are the runtimes using each of the 3 algorithms to compute the condition number of $(I+A)^{1/t}$ for the various values

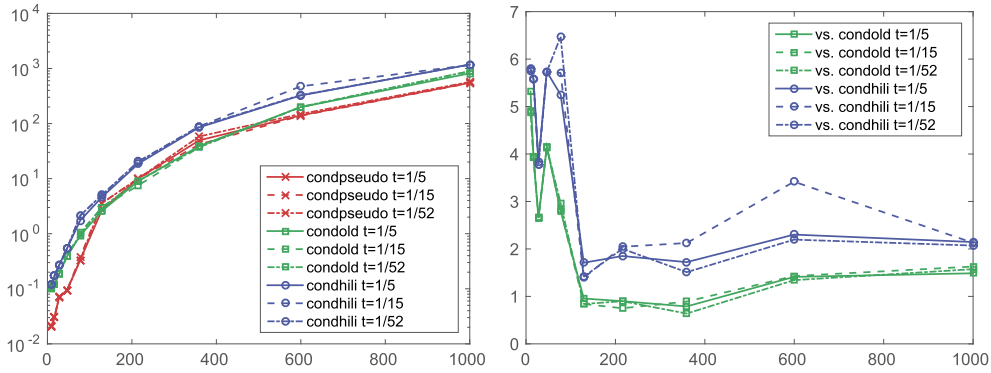

Fig. 5. Runtime in seconds and resulting speedup when computing the matrix function corresponding to $f(x) = (1+x)^t$ for $t = 1/5,\ 1/15,\ 1/52$ using `condpseudo`, `condold`, and `condhili` as n varies between 10 and 1000. The x-axis shows n, the size of the test matrix, whilst the y-axis denotes the runtime and speedup, respectively. *Left:* Runtime in seconds when running each algorithm. *Right:* Speedup when using `condpseudo` compared to `condold` and `condhili`.

of t whilst the right-hand plot shows the speedup when using `condpseudo` relative to the other methods. The x-axis shows n, the size of the matrices, whilst the y-axis shows the runtime in seconds (left-hand plot) and the speedup obtained (right-hand plot). We see that `condpseudo` is much cheaper than the alternatives for fairly small matrices and appears to settle at around 1.5 times faster than `condold` and 2 times faster than `condhili`, respectively, on this machine. This would suggest that using `condpseudo` is beneficial for applications where low-accuracy solutions are required and is particularly good in situations where lots of small matrix functions need to be computed.

6. Conclusions

The main results in this paper are as follows. We have obtained an explicit expression for the remainder term of a matrix function Taylor polynomial (Theorem 2.2). Combining this with use of the ϵ-pseudospectrum of A leads to upper bounds on the condition numbers of $f(A)$. Our numerical experiments demonstrated that our bounds can be used for practical computations: they provide a cheap upper bound on the condition number which is often only a few orders of magnitude too large. This means that our bounds could be used as a quick estimate of the condition number and if this estimate is too large, for instance if the estimate suggests that an insufficient number of correct significant figures might be obtained in computing $f(A)$, then existing methods can be used to obtain the condition number more accurately.

Another benefit of our approach is that it can easily be applied to bound the condition number of complicated matrix functions such as $\cos(\sqrt{A})$ without modification, as there are currently no specialized methods for computing such quantities.

Our results may also have further useful applications in the development of matrix function algorithms by allowing us to estimate the size of the remainder terms for Padé approximants, for example. We may also be able to glean further insight into

the behaviour of existing algorithms to compute matrix functions (see the discussion of Corollary 3.4). This will be the subject of future work.

Acknowledgements

We are very grateful to Nick Higham for numerous helpful comments on earlier drafts of this work and to the referee who helped to significantly increase the quality of the manuscript.

References

[1] Lars V. Ahlfors, Complex Analysis, third edition, McGraw-Hill, New York, ISBN 978-0-0700-0657-7, 1979.
[2] Awad H. Al-Mohy, Nicholas J. Higham, The complex step approximation to the Fréchet derivative of a matrix function, Numer. Algorithms 53 (1) (2010) 133–148, http://dx.doi.org/10.1007/s11075-009-9323-y.
[3] Awad H. Al-Mohy, Nicholas J. Higham, Samuel D. Relton, New algorithms for computing the matrix sine and cosine separately or simultaneously, SIAM J. Sci. Comput. 37 (1) (2015) A456–A487, http://dx.doi.org/10.1137/140973979.
[4] Philip I. Davies, Nicholas J. Higham, A Schur–Parlett algorithm for computing matrix functions, SIAM J. Matrix Anal. Appl. 25 (2) (2003) 464–485, http://dx.doi.org/10.1137/S0895479802410815.
[5] David Elliott, Truncation errors in Padé approximations to certain functions: an alternative approach, Math. Comp. (ISSN 0025-5718) 21 (99) (1967) 398–406.
[6] F.R. Gantmacher, The Theory of Matrices, vol. 1, Chelsea, New York, ISBN 0-8284-0131-4, 1959.
[7] Nicola Guglielmi, Michael L. Overton, Fast algorithms for the approximation of the pseudospectral radius of a matrix, SIAM J. Matrix Anal. Appl. 32 (4) (2011) 1166–1192.
[8] Kurt Hensel, Über Potenzreihen von Matrizen, J. Reine Angew. Math. 155 (42) (1926) 100–110.
[9] Nicholas J. Higham, The Matrix Computation Toolbox, http://www.maths.manchester.ac.uk/~higham/mctoolbox.
[10] Nicholas J. Higham, The Matrix Function Toolbox, http://www.maths.manchester.ac.uk/~higham/mftoolbox.
[11] Nicholas J. Higham, Functions of Matrices: Theory and Computation, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, ISBN 978-0-898716-46-7, 2008.
[12] Nicholas J. Higham, Lijing Lin: An improved Schur–Padé algorithm for fractional powers of a matrix and their Fréchet derivatives, SIAM J. Matrix Anal. Appl. 34 (3) (2013) 1341–1360, http://dx.doi.org/10.1137/130906118.
[13] Nicholas J. Higham, Samuel D. Relton, Higher order Fréchet derivatives of matrix functions and the level-2 condition number, SIAM J. Matrix Anal. Appl. 35 (3) (2014) 1019–1037, http://dx.doi.org/10.1137/130945259.
[14] S.-H. Lui, A pseudospectral mapping theorem, Math. Comp. 72 (244) (2003) 1841–1854.
[15] Roy Mathias, Approximation of matrix-valued functions, SIAM J. Matrix Anal. Appl. 14 (4) (1993) 1061–1063.
[16] S.C. Reddy, P.J. Schmid, D.S. Henningson, Pseudospectra of the Orr–Sommerfeld operator, SIAM J. Appl. Math. 53 (1993) 15–47, http://dx.doi.org/10.1137/0153002.
[17] Walter Rudin, Real and Complex Analysis, third edition, McGraw-Hill, New York, ISBN 0070542341, 1986.
[18] Steven M. Serbin, Rational approximations of trigonometric matrices with application to second-order systems of differential equations, Appl. Math. Comput. 5 (1) (1979) 75–92, http://dx.doi.org/10.1016/0096-3003(79)90011-0.
[19] Lloyd Nicholas Trefethen, Mark Embree, Spectra and Pseudospectra: The Behavior of Nonnormal Matrices and Operators, Princeton University Press, 2005.
[20] H.W. Turnbull, A matrix form of Taylor's theorem, Proc. Edinb. Math. Soc. (2) 2 (1930) 33–54, http://dx.doi.org/10.1017/S0013091500007537.
[21] Edouard Weyr, Note sur la théorie de quantités complexes formées avec n unités principales, Bull. Sci. Math. II 11 (1887) 205–215.