A CHAIN RULE FOR MATRIX FUNCTIONS AND APPLICATIONS

ROY MATHIAS*

July 22, 1997

Abstract. Let f be a not necessarily analytic function and let $A(t)$ be a family of $n \times n$ matrices depending on the parameter t. Conditions for the existence of the first and higher derivatives of $f(A(t))$ are presented together with formulae that represent these derivatives as a submatrix of $f(B)$ where B is a larger block Toeplitz matrix. This block matrix representation of the first derivative is shown to be useful in the context of condition estimation for matrix functions. The results presented here are slightly stronger than those in the literature and are proved in a considerably simpler way.

Key words. derivative, matrix function, condition estimation, Jordan structure

AMS(MOS) subject classifications. 15A99, 47A55, 47A56

1. Introduction.

Let f be an analytic function. It is well known that

$$f\begin{pmatrix} \lambda & w \\ 0 & \lambda \end{pmatrix} = \begin{pmatrix} f(\lambda) & wf'(\lambda) \\ 0 & f(\lambda) \end{pmatrix}.$$

We generalize this by showing that

$$f\begin{pmatrix} A & W \\ 0 & A \end{pmatrix} = \begin{pmatrix} f(A) & \frac{d}{dt}f(A+tW)\Big|_{t=0} \\ 0 & f(A) \end{pmatrix} \tag{1.1}$$

when A and W are square matrices. One can generalize this idea to obtain formulae for higher derivatives.

Our results improve on the results in the literature[1] in several ways. Firstly, [3, 2, 7] all require that $A(t)$ be continuously differentiable in order to conclude that $f(A(t))$ is merely differentiable. Secondly, our method of proof and our expression for the derivative are considerably simpler than those in [3, 2, 7]. Finally, our formula for the derivative is more useful (easier to evaluate and probably more accurate) for numerical computations – see Section 5 and the discussion following Theorem 2.1.

Throughout we let D denote an open subset of $\mathbb{C}$ or $\mathbb{R}$. We let M_n denote the set of $n \times n$ complex matrices and $M_n(D,m)$ denote the set of $n \times n$ matrices that have spectrum contained in D and largest Jordan block of size at most m. Let f be $m-1$ times continuously differentiable on D. Given $A \in M_n(D,m)$ we define $f(A)$ by

$$f(A) = r_{A,f}(A) \tag{1.2}$$

where $r_{A,f}$ is any polynomial that interpolates f and its derivatives at the roots of the minimal polynomial of A. That is, if λ is an eigenvalue of A of index p then

$$f^{(i)}(\lambda) = r_{A,f}^{(i)}(\lambda) \quad i = 0, 1, \ldots, p-1.$$

* Department of Mathematics, College of William & Mary, Williamsburg, VA 23187. e-mail: na.mathias@na-net.ornl.gov. This research was supported in part by National Science Foundation grant DMS-9201586 and a Summer Research Grant from the College of William & Mary.

[1] The book "Matrix Differential Calculus with Applications in Statistics and Econometrics" [11] does not address the subject of this paper.

We discuss the many ways to define $f(A)$ later in the section. By considering the Jordan canonical form of A one can check that the right hand side of (1.2) is indeed independent of the interpolating polynomial chosen – see for example [7, Theorem 6.1.9 (b)] for the details.

We now give three useful properties of functions defined on matrices by (1.2). A simple generalization of [7, Theorem 6.1.28], using the continuity of the divided differences that arise in the interpolation problem, yields

LEMMA 1.1. *Let f be $m-1$ times continuously differentiable on D. Then f is continuous on $M_n(D,m)$.*

This fact is crucial in obtaining (1.1). The definition (1.2) implies the desirable property

$$f(SAS^{-1}) = S\ f(A)\ S^{-1}. \tag{1.3}$$

This will also be used in the proof of (1.1). An immediate corollary of the definition (1.2) is that $f(A)$ depends on f only through its first few derivatives on the spectrum of A:

LEMMA 1.2. *Let $A \in M_n(D,m)$. If for each $k = 1,2,\ldots,m$*

$$f^{(i)}(\lambda) = g^{(i)}(\lambda) \quad i = 0,1,\ldots,k-1$$

for all eigenvalues λ of A of index k then

$$f(A) = g(A).$$

There are a number of ways to extend a scalar valued function to matrices. Rinehart discusses 8 different definitions and shows that many are identical and that all but one are essentially the same in the sense that if for some function f and matrix A two definitions are applicable then the resulting value of $f(A)$ is the same in either case [15]. Some of these definitions are also mentioned in [7, Problems 6.1.14-15, 6.2.1 and Theorem 6.2.28].

In [7, Definition 6.2.4] the notion of a *primary matrix function* derived from a scalar stem function was defined – it is essentially the same as our definition of $f(A)$. However, the starting point in [7] was (1.3) and the requirement that f be a continuous function on $M_n(D,n)$. The relation (1.2) was proved as consequence of these two requirements. The reason for our approach is that if $m < n$ then we can consider functions that are defined on $M_n(D,m)$ but not on $M_n(D,n)$.

One could define $f(A)$ via a contour integral or via a power series, assuming in either case that the scalar function could be expressed in the same way. If one were to use these definitions then one could derive formulae for the derivative of $f(A(t))$ quite easily. However, if $A(t)$ is a Hermitian family of matrices then its spectrum would be real for all t and so it would be reasonable to consider $f(A(t))$ where f is defined only on a subset of the real line rather than a open subset of $\mathbb{C}$, and so f could differentiable without being infinitely differentiable. The question of differentiating such functions of a matrix argument arises in the study of monotone matrix functions (see, e.g., [7,

Section 6.6]). For such functions we would not be able to define $f(A)$ by a contour integral or a power series.

Section 2 contains our main result. Theorem 2.1 is a formal statement of the formula (1.1). This is perhaps the most important result in the paper. The question of differentiating $f(A(t))$ has also been considered by Horn and Johnson [7, Section 6.6], Daleckiĭ and Kreĭn [3] (Hermitian case only), and Daleckiĭ [2]. We compare Theorem 2.1 with their results.

In Section 3 we give an upper bound on the size of the Jordan blocks of certain block upper triangular matrices. This bound is used in Sections 2 and 4, and is only necessary because we want to consider functions f that are not infinitely differentiable and we want to require only the weakest possible differentiability conditions on f. We generalize Theorem 2.1 to higher derivatives in Theorem 4.1.

In Section 5 we present an application of Theorem 2.1.

2. The First Derivative. In this section we give a basic formula for the first derivative of $f(A(t))$. In Section 4 we generalize it to the kth derivative of $f(A(t))$. The following theorem is our basic result.

THEOREM 2.1. *Let f be $2m-1$ times continuously differentiable on D. Let $A(t)$ be differentiable at t_0 and assume that $A(t) \in M_n(D,m)$ for all t in some neighborhood of t_0. Then*

$$\frac{d}{dt} f(A(t))\bigg|_{t=t_0} = \left[f\begin{pmatrix} A(t_0) & A'(t_0) \\ 0 & A(t_0) \end{pmatrix} \right]_{12}. \tag{2.4}$$

The $_{12}$ on the right hand side means 'take the $1,2$ block of the matrix' on the right hand side.

Proof. Take $\epsilon \neq 0$ and let

$$S = \begin{pmatrix} I & \epsilon^{-1} I \\ 0 & I \end{pmatrix}.$$

Then

$$\begin{aligned} f\begin{pmatrix} A(t_0) & \frac{A(t_0+\epsilon)-A(t_0)}{\epsilon} \\ 0 & A(t_0+\epsilon) \end{pmatrix} &= Sf(S^{-1}\begin{pmatrix} A(t_0) & \frac{A(t_0+\epsilon)-A(t_0)}{\epsilon} \\ 0 & A(t_0+\epsilon) \end{pmatrix} S)S^{-1} \\ &= Sf(\begin{pmatrix} A(t_0) & 0 \\ 0 & A(t_0+\epsilon) \end{pmatrix})S^{-1} \\ &= S\begin{pmatrix} f(A(t_0)) & 0 \\ 0 & f(A(t_0+\epsilon)) \end{pmatrix})S^{-1} \\ &= \begin{pmatrix} f(A(t_0)) & \frac{f(A(t_0+\epsilon))-f(A(t_0))}{\epsilon} \\ 0 & f(A(t_0+\epsilon)) \end{pmatrix}. \end{aligned}$$

Now let $\epsilon \to 0$. Because f is $2m-1$ times continuously differentiable and the largest Jordan block of the matrix on the left hand side is at most $2m$ (Lemma 3.1) the continuity of f (Lemma 1.1) implies that the limit of the left hand side exists and is

$$f\begin{pmatrix} A(t_0) & A'(t_0) \\ 0 & A(t_0) \end{pmatrix}.$$

Since the limit on the left hand side exists so does the limit on the right hand side. The $1,2$ block of this limit is $\frac{d}{dt}f(A(t))|_{t=t_0}$. This gives the desired result. □

We have used Lemma 3.1, a bound on Jordan block size, in proving this result. If we had made the stronger assumption that f is $2n-1$ times continuously differentiable (rather than merely $2m-1$ times) then it would not have been necessary to use Lemma 3.1.

Typically one will know only that the size of the largest Jordan block of $A(t)$ is bounded by n so one would usually apply this result with $m=n$. That is, in general f must be $2n-1$ times continuously differentiable in order that $f(A(t))$ be differentiable. If $A(t)$ is Hermitian for all t, then it is also diagonalizable, and hence we may apply the result with $m=1$ and we can conclude that f need only be continuously differentiable in order that $f(A(t))$ be differentiable. In this case, or more generally when $A(t)$ is diagonalizable, the derivative can be expressed in a form involving a Hadamard product [7, Theorem 6.6.30].

Let us compare our result with those in the literature – Daleckiĭ and Krein [3, Theorem 1] (Hermitian case only), Daleckiĭ [2], and Horn and Johnson [7, Theorem 6.6.14]. For comparison we state part of [7, Theorem 6.6.14] which is representative of the other two results also.

THEOREM 2.2. *Let f be $2n-1$ times continuously differentiable on D. Let $A(t)$ be continuously differentiable on D. Then*

1. *$f(A(t))$ is continuously differentiable on D*
2. *Let $t_0 \in D$ be given and let $p_{A(t_0)\oplus A(t_0)}(\cdot)$ be the Newton interpolating polynomial that interpolates f and its derivatives at the zeros of the characteristic polynomial of $A(t_0)\oplus A(t_0)$. Then*

$$\left.\frac{d}{dt}f(A(t))\right|_{t=t_0} = \left.\frac{d}{dt}p_{A(t_0)\oplus A(t_0)}(A(t))\right|_{t=t_0}.$$

3. *For each $t \in D$ let $\lambda_1(t),\dots,\lambda_{\mu(t)}(t)$ denote the distinct eigenvalues of $A(t)$ and let $r_1(t),\dots,r_{\mu(t)}(t)$ denote the their respective multiplicities as zeros of the minimal polynomial of $A(t)$. Let $A_1(t),\dots,A_{\mu(t)}(t)$ denote the Frobenius covariants of $A(t)$ (defined in [7, (6.1.40)]) and let $\Delta f(u,v)$ denote the divided difference $(f(u)-f(v))/(u-v)$. Then*

$$\frac{d}{dt}f(A(t)) = \sum_{j,k=1}^{\mu(t)}\sum_{l=0}^{r_j(t)-1}\sum_{m=0}^{r_k(t)-1}\frac{1}{l!m!}\left.\frac{\partial^{l+m}}{\partial u^l\partial v^m}\Delta f(u,v)\right|_{u=\lambda_j(t),v=\lambda_k(t)}$$

$$\times A_j(t)[A(t)-\lambda_j(t)I]^l\frac{d}{dt}A(t)\ A_k(t)[A(t)-\lambda_k(t)I]^m.$$

All the results in [3, 2, 7] require that $A(t)$ be continuously differentiable at t_0 in order to conclude that $f(A(t))$ is differentiable at $t=t_0$. Our result is stronger than theirs in this respect as we only require that $A(t)$ be differentiable at t_0. Horn and Johnson go on to show that under the stronger assumption of continuous differentiability

$f(A(t))$ is also continuously differentiable[2]. In Corollary 2.3 we show that the formula (2.4) easily yields the continuous differentiability of $f(A(t))$ when $A(t)$ is continuously differentiable. In fact the continuous differentiability of $f(A(t))$ seems quite natural given the formula (2.4), while it seems rather surprising if one looks at a formula for the derivative like those in [7, 3, 2] which involve Frobenius covariants or eigen-projections– quantities which may not even be continuous.

Theorem 2.1 shows that if one can evaluate f at a matrix then one can also compute the derivative of $f(A(t))$ using the same method – we exploit this in the last section. From a computational point of view our formula (2.4) is superior to those in [2, 3, 7]. In particular there is no need to know the eigenvalues of $A(t_0)$ as is required by the formula in part 2 of Theorem 2.2. Part 3 of Theorem 2.2 requires that one also know the Frobenius covariants/eigen-projections of $A(t_0)$. Having the formula depend on the eigenvalues and possibly eigenprojections could be a source of serious errors in numerical computation since the eigenvalues and eigenprojections may be very ill-conditioned.

Our proof of Theorem 2.1 is much simpler than the proofs of the corresponding results in [7, 3, 2] because most of the work is in proving that f is continuous on $M_{2n}(D, 2m)$. Another nice feature of the formula (2.4) is that it allows one to obtain a similar formula for higher derivatives by a simple inductive argument. We indicate how to this at the beginning of Section 4.

Theorem 2.1 covers the Hermitian and non-Hermitian cases together. The arguments in [2, 3, 7] do not. So it may appear that our approach is superior in this respect. It is not. If one were to develop the arguments in [2] or [7, proof of Theorem 6.6.14] more carefully then one would see that the differentiability of $f(A(t))$ is guaranteed by f having $2m_i - 1$ continuous derivatives at each eigenvalue λ_i of $A(t_0)$, where m_i is such that for all t in some neighborhood of t_0 every Jordan block of corresponding to an eigenvalue in a neighborhood of λ_i of $A(t)$ has size at most m_i.[3] In particular, the more careful argument would cover the Hermitian case. (This more careful approach still requires the continuous differentiability of $A(t)$.)

A possible weakness of all these results, Theorem 2.1 included, is that they require f to be *continuously* differentiable in order to conclude that $f(A(t))$ is differentiable. Whereas, if $A(t)$ were a scalar function then it would be sufficient that f be merely differentiable.

Now we show that the continuous differentiability of $A(t)$ guarantees that of $f(A(t))$.

COROLLARY 2.3. *Let f be $2n-1$ times continuously differentiable on D and $A(t) \in M_n(D, n)$ be a continuously differentiable function of t. Then $f(A(t))$ is continuously differentiable.*

Proof. From Theorem 2.1 we know that the derivative of $f(A(t))$ is the 1,2 block of $f(\hat{A}(t))$ where

$$\hat{A}(t) = \begin{pmatrix} A(t) & A'(t) \\ 0 & A(t) \end{pmatrix}.$$

[2] One can check that the continuous differentiability of $A(t)$ is used in an essential way in proving the differentiability of $f(A(t))$. See [7, top of p. 525] for example.

[3] This point has been noted [2, between lines (18) and (19)].

The matrix $\hat{A}(t)$ is a continuous function of t, since $A(t)$ is continuously differentiable. Since f is $2n-1$ times continuously differentiable we know that $f(\hat{A}(t))$ is continuous, and thus,

$$\frac{d}{dt}f(A(t)) = [f(\hat{A}(t))]_{12}$$

is also continuous. □

We shall say no more about continuous differentiability.

3. Bounds on Jordan Block Size. It is useful to have a bound on the size of the Jordan blocks of block upper triangular matrices. The bound can be derived from results due to Friedland and Hershkowitz [4, §3] and Hershkowitz, Rothblum and Schneider [5, Theorem 5.9]. Meyer and Rose also prove this result [13, Theorem 2.1]. For completeness we include a simple proof, which is different from those in the abovementioned papers.

LEMMA 3.1. *Let A be a block upper triangular matrix with square main diagonal blocks $A_{ii}, i = 1, 2, \ldots, m$ that are not necessarily of the same size. Fix $\lambda \in \mathbb{C}$ and let k_i be the index of λ in A_{ii}. Then the index of λ in A is at most $k_1 + k_2 + \cdots + k_m$.*

Proof. It is sufficient to consider the case $\lambda = 0$ and $m = 2$. The general case can be derived from this by considering $A - \lambda I$ and by using induction on m.

To show that the index of 0 in A is at most $k_1 + k_2$ it is sufficient to show that

$$\text{rank } (A^{k_1+k_2}) = \text{rank } (A^{k_1+k_2+1}).$$

This is implied by

$$\text{rank } (A^{k_1+k_2}) \leq \text{rank } (A^{k_1+k_2+1}) \tag{3.1}$$

since rank $(XY) \leq$ rank (X) for any matrices X and Y for which XY is defined. We shall prove (3.1).

Let r_i be the number of non-zero eigenvalues of A_{ii}. Then using the block upper triangularity of A we have

$$\text{rank } (A^{k_1+k_2+1}) \geq \text{rank } (A_{11}^{k_1+k_2+1}) + \text{rank } (A_{22}^{k_1+k_2+1}) \geq r_1 + r_2. \tag{3.2}$$

Let $k = k_1 + k_2$. Then

$$\begin{aligned}
A^{k_1+k_2} &= \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix}^k \\
&= \begin{pmatrix} A_{11}^k & \sum_{j=0}^k A_{11}^j A_{12} A_{22}^{k-j} \\ 0 & A_{22}^k \end{pmatrix} \\
&= \begin{pmatrix} A_{11}^k & \sum_{j=k_1}^k A_{11}^j A_{12} A_{22}^{k-j} \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & \sum_{j=0}^{k_1} A_{11}^j A_{12} A_{22}^{k-j} \\ 0 & A_{22}^k \end{pmatrix} \\
&= \begin{pmatrix} A_{11}^{k_1} & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} A_{11}^{k_2} & \sum_{j=0}^{k_2} A_{11}^j A_{12} A_{22}^{k-j} \\ 0 & 0 \end{pmatrix} \\
&\quad + \begin{pmatrix} 0 & \sum_{j=0}^{k_1} A_{11}^j A_{12} A_{22}^{k_1-j} \\ 0 & A_{22}^{k_1-1} \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 0 & A_{22}^{k_2} \end{pmatrix}
\end{aligned}$$

Since the index of 0 in A_{11} is k_1 it follows that the rank of $A_{11}^{k_1}$ is r_1 and hence the rank of the first term in the sum is at most r_1. In the same way the rank of the second term is at most r_2. Since rank is subadditive we have

$$\text{rank } (A^{k_1+k_2}) \le r_1 + r_2 \le \text{rank } (A^{k_1+k_2+1})$$

as required. The second inequality is from (3.2). □

4. Higher Derivatives. Now let us consider higher derivatives. One approach is to use induction and Theorem 2.1. This would give us

$$\frac{d^2}{dt^2} f(A(t))\Big|_{t=t_0} = \left[f \begin{pmatrix} A(t_0) & A'(t_0) & A'(t_0) & A''(t_0) \\ 0 & A(t_0) & 0 & A'(t_0) \\ 0 & 0 & A(t_0) & A'(t_0) \\ 0 & 0 & 0 & A(t_0) \end{pmatrix} \right]_{14}$$

for the second derivative. Since we have a $4n \times 4n$ matrix on the right hand side one might expect that $4n-1$ continuous derivatives are required of f, but a careful analysis of the Jordan structure of the $4n \times 4n$ matrix shows that $3n-1$ derivatives are sufficient. This approach can be generalized to higher derivatives and one can derive Theorem 4.1 from it, but this is a rather round-about and unnatural development.

Given $n \times n$ matrices $A_0, A_1, \dots, A_k$ let $T[A_0, A_1, \dots, A_k]$ denote the $n(k+1) \times n(k+1)$ block upper triangular block Toeplitz matrix with i, j block equal to A_{j-i} for $j \ge i$. So for example

$$T[A_0, A_1, A_2] = \begin{pmatrix} A_0 & A_1 & A_2 \\ 0 & A_0 & A_1 \\ 0 & 0 & A_0 \end{pmatrix}.$$

THEOREM 4.1. *Let $A(t)$ be k times differentiable at t_0 and assume that $A(t) \in M_n(D, m)$ for all t in some neighborhood of t_0. Assume that f is $(k+1)m-1$ times continuously differentiable on D. Then $f(A(t))$ is k times differentiable at $t = t_0$ and*

$$f(T[A(t_0), \frac{A^{(1)}(t_0)}{1!}, \quad \dots \quad, \frac{A^{(k)}(t_0)}{k!}]) = T[f(A(t_0)), \frac{d}{dt} f(A(t_0)), \dots, \frac{1}{k!}\frac{d^k}{dt^k} f(A(t_0))]. \tag{4.1}$$

Proof. Take $0 = \epsilon_0 < \epsilon_1 < \cdots < \epsilon_k$. Let $\Delta_t^j A(\epsilon_i, \epsilon_{i+1}, \dots, \epsilon_{i+j})$ denote the jth divided difference of A at the $j+1$ points $t+\epsilon_i, \dots, t+\epsilon_{i+j}$. That is, $\Delta_t^0 A = A(t)$ and for $j > 0$

$$\Delta_t^j A(\epsilon_i, \epsilon_{i+1}, \dots, \epsilon_{i+j}) = \frac{\Delta_t^{j-1} A(\epsilon_{i+1}, \dots, \epsilon_{i+j}) - \Delta_t^{j-1} A(\epsilon_i, \dots, \epsilon_{i+j-1})}{(t+\epsilon_{i+j}) - (t-\epsilon_i)}.$$

Let $T_A(\epsilon), T_f(\epsilon)$ and $S(\epsilon)$ denote the $(k+1)n \times (k+1)n$ block upper triangular matrices with i, j block equal to $\Delta_t^{j-i} A(\epsilon_i, \dots, \epsilon_j)$, $\Delta_t^{j-i}(f(A(\epsilon_i, \dots, \epsilon_j))$ and

$$\left[\prod_{l=i}^{j-1} (\epsilon_{j-1} - \epsilon_{l-1}) \right]^{-1} I$$

respectively, for $i \leq j$. If $i > j$ the the ij block is 0 because the matrix is block upper triangular. Let $D(\epsilon)$ be the block diagonal matrix with i,i block equal to $A(t_0+(i-1)\epsilon)$. All these matrices depend on ϵ, but we suppress this dependence in the case of S for simplicity of notation. Note also that in the limit as ϵ goes to 0

$$T_A(\epsilon) \to T[A(t_0), \frac{A^{(1)}(t_0)}{1!}, \ldots, \frac{A^{(k)}(t_0)}{k!}].$$

We now demonstrate that

$$T_A(\epsilon)S = SD(\epsilon) \tag{4.2}$$

by induction on k. The result is immediate when $k = 0$ since then $S = I$ and $T_A(\epsilon) = D(\epsilon)$.

Let us assume that (4.2) is true for $k-1$ and prove it for k. Since by assumption the result is true for $k-1$, every block on the right hand side must be the same as that n the left hand side except perhaps for the $1, k+1$ block. We shall show that this block is also the same by explicitly computing it. The $1, k+1$ block on the left hand side is

$$\begin{aligned}
(T_A(\epsilon)S)_{1,k+1} &= \sum_{j=1}^{k+1} \Delta_{t_0}^{j-1} A(\epsilon_0, \ldots, \epsilon_{j-1}) \left[\prod_{l=j}^{k} (\epsilon_k - \epsilon_{l-1})\right]^{-1} \\
&= \left[\prod_{l=1}^{k} (\epsilon_k - \epsilon_{l-1})\right]^{-1} \sum_{j=1}^{k+1} \Delta_{t_0}^{j-1} A(\epsilon_0, \ldots, \epsilon_{j-1}) \prod_{l=1}^{j-1} (\epsilon_k - \epsilon_{l-1}) \\
&= \left[\prod_{l=1}^{k} (\epsilon_k - \epsilon_{l-1})\right]^{-1} \sum_{j=1}^{k+1} \Delta_{t_0}^{j-1} A(\epsilon_0, \ldots, \epsilon_{j-1}) \prod_{l=1}^{j-1} (t_0 + \epsilon_k - (t_0 + \epsilon_{l-1})) \\
&= \left[\prod_{l=1}^{k} (\epsilon_k - \epsilon_{l-1})\right]^{-1} A(t_0 + \epsilon_k)
\end{aligned}$$

which is the $1, k+1$ block of the right hand side, as desired. The last equality follows from the fact that the penultimate quantity is a multiple of the Newton form of the polynomial that interpolates $A(t)$ at the points $t_0 + \epsilon_0, t_0 + \epsilon_1, \ldots, t_0 + \epsilon_k$ evaluated at the point $t_0 + \epsilon_k$. This can be found in most numerical analysis texts, see for example [1, equation (3.11)].

Since S is non-singular it follows from (4.2) that $S^{-1}T_A(\epsilon)S = D(\epsilon)$. Thus we have

$$f(T_A(\epsilon)) = Sf(S^{-1}T_A(\epsilon)S)S^{-1} = Sf(D(\epsilon))S^{-1} = T_f(\epsilon). \tag{4.3}$$

One can check that

$$\lim_{\epsilon \to 0} T_A(\epsilon) = T[A(t_0), \frac{A^{(1)}(t_0)}{1!}, \ldots, \frac{A^{(k)}(t_0)}{k!}].$$

Lemma 3.1 ensures that the largest Jordan blocks of $T_A(\epsilon)$ is of size at most $(k+1)m$ and so Lemma 1.1 ensures that $f(T_A(\epsilon))$ is continuous at $\epsilon = 0$. That is the limit as

$\epsilon \to 0$ of the term on the extreme left in (4.3) exists and so the limit of the extreme right term must also exist and be the same. If

$$\lim_{\epsilon \to 0} \Delta_{t_0}^j f(A(\epsilon_i, \ldots, \epsilon_{i+j})$$

exists then $f(A(t))$ is necessarily j times differentiable at t_0 and the limit is the dervative. This gives the result. □

Notice that the right hand side of (4.1) depends on f only through $f^{(i)}(\lambda)$ for $i = 0, 1, \ldots, (k+1)m - 1$ and λ in the spectrum of $A(t_0)$. Consequently, if

$$f^{(i)}(\lambda) = g^{(i)}(\lambda) \quad i = 0, 1, \ldots, (k+1)m - 1$$

for all λ in the spectrum of $A(t_0)$ then

$$\frac{d^j}{dt^j} f(A(t))\bigg|_{t=t_0} = \frac{d^j}{dt^j} g(A(t))\bigg|_{t=t_0} \qquad j = 0, 1, \ldots, k.$$

In the case $k = 2$ this observation is [7, Theorem 6.6.14, part 4]. If we further specialize to the case where g is the polynomial that interpolates f and its derivatives at the eigenvalues (counting multiplicities) of $A(t_0) \oplus A(t_0)$ then we obtain part 3 of the same theorem in [7].

5. Applications to Condition Estimation. Often one wishes to compute the condition number for the problem of computing $f(A)$. That is, one wishes to find

$$\inf_{\epsilon \downarrow 0} \max_{\|E\| \le \epsilon} \frac{\|f(A+E) - f(A)\|}{\epsilon} \tag{5.1}$$

for some norm $\|\cdot\|$. (Actually, the relative condition number, i.e., the quantity in (5.1) multiplied by the factor $\|A\|/\|f(A)\|$ is more commonly used. It is easily obtained given (5.1) and so we will consider only (5.1).) One can show that (5.1) is equal to

$$\max_{\|E\| \le \epsilon} \|L_f(A; E)\| \tag{5.2}$$

where $L_f(A; \cdot)$ is the Fréchet derivative of f at A and can be evaluated by

$$L_f(A; E) = \frac{d}{dt} f(A + tE)\bigg|_{t=0}. \tag{5.3}$$

If one can evaluate $L_f(A; E)$ for various values of E and if one takes $\|\cdot\|$ to be the Frobenius norm then one can use a power method [8, 12] or a Lanczos-type method [12] to estimate the quantity in (5.2). However, we know that

$$L_f(A; E) = \left[f \begin{pmatrix} A & E \\ 0 & A \end{pmatrix} \right]_{12}. \tag{5.4}$$

The utility of this observation is that there are special methods to compute $f(X)$ when f is a function with special properties – for example the sine or cosine [16],

the exponential [14], the logarithm [8], the square root [6] and the matrix sign ([9] and the references therein). These special methods immediately yield methods for computing the directional derivative. Furthermore, one can use error analysis and perturbation theory for the function f to obtain error analysis and perturbation theory for its derivative. We illustrate this with the matrix sign function.

The matrix sign function is the matrix function obtained by taking D to be the complex plane excluding the imaginary axis and f to be defined by $f(z) = \text{sign } (Re(z))$. It is defined for any matrix with no eigenvalues on the imaginary axis. Note that f is infinitely differentiable on D.

One way to compute sign (A) is by the Newton iteration

$$A_0 = A, \quad A_{i+1} = \frac{1}{2}(A_i + A_i^{-1}), \quad i = 0, 1, \ldots. \tag{5.5}$$

which is globally convergent to sign (A), assuming of course that A has no eigenvalues on the imaginary axis. This iteration can be accelerated by scaling, see [10] for the details. For simplicity we omit scaling here. The iteration is quadratically convergent to $S = \text{sign } (A)$. One can show that

$$\|A_{i+1} - S\| \leq \frac{1}{8}\|S\| \, \|A_i - A_i^{-1}\|^2 + O(\|A_i - A_i^{-1}\|^3) \tag{5.6}$$

where $\|\cdot\|$ is the spectral norm (or any other submultiplicative norm). The main ideas in the proof of (5.6) are the use of the Neumann series for the inverse and the fact all the quantities that arise (A_i, A_i^{-1} and S) are polynomials in A and therefore commute. When $\|A_i - A_i^{-1}\|$ is small we have $\|S\| \approx \|A_{i+1}\|$ and so (5.6) gives an approximate upper bound on the error in A_{i+1} as an approximation to S.

One can compute $L_{\text{sign}}\,(A, E)$ by applying the Newton iteration (5.5) to

$$B_0 \equiv B \equiv \begin{pmatrix} A & E \\ 0 & A \end{pmatrix}. \tag{5.7}$$

By induction we have

$$(B_i)_{11} = (B_i)_{22} = A_i.$$

Let $E_i = (B_i)_{12}$. Explicitly computing $B_{i+1} = (B_i + B_i^{-1})/2$ gives

$$E_{i+1} = (B_{i+1})_{12} = \frac{1}{2}(E_i - A_i^{-1}E_iA_i^{-1}). \tag{5.8}$$

This iteration for E_i is precisely what was derived in [8, Theorem 3.3]. One can obtain a stopping criterion by applying the error bound (5.6) to the matrices B_i. In particular

$$\begin{aligned}
\|E_{i+1} - L_{\text{sign}}\,(A; E)\| &= \|(B_{i+1})_{12} - (\text{sign } (B))_{12}\| \\
&\leq \|B_{i+1} - \text{sign } (B)\| \\
&\leq \frac{1}{8}\|\text{sign } (B)\| \, \|B_i - B_i^{-1}\|^2 + O(\|B_i - B_i^{-1}\|^3) \\
&\leq \frac{1}{8}(\|S\| + \|L_{\text{sign}}\,(A, E)\|) \, (\|A_i - A_i^{-1}\| + \|E_i + A_i^{-1}E_iA_i^{-1}\|)^2 \\
&\quad + O((\|A_i - A_i^{-1}\| + \|E_i + A_i^{-1}E_iA_i^{-1}\|)^3).
\end{aligned}$$

Notice that

$$\|S\| + \|L_{\text{sign}}\ (A, E)\| \approx \|A_{i+1}\| + \|E_{i+1}\|$$

so we have an approximate upper bound on $\|E_{i+1} - L_{\text{sign}}\ (A; E)\|$ in terms of the known quantities A_i, A_i^{-1}, A_{i+1} and E_{i+1}. This is useful because we generally do not need to compute $L_{\text{sign}}(A; E)$ as accurately as $sign(A)$ and so can stop the iteration (5.8) before the iteration (5.5). Although the iteration (5.8) is not new, the bound on $\|E_{i+1} - L_{\text{sign}}(A; E)\|$ is new.

Acknowledgement: The present proof of Theorem 4.1 is based on an idea provided by an anonymous referee. The original proof was very roundabout and unnatural.

REFERENCES

[1] R. Burden and J. Faires. *Numerical Analysis*. PWS-Kent, Boston, 1993.

[2] Ju. L. Daleckiĭ. Differentiatiation of non-hermitian matrix functions depending on a parameter. *American Mathematical Society Translations, Series 2*, 47:73–87, 1965.

[3] Ju. L. Daleckiĭ and S. G. Kreĭn. Integration and differentiatiation of functions of Hermitian matrices and applications to the theory of perturbations. *American Mathematical Society Translations, Series 2*, 47:1–30, 1965. Russian version published in 1958.

[4] S. Friedland and D. Hershkowitz. The rank of powers of matrices in a block triangular form. *Linear Algebra Appl.*, 107:17–22, 1988.

[5] D. Hershkowitz, U. Rothblum, and H. Schneider. The combinatorial structure of the generalized nullspace of a block triangular matrix. *Linear Algebra Appl.*, 116:9–26, 1989.

[6] Nicholas J. Higham. Newton's method for the matrix square root. *Math. Comp.*, 46(174):537–549, April 1986.

[7] R. A. Horn and C. R. Johnson. *Topics in Matrix Analysis*. Cambridge University Press, New York, 1991.

[8] C. Kenney and A. J. Laub. Polar decomposition and matrix sign function condition estimates. *SIAM J. Sci. Stat. Comp.*, 12(3):488–504, 1991.

[9] C. Kenney and A. J. Laub. Rational iteration methods for the matrix sign function. *SIAM J. Matrix Anal. Appl.*, 12(2):273–291, 1991.

[10] C. Kenney and A. J. Laub. On scaling Newton's method for the polar decomposition and the matrix sign function. *SIAM J. Matrix Anal. Appl.*, 13(3):688–706, 1992.

[11] Magnus and Neudecker. *Matrix Differential Calculus with Applications in Statistics and Econometrics*. Wiley, New York, 1988.

[12] R. Mathias. Evaluating the Fréchet derivative of the matrix exponential. *Numerische Math.*, 63:213–226, 1992.

[13] C. Meyer and N Rose. The index and Drazin inverse of block triangular matrices. *SIAM J. Appl. Math.*, 33(1):1–7, 1976.

[14] C. Moler and C. Van Loan. Nineteen dubious ways to compute the matrix exponential. *SIAM Review*, 20(4):801–836, 1978.

[15] R. F. Rinehart. The equivalence of definitions of a matric function. *Amer. Math. Monthly*, 62:395–413, 1955.

[16] S. Serbin and S. Blalock. An algorithm for computing the matrix cosine. *SIAM J. Sci. Stat. Comp.*, 1:198–204, 1980.