SIAM J. MATRIX ANAL. APPL.
Vol. 10, No. 2, pp. 191–209, April 1989

005

CONDITION ESTIMATES FOR MATRIX FUNCTIONS*

CHARLES KENNEY† AND ALAN J. LAUB†

Abstract. A sensitivity theory based on Fréchet derivatives is presented that has both theoretical and computational advantages. Theoretical results such as a generalization of Van Loan's work on the matrix exponential are easily obtained: matrix functions are least sensitive at normal matrices. Computationally, the central problem is to estimate the norm of the Fréchet derivative, since this is equal to the function's condition number. Two norm-estimation procedures are given; the first is based on a finite-difference approximation of the Fréchet derivative and costs only two extra function evaluations. The second method was developed specifically for the exponential and logarithmic functions; it is based on a trapezoidal approximation scheme suggested by the chain rule for the identity $e^X = (e^{X/2^n})^{2^n}$. This results in an infinite sequence of coupled Sylvester equations that, when truncated, is uniquely suited to the "scaling and squaring" procedure for e^X or the "inverse scaling and squaring" procedure for $\log X$.

Both the trapezoid approximation method and the more general finite-difference approach yield excellent condition estimates for a large class of problems taken from the literature. The problems in this set illustrate that condition estimates based on the Fréchet derivative have the virtue of reliability and general applicability.

Key words. condition estimation, matrix-valued function, exponential, logarithm, Fréchet derivative

AMS(MOS) subject classifications. 65F35, 65F30, 15A12

1. Introduction. In this paper, we are concerned with the effects of perturbations on matrix functions

$$F(X) \equiv \sum_{n=0}^{\infty} a_n X^n \tag{1.1}$$

where $a_n \in \mathbb{R}$ and $X \in \mathbb{R}^{p \times p}$. We assume that the scalar power series

$$F(x) = \sum_{n=0}^{\infty} a_n x^n \tag{1.2}$$

is absolutely convergent for $|x| < r$ for some $r > 0$. We are interested in estimating the "worst case" perturbations as defined by the condition numbers [28]

$$\begin{aligned} K_\delta = K_\delta(F,X) &\equiv \max_{\|Z\| \leq 1} \frac{\|F(X+\delta Z) - F(X)\|}{\delta}, \\ K = K(F,X) &\equiv \lim_{\delta \to 0^+} K_\delta(F,X) \end{aligned} \tag{1.3}$$

where we assume that $\delta > 0$ and $\|X\| + \delta < r$, so that $F(X + \delta Z)$ is well defined. We shall use the Frobenius matrix norm

$$\|M\|^2 \equiv \sum M_{ij}^2 \tag{1.4}$$

throughout the paper unless explicitly noted otherwise, since this norm has nice properties vis-à-vis the Kronecker matrix product.

* Received by the editors February 9, 1988; accepted for publication (in revised form) September 13, 1988. This research was supported by the Office of Naval Research under contract N00014-85-K-0553 and by the National Science Foundation (and the Air Force Office of Scientific Research) under grant ECS87-18897.

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

The condition number $K(F, X)$ of F at X is determined by the Fréchet derivative of F at X: we say that a linear mapping $L : \mathbb{R}^{p\times p} \rightarrow \mathbb{R}^{p\times p}$ is the Fréchet derivative of F at X (see [2], [12]) if for all Z in $\mathbb{R}^{p\times p}$

$$\lim_{\delta\rightarrow 0}\left\|\frac{F(X+\delta Z)-F(X)}{\delta}-L(Z)\right\|=0. \tag{1.5}$$

When it is convenient to explicitly indicate the dependence of L on X, we will write $L(Z, X)$ instead of $L(Z)$. For brevity, we will refer to L as the derivative of F.

Example 1. The squaring function $F(X) = X^2$ satisfies $(F(X + \delta Z) - F(X))/\delta = XZ + ZX + \delta Z^2$, so its derivative at X is given by $L(Z) = XZ + ZX$.

Example 2. The derivative at X of the exponential function $F(X) = e^X$ is given by [31]

$$L(Z)=\int_0^1 e^{X(1-s)}Ze^{Xs}\,ds. \tag{1.6}$$

Other examples are given in Appendix B.

From the definition of the Fréchet derivative (see [31, Thm. 5]), we have

$$K(F,X)=\|L(\cdot,X)\|\equiv\max_{Z\neq 0}\frac{\|L(Z,X)\|}{\|Z\|}. \tag{1.7}$$

Because of this, most of our effort is devoted to studying L and methods for estimating its norm.

In § 2, the eigenvalues of X are used to obtain a lower bound on $K(F, X)$; this lower bound is in fact equal to $K(F, X)$ when X is normal. Thus matrix functions exhibit minimal sensitivity when they are evaluated at normal matrices, an effect demonstrated by Van Loan [31] for the exponential function $F(X) = e^X$. Similar results are given for large scale perturbations.

In § 2, we also lay the groundwork for estimating the norm of L via the power method: given Z_0 of unit norm, let

$$W\equiv L(Z_0,X), \tag{1.8}$$

$$Z_1\equiv L(W,X^T). \tag{1.9}$$

For suitably chosen Z_0, $\|Z_1\|^{1/2} \cong \|L(\cdot, X)\|$, and more accurate estimates can be obtained by repeating the cycle with $Z_0 = Z_1/\|Z_1\|$.

The main problem with this approach is that evaluating $L(Z)$ directly may be rather difficult. For example, in the case of the matrix exponential, it is not at all clear how we should go about evaluating the integral representation in (1.6). In § 3, we consider the problem of forming $L(Z)$ for both the exponential and logarithmic matrix functions. For the exponential problem, $L(Z)$ can be accurately approximated by using a compound trapezoid approximation in (1.7); this approach can be efficiently implemented during the squaring phase of the "scaling and squaring" method of evaluating e^X. This association with scaling and squaring is quite natural because the trapezoid approximation can be derived from the chain rule for the identity $e^X = (e^{X/2^n})^{2^n}$. For the logarithmic problem, a similar approximation can be done during the square root phase of the "inverse scaling and squaring" method of evaluating $\log X$.

While these sensitivity estimation procedures can be easily incorporated into standard packages, such as MATEXP by Ward [32], the numerical effort involved in using them can vary considerably depending on the amount of scaling to be done. For example, one power method cycle of evaluating L and L^T for the matrix exponential can range in cost from $\frac{1}{4}$ to as much as three times the effort needed to evaluate e^X.

By contrast, there is another way of evaluating L such that, independent of the function F, $\|L\|$ can be estimated at a cost of only two extra function evaluations. The idea behind this method is to use the relation

$$\frac{F(X+\delta Z)-F(X)}{\delta}=L(Z,X)+O(\delta) \tag{1.10}$$

as a means of approximating $L(Z,X)$. Thus the power method steps (1.8) and (1.9) can be approximated by

$$W=\frac{F(X+\delta Z_0)-F(X)}{\delta}, \tag{1.11}$$

$$Z_1=\frac{F(X^T+\delta W)-F(X^T)}{\delta}, \tag{1.12}$$

for δ sufficiently small.

To provide a practical assessment of the trapezoid and finite-difference condition estimators, a large set of problems from [3], [7], [16], [25], and [32] was tested numerically; a selected subset of the results is given in § 4. For almost all of the examples, our condition estimates, based on one power method cycle, were within 90 percent of the actual condition number and none of the estimates was less than 25 percent of the actual condition number (see Tables 1 and 2). Of particular interest is an example considered by Ward [32, Example 3] that has shown that the sensitivity estimation scheme employed in the subroutine MATEXP can give very conservative bounds. In this case, Ward's method predicts that not more than 12 digits of accuracy would be lost in the computation of the matrix exponential, whereas one cycle of the power method predicted that at most four digits would be lost; in fact, the numerically computed result had lost exactly four digits of accuracy. This illustrates that condition estimates based on the Fréchet derivative appear to be extremely reliable.

2. General perturbation results. For $\|Z\|=1$ and $\|X\|+\delta<r$, we may write by (1.2),

(2.1)

$$F(X+\delta Z)=F(X)+\delta\sum_{n=1}^{\infty}a_n\sum_{k=0}^{n-1}X^kZX^{n-1-k}+\cdots$$
$$+\delta^m\sum_{n=m}^{\infty}a_n\sum_{k_1+\cdots+k_m=0}^{n-m}X^{k_1}ZX^{k_2}Z\cdots X^{k_m}ZX^{n-m-k_1-\cdots-k_m}+\cdots$$

where the absolute convergence of the series justifies the rearrangement of the terms in (2.1). From (2.1), and (1.5),

$$L(Z,X)=\sum_{n=1}^{\infty}a_n\sum_{k=0}^{n-1}X^kZX^{n-1-k}. \tag{2.2}$$

The discussion in the previous section has shown that the condition number (1.3) satisfies

$$K(F,X)=\|L(\cdot,X)\|=\max_{Z\neq 0}\frac{\|L(Z,X)\|}{\|Z\|}.$$

We use the Frobenius norm (1.4) because of its natural connection to the spectral or two-norm of the Kronecker form of the Fréchet derivative. Let Vec A denote the vector formed by stacking the columns of a matrix A, and define the Kronecker product of two matrices A and B by (see [15]) $A \otimes B \equiv [a_{ij}B]$. Then the Frobenius norm of a matrix Z is equal to the two-norm of Vec Z:

$$\|Z\| = \|\text{Vec } Z\|_2. \tag{2.3}$$

Also, $\text{Vec } (AZB) = (B^T \otimes A) \text{ Vec } Z$. Thus,

$$\text{Vec } L(Z,X) = D(X) \text{ Vec } Z \tag{2.4}$$

where $D(X)$ is the Kronecker form of the Fréchet derivative

$$D(X) \equiv \sum_{n=1}^{\infty} a_n \sum_{k=0}^{n-1} (X^T)^{n-1-k} \otimes X^k. \tag{2.5}$$

By (2.3) and (2.4), we have

$$\max_{Z \neq 0} \frac{\|L(Z,X)\|}{\|Z\|} = \max_{Z \neq 0} \frac{\|\text{Vec } L(Z,X)\|_2}{\|\text{Vec } Z\|_2} = \max_{Z \neq 0} \frac{\|D(X) \text{ Vec } Z\|_2}{\|\text{Vec } Z\|_2},$$

so that the Frobenius norm of the Fréchet derivative is equal to the two-norm of its Kronecker matrix form:

$$\|L(\cdot,X)\| = \|D(X)\|_2. \tag{2.6}$$

The importance of this identity lies in the fact that the two-norm of a real matrix A is the square root of the largest eigenvalue λ of A^TA, and hence can be estimated by using the power method. For a given vector v_0 with $\|v_0\|_2 = 1$, compute the vectors $u_k \equiv Av_k$, $\tilde{v}_{k+1} \equiv A^Tu_k$, $v_{k+1} \equiv \tilde{v}_{k+1}/\|\tilde{v}_{k+1}\|_2$ for $k = 0, 1, 2, \cdots$. If v_0 is not orthogonal to the eigenspace E_λ of A^TA corresponding to λ where $\lambda^{1/2} = \|A\|_2$, then $\|\tilde{v}_k\|^{1/2} \to \|A\|_2$; and unless v_0 is poorly chosen, $\|\tilde{v}_1\|^{1/2} \cong \|A\|_2$. That is, one cycle of the power method provides an approximation of $\|A\|_2$ that is usually sufficient for the purposes of condition estimation [8].

Using $(A \otimes B)^T = A^T \otimes B^T$ and (2.5), we have

$$\begin{aligned}(D(X))^T &= \sum_{n=1}^{\infty} a_n \sum_{k=0}^{n-1} X^{n-1-k} \otimes (X^T)^k \\ &= D(X^T).\end{aligned} \tag{2.7}$$

Define v_1 by $u_0 \equiv D(X)v_0$, $\tilde{v}_1 \equiv (D(X))^Tu_0$, $v_1 \equiv \tilde{v}_1/\|\tilde{v}_1\|_2$. From (2.4) and (2.7) this is equivalent to forming Z_1 by

$$W \equiv L(Z_0,X), \quad \tilde{Z}_1 \equiv L(W,X^T), \quad Z_1 \equiv \tilde{Z}_1/\|\tilde{Z}_1\|, \tag{2.8}$$

where $v_0 = \text{Vec } (Z_0)$ and $v_1 = \text{Vec } (Z_1)$. This is fortunate, because it means that we can avoid dealing with the $p^2 \times p^2$ Kronecker matrix $D(X)$ when estimating the condition of F at X by the power method. Instead, we may use the more compact formulation (2.8).

Now we establish a lower bound on $K(F, X)$ and show that this lower bound is in fact equal to $K(F, X)$ when X is normal.

LEMMA 2.1. *Let v and w be nonzero vectors such that $Xv = \lambda v$ and $X^T w = \mu w$. Then $w \otimes v$ is an eigenvector of $D(X)$ with associated eigenvalue υ where*

$$\upsilon = F'(\lambda) \qquad \text{for } \lambda = \mu,$$

$$\upsilon = \frac{F(\lambda) - F(\mu)}{\lambda - \mu} \qquad \text{for } \lambda \neq \mu. \tag{2.9}$$

Proof. Since $(A \otimes B)(C \otimes D) = (AC) \otimes (BD)$ for any compatible matrices A, B, C, D (see [15]), we have

$$\begin{aligned} D(X)(w \otimes v) &= \sum_{n=1}^{\infty} a_n \sum_{k=0}^{n-1} ((X^T)^{n-1-k} \otimes X^k)(w \otimes v) \\ &= \sum_{n=1}^{\infty} a_n \sum_{k=0}^{n-1} ((X^T)^{n-1-k} w) \otimes (X^k v) \\ &= \sum_{n=1}^{\infty} a_n \left(\sum_{k=0}^{n-1} \mu^{n-1-k} \lambda^k \right) (w \otimes v). \end{aligned}$$

Now, if $\mu = \lambda$, then $\sum_{k=0}^{n-1} \mu^{n-1-k} \lambda^k = n\lambda^{n-1}$ and

$$D(X)(w \otimes v) = \sum_{n=1}^{\infty} n a_n \lambda^{n-1} (w \otimes v) = F'(\lambda)(w \otimes v).$$

Otherwise, if $\mu \neq \lambda$, then $\sum_{k=0}^{n-1} \mu^{n-1-k} \lambda^k = (\lambda^n - \mu^n)/(\lambda - \mu)$ and

$$D(X)(w \otimes v) = \sum_{n=1}^{\infty} a_n \frac{\lambda^n - \mu^n}{\lambda - \mu} (w \otimes v) = \frac{F(\lambda) - F(\mu)}{\lambda - \mu} (w \otimes v). \qquad \square$$

COROLLARY 2.2. *Let $\upsilon_{\max}$ be defined by*

$$\upsilon_{\max} = \max_{\lambda, \mu \in \Lambda(X)} \left| \frac{F(\lambda) - F(\mu)}{\lambda - \mu} \right| \tag{2.10}$$

where $\Lambda(X)$ denotes the set of eigenvalues of X and the ratio in (2.10) is taken to be $|F'(\lambda)|$ when $\lambda = \mu$. Then the condition number of F at X is bounded below by $\upsilon_{\max}$: $\upsilon_{\max} \leqq K(F, X)$.

Proof. By (1.3), (1.7), and (2.6) we have that $K(F, X) = \|L(\cdot, X)\| = \|D(X)\|_2$, but the two-norm of $D(X)$ is bounded below by the absolute value of any eigenvalue of $D(X)$. Hence by Lemma 2.1 we must have $\upsilon_{\max} \leqq \|D(X)\|_2$. □

LEMMA 2.3. *If $X \in \mathbb{R}^{p \times p}$ is normal, that is, $X^T X = X X^T$, then $D(X)$ is normal.*

Proof. Use $(A \otimes B)(C \otimes D) = (AC) \otimes (BD)$ together with $(A \otimes B)^T = A^T \otimes B^T$ and (2.5) to show that $D(X)$ and $D^T(X)$ commute. □

COROLLARY 2.4. *If X is normal, then the condition number of F at X is equal to $\upsilon_{\max}$ in (2.10):*

$$K(F, X) = \max_{\lambda, \mu \in \Lambda(X)} \left| \frac{F(\lambda) - F(\mu)}{\lambda - \mu} \right|.$$

Proof. By Lemma 2.3, $D(X)$ is normal. Thus its two-norm is equal to its spectral radius that, by Lemma 2.1, is just $\upsilon_{\max}$ in (2.10). □

The fact that the lower bound in Corollary 2.2 is attained for normal matrices indicates that the condition number of F is as small as possible when X is normal. This effect has been demonstrated for the exponential function $F(X) = e^X$ in the sensitivity study of Van Loan [31], by using the explicit representation (1.6) together with the property that when X is normal, $\| e^X \|_2 = e^{\alpha(X)}$ where $\alpha(X) \equiv \max_{\lambda \in \Lambda(X)} \operatorname{Re}(\lambda)$.

The preceding dealt with linear perturbation theory of the matrix function $F(X) = \sum_{n=0}^{\infty} a_n X^n$, by considering the limiting behavior of the finite-difference operator

$$DF(Z,X,\delta) \equiv \frac{F(X+\delta Z)-F(X)}{\delta}$$

as $\delta \to 0^+$. We conclude this section with similar results on the behavior of F with respect to large (i.e., nondifferential) perturbations, and our goal will be to bound $K_\delta(F,X)$ in (1.3).

LEMMA 2.5. *Let $\|X\| + \delta < r$. Let v and w be normalized eigenvectors such that $Xv = \lambda_1 v$ and $w^H X = \lambda_2 w^H$. Define $Z = vw^H$. Then $\mu Z = (F(X+\delta Z) - F(X))/\delta$ where*

$$\mu = \frac{F(\lambda_1)-F(\lambda_2)}{\lambda_1 - \lambda_2} \quad \text{if } \lambda_1 \neq \lambda_2, \tag{2.11}$$

$$\mu = \frac{F(\lambda_1 + w^H v\delta)-F(\lambda_1)}{w^H v\delta} \quad \text{if } \lambda_1 = \lambda_2. \tag{2.12}$$

The right-hand side of (2.12) *is taken to be $F'(\lambda_1)$ if $w^H v = 0$. As a consequence, we have the lower bound*

$$\max |\mu| \leqq K_\delta(F,X). \tag{2.13}$$

Proof. The proof is essentially the same as that used in Lemma 2.1. □

The next lemma gives a simple upper bound for $K_\delta(F,X)$ in terms of the function F_+ defined by the associated "positive" series $F_+(x) \equiv \sum_{n=0}^{\infty} |a_n| x^n$.

LEMMA 2.6. *Let $\|Z\| \leqq 1$ and $\|X\| + \delta < r$. Then*

$$K_\delta(F,X) \leqq \frac{F_+(\|X\|+\delta)-F_+(\|X\|)}{\delta} \leqq F'_+(\|X\|+\delta). \tag{2.14}$$

Proof. From (2.1) and $\|Z\| \leqq 1$,

$$\begin{aligned}\|F(X+\delta Z)-F(X)\| &\leqq \delta \sum_{n=1}^{\infty} n|a_n|\,\|X\|^{n-1} + \cdots + \delta^m \sum_{n=m}^{\infty} \binom{n}{m} |a_n|\,\|X\|^{n-m} + \cdots \\ &= \delta F'_+(\|X\|) + \cdots + \frac{\delta^m}{m!} F_+^{m}(\|X\|) + \cdots \\ &= F_+(\|X\|+\delta) - F_+(\|X\|).\end{aligned}$$

Thus

$$\frac{\|F(X+\delta Z)-F(X)\|}{\delta} \leqq \frac{F_+(\|X\|+\delta)-F_+(\|X\|)}{\delta} = F'_+(\|X\|+\rho) \leqq F'_+(\|X\|+\delta)$$

for some $0 \leqq \rho \leqq \delta$ by the mean value theorem and the fact that F'_+ is nondecreasing. □

For example, if $F(x) = e^x$, then Lemma 2.6 gives

$$\| e^{X+\delta Z} - e^X \| / \delta \leqq e^{\|X\|+\delta} - e^{\|X\|} / \delta \leqq e^{\|X\|+\delta}.$$

This upper bound can be very conservative in some cases. However, the next lemma shows that there are situations where the upper bound in Lemma 2.6 coincides with the lower bound in Lemma 2.5 to give an exact value for $K_\delta(F, X)$.

LEMMA 2.7. *Let $X = X^T \geqq 0$ and let the series* (1.2) *have nonnegative coefficients, $a_n \geqq 0$, so that $F = F_+$. Then $K_\delta(F, X) = (F(\|X\| + \delta) - F(\|X\|))/\delta$.*

Proof. Since X is nonnegative definite symmetric there exists a real eigenvector v such that $Xv = \lambda v$ with $v^Tv = 1$ where $\lambda = \|X\|$. By Lemma 2.5, $\mu Z = (F(X + \delta Z) - F(X))/\delta$ where $Z = vv^T$ and $\mu = (F(\lambda + \delta) - F(\lambda))/\delta$. Note that $\mu > 0$, since F is nondecreasing. Thus, $K_\delta(F, X) \geqq \|\mu Z\| = \mu$, since $\|Z\| = \|vv^T\| = 1$. On the other hand, $\mu = (F(\|X\| + \delta) - F(\|X\|))/\delta$, since $\|X\| = \lambda$. Thus, since $F = F_+$, we have by Lemma 2.6 that $K_\delta(F, X) \leqq \mu$. This shows that we must have $K_\delta(F, X) = \mu = (F(\|X\| + \delta) - F(\|X\|))/\delta$. □

The next lemma shows that for $F = F_+$ and large δ, Lemma 2.7 is approximately true, not just for symmetric nonnegative definite matrices, but for *any* matrix X.

LEMMA 2.8. *Let $F = F_+$ with radius of convergence $r = \infty$. Then for $\delta > 2\|X\|$ and any real matrix X, we have*

$$\frac{F(\delta - \|X\|) - F(\|X\|)}{\delta} \leqq K_\delta(F, X) \leqq \frac{F(\delta + \|X\|) - F(\|X\|)}{\delta}. \tag{2.15}$$

Proof. The right-hand side inequality of (2.15) is simply a restatement of (2.14) in Lemma 2.6. To prove the left-hand side inequality in (2.15), let $Z_1 \equiv e_1e_1^T$ where $e_1 \equiv (1, 0, \cdots, 0)^T$ and set $Z = (1 - \varepsilon)Z_1 - X/\delta$ where ε is chosen so that $\|Z\| = 1$. Then $1 = \|Z\| \leqq 1 - \varepsilon + \|X\|/\delta$ so $\varepsilon \leqq \|X\|/\delta$. Now $X + \delta Z = \delta(1 - \varepsilon)Z_1$, so $F(X + \delta Z) = F(\delta(1 - \varepsilon))Z_1$, since $F(\alpha Z_1) = F(\alpha)Z_1$ for any scalar α. Moreover, since $\delta\varepsilon \leqq \|X\|$, $\|F(X + \delta Z)\| = F(\delta(1 - \varepsilon)) \geqq F(\delta - \|X\|)$ because $F = F_+$ is nondecreasing. However, $\|F(X + \delta Z)\| - \|F(X)\| \leqq \|F(X + \delta Z) - F(X)\|$. Thus,

$$F(\delta - \|X\|) - F(\|X\|) \leqq \|F(X + \delta Z)\| - \|F(X)\| \leqq \|F(X + \delta Z) - F(X)\|$$

because $\|F(X)\| \leqq F(\|X\|)$. Dividing by δ in the above completes the proof. □

Example. Let $F(X) = e^X$ with $\|X\| = 1$; then by Lemma 2.8, we have that

$$\frac{e^{\delta - 1} - e}{\delta} \leqq K_\delta(F, X) \leqq \frac{e^{\delta + 1} - e}{\delta},$$

which determines K_δ to within a factor of e^2 for large δ.

3. Exponential and logarithmic linear perturbation theory. In this section, we treat the problem of approximating the Fréchet derivatives of the exponential and logarithmic matrix functions.

The earliest representation of the exponential derivative appears to be due to Hausdorff [17]:

$$L(Z, X) = e^X \sum_{n=0}^{\infty} \frac{1}{(n+1)!}\{Z, X^n\} = e^X\{Z, (e^X - I)X^{-1}\} \tag{3.1}$$

where the nested Lie product $\{\cdot, \cdot\}$ is defined by $\{Z, X^n\} \equiv [\cdots[[Z, X], X], \cdots, X]$ with n factors of X appearing; $[Z, X]$ denotes the Lie bracket $ZX - XZ$. In the rightmost side of (3.1), the expression $(e^X - I)X^{-1}$ should be interpreted as the series $\sum_{m=0}^{\infty} X^m/(m + 1)!$ when X is not invertible. This Lie product expansion for the exponential derivative arose in connection with the Baker–Campbell–Hausdorff formula (see [23, pp. 656–658], [4])

$$e^Xe^Y = \exp\left(X + Y + \frac{1}{2}[X, Y] + \frac{1}{12}[[X, Y], Y - X] + \cdots\right).$$

From (2.2), with $F(X) = e^X$, we obtain another series representation:

$$L(Z,X) = \sum_{n=1}^{\infty} \frac{1}{n!} \sum_{k=0}^{n-1} X^k Z X^{n-1-k},$$

but this series and (3.1) are too hard to work with numerically. A much more useful representation is given by Van Loan in [31]: $L(Z, X) = \int_0^1 e^{X(1-s)} Z e^{Xs}\, ds$. This may be approximated by using the trapezoid rule (see [10])

$$(3.2) \qquad L(Z,X) \cong L_n(Z,X) \equiv \frac{1}{2^{n+1}}\left[e^X Z + 2 \sum_{k=1}^{2^n-1} e^{kX/2^n} Z e^{(2^n-k)X/2^n} + Z e^X\right]$$

for $n = 0, 1, 2, \cdots$.

We have selected this method of approximation because it is uniquely suited to one of the most successful methods of computing e^X, namely the scaling and squaring method [24], [32]. In this method, X is scaled by a power of two, say 2^n, so that $e^{X/2^n}$ is easily evaluated by using, for example, a Padé approximation. The result is then squared n times: $e^X = (e^{X/2^n})^{2^n}$. During the squaring phase, we have available to us sequentially the computed values of the matrices $e^{X/2^n}$, $e^{X/2^{n-1}}$, $\cdots$, $e^{X/2}$, and e^X. This raises the possibility of evaluating the trapezoid approximant, $L_n(Z, X)$, for a given matrix Z, during the computation of e^X. This would not be practical if we implemented (3.2) directly, but fortunately there is an equivalent formulation for $L_n(Z, X)$ that is much easier to evaluate and only requires the matrices $e^{X/2^{n-k}}$ as they become available. Let

$$(3.3) \qquad W_0 \equiv (e^{X/2^n} Z + Z e^{X/2^n})/2^{n+1},$$

and for $j = n, n-1, \cdots, 1$, define

$$(3.4) \qquad W_{n+1-j} \equiv e^{X/2^j} W_{n-j} + W_{n-j} e^{X/2^j}.$$

Then from (3.2)–(3.4), $L_n(Z, X) = W_n$.

We will show that if n is large, then $L_n(Z, X)$ is near $L(Z, X)$ and $\|L_n(\cdot, X)\|$ provides a good estimate of $\|L(\cdot, X)\|$. For example, if n is large enough so that $\|e^{X/2^n} - I\| < \frac{1}{4}$, then our results give

$$0.950\|L_n(\cdot,X)\| \leqq \|L(\cdot,X)\| \leqq 1.055\|L_n(\cdot,X)\|.$$

Somewhat surprisingly, it seems to be the case that the easiest way to determine how well $L_n(\cdot, X)$ approximates $L(\cdot, X)$ is to study the approximation of $L^{-1}(\cdot, X)$ by $L_n^{-1}(\cdot, X)$. Both $L^{-1}(\cdot, X)$ and $L_n^{-1}(\cdot, X)$ arise naturally in the study of the inverse exponential or logarithmic problem $e^X \rightarrow X$. Such problems occur, for example, in a control theory setting wherein discrete samples from a continuous system are used to identify system parameters. See [20], [27], and [29]. Since the logarithm is a multivalued function, we need some restriction on X to ensure the existence of a unique real solution X to the problem $e^X = A$ (cf. [9], [13], [18], [33]). To do so, we shall assume throughout this section that $A \in \mathbb{R}^{p \times p}$ has no eigenvalues on the negative real axis including zero. This is sufficient to ensure that there exists a unique real matrix $X \in \mathbb{R}^{p \times p}$ such that $e^X = A$ with the eigenvalues of X confined to the strip $-\pi < \operatorname{Im}(z) < \pi$. (See Appendix A.)

Under the above assumptions,

$$(3.5) \qquad \log A = 2^n \log A^{1/2^n}$$

where $A^{1/2^n}$ denotes the unique real nth square root of A (see [11]) whose eigenvalues, $\lambda = \lambda(A^{1/2^n})$ lie in the sector $-\pi/2^n < \arg(\lambda) < \pi/2^n$. (See Appendix A.) This forms the basis of the "inverse scaling and squaring" method for approximating $\log A$. Take n square roots of A, so that $A^{1/2^n}$ is near the identity. Then $\log A^{1/2^n}$ can be computed by using, for example, a Padé approximation in the variable $Y \equiv I - A^{1/2^n}$. Multiplying the result by 2^n, we obtain $\log A$ as in (3.5). (See § 4 for more details.)

By Lemma A1, in Appendix A, the derivative of the logarithmic function $e^X \rightarrow X$ is the inverse of $L(\cdot, X)$ in (1.6) provided $L(\cdot, X)$ is invertible. However, $L(\cdot, X)$ is invertible if and only if the associated Kronecker matrix $D(X)$ given by (2.5) is nonsingular. By Lemma 2.1, $D(X)$ is singular for the exponential function if and only if $e^\lambda = 0$ or $(e^\lambda - e^\mu)/(\lambda - \mu) = 0$ for $\lambda, \mu \in \Lambda(X)$, $\mu \neq \lambda$. However, e^λ is never zero and $e^\lambda = e^\mu$ with $\mu \neq \lambda$ means that $\lambda = \mu + 2\pi i k$ for some nonzero integer k, which would violate the condition that $-\pi < \operatorname{Im}(\lambda), \operatorname{Im}(\mu) < \pi$. Thus $D(X)$ is nonsingular and $L(\cdot, X)$ is invertible whenever $\Lambda(X)$ is confined to the strip $-\pi < \operatorname{Im}(z) < \pi$. This strip condition also implies that $L_n(\cdot, X)$ is invertible. To see this, note that $L_n^{-1}(W, X)$, for a given matrix W, can be found by inverting the procedure in (3.3)–(3.4). That is, for $A = e^X$, set $W_n = W$ and solve sequentially for $W_{n-1}, \cdots, W_0$ and Z in

$$W_{n+1-j} = A^{1/2^j} W_{n-j} + W_{n-j} A^{1/2^j}, \tag{3.6}$$

$$2^{n+1} W_0 = A^{1/2^n} Z + Z A^{1/2^n}. \tag{3.7}$$

Then $L_n^{-1}(W, X) = Z$ because $L_n(Z, X) = W$ by (3.2)–(3.4).

From this we see that $L_n^{-1}(\cdot, X)$ is invertible whenever the Sylvester equations (3.6) and (3.7) are uniquely solvable. However, the strip condition on X forces the eigenvalues of $A^{1/2^j}$, for $j \geqq 1$, to lie in the open right-half complex plane. Consequently $\mu + \lambda \neq 0$ for $\mu, \lambda \in \Lambda(A^{1/2^j})$ and (3.6), (3.7) have unique solutions [21].

The sequence $W_0, \cdots, W_n$ has a nice representation that forms the basis of our analysis of the relationship between L and L_n and which originally inspired our work in this area.

LEMMA 3.1. *Let W_j be defined by* (3.6), (3.7) *with $W = W_n \equiv L(\hat{Z}, X)$, i.e., $W_n = \int_0^1 e^{X(1-s)} \hat{Z} e^{Xs}\, ds$. Then*

$$W_{n-j} = \int_0^{1/2^j} e^{X(1/2^j - s)} \hat{Z} e^{Xs}\, ds \tag{3.8}$$

for $j = 1, \cdots, n$.

Proof. We show that (3.8) is valid for $j = 1$; for $j > 1$ use similar arguments. By (3.6),

$$W_n = A^{1/2} W_{n-1} + W_{n-1} A^{1/2},$$

which we may rewrite, using $A = e^X$, as $W_n = e^{X/2} W_{n-1} + W_{n-1} e^{X/2}$. Under the assumption that $\Lambda(X)$ lies in the strip $-\pi < \operatorname{Im}(z) < \pi$, this equation has a unique solution. Thus it is sufficient to show that $W_n = e^{X/2} \tilde{W}_{n-1} + \tilde{W}_{n-1} e^{X/2}$ where $\tilde{W}_{n-1} \equiv \int_0^{1/2} e^{X(1/2-s)} \hat{Z} e^{Xs}\, ds$. But

$$\begin{aligned} e^{X/2} \tilde{W}_{n-1} + \tilde{W}_{n-1} e^{X/2} &= \int_0^{1/2} e^{X(1-s)} \hat{Z} e^{Xs}\, ds + \int_0^{1/2} e^{X(1/2-s)} \hat{Z} e^{X(s+1/2)}\, ds \\ &= \int_0^{1/2} e^{X(1-s)} \hat{Z} e^{Xs}\, ds + \int_{1/2}^{1} e^{X(1-s)} \hat{Z} e^{Xs}\, ds \\ &= \int_0^1 e^{X(1-s)} \hat{Z} e^{Xs}\, ds \equiv W. \qquad \square \end{aligned}$$

We need two technical lemmas to prove our main result (Theorem 3.4).

LEMMA 3.2. *Let* $[C,B]$ *denote the Lie product* $[C,B]\equiv CB-BC$. *Then we may write*

$$\int_0^1 e^{B(1-s)}Ce^{Bs}\,ds=\frac{1}{2}(e^BC+Ce^B)-\frac{1}{2}\int_0^1[e^{B(1-s)},[e^{Bs},C]]\,ds.$$

Proof. Expand the nested Lie product on the right-hand side and use the identity

$$\int_0^1 e^{B(1-s)}Ce^{Bs}\,ds=\int_0^1 e^{Bs}Ce^{B(1-s)}\,ds. \qquad \square$$

LEMMA 3.3. *Let* $\mu(B)\equiv\frac{1}{2}\lambda_{\max}(B+B^T)$. *Then we have*

$$\left\|\int_0^1 e^{B(1-s)}Ce^{Bs}\,ds-\frac{1}{2}(e^BC+Ce^B)\right\|\leqq\frac{1}{3}\|C\|\,\|B\|^2e^{\mu(B)}. \tag{3.9}$$

Proof. Use the methods of Lemma 3 of [24, Appendix 2]. □

Using the preceding three lemmas we can now prove our main result on the approximation of $L^{-1}(\cdot,X)$ by $L_n^{-1}(\cdot,X)$.

THEOREM 3.4. *Let n be large enough so that* $\omega\equiv\|I-e^{X/2^n}\|<1$. *Then for any* $W\in\mathbb{R}^{p\times p}$, *we have*

$$\|L^{-1}(W,X)-L_n^{-1}(W,X)\|\leqq\frac{1}{3}\left(\frac{1}{1-\omega}\log\left(\frac{1}{1-\omega}\right)\right)^2\|L^{-1}(W,X)\|. \tag{3.10}$$

Proof. Let $L_n(Z,X)=W$ and $L(\hat{Z},X)=W$ so that $Z=L_n^{-1}(W,X)$ and $\hat{Z}=L^{-1}(W,X)$. Now define $W_0,W_1,\cdots,W_{n-1}$ by (3.6), (3.7), so that by the definition of L_n^{-1} in (3.6), (3.7)

$$W_0=\frac{(e^{X/2^n}Z+Ze^{X/2^n})}{2^{n+1}}. \tag{3.11}$$

However, by Lemma 3.1,

$$W_0=\int_0^{1/2^n}e^{X(1/2^n-s)}\hat{Z}e^{Xs}\,ds. \tag{3.12}$$

By the change of variables, $s\to 2^ns$,

$$\int_0^{1/2^n}e^{X(1/2^n-s)}\hat{Z}e^{Xs}\,ds=\frac{1}{2^n}\int_0^1 e^{(X/2^n)(1-s)}\hat{Z}e^{(X/2^n)s}\,ds. \tag{3.13}$$

Now by Lemma 3.2 with $B=X/2^n$ and $C=\hat{Z}$,

$$\begin{aligned}\frac{1}{2^n}\int_0^1 e^{(X/2^n)(1-s)}\hat{Z}e^{(X/2^n)s}\,ds&=\frac{1}{2^{n+1}}(e^{X/2^n}\hat{Z}+\hat{Z}e^{X/2^n})\\&\quad-\frac{1}{2^{n+1}}\int_0^1[e^{(X/2^n)(1-s)},[e^{(X/2^n)s},\hat{Z}]]\,ds.\end{aligned} \tag{3.14}$$

Combining (3.11)–(3.14), we obtain

$$e^{X/2^n}Z+Ze^{X/2^n}=e^{X/2^n}\hat{Z}+\hat{Z}e^{X/2^n}-\int_0^1[e^{(X/2^n)(1-s)},[e^{(X/2^n)s},\hat{Z}]]\,ds.$$

This may be written as $\Omega(\hat{Z} - Z) = \int_0^1 [e^{(X/2^n)(1-s)}, [e^{(X/2^n)s}, \hat{Z}]]\,ds$ where $\Omega(V) \equiv e^{X/2^n}V + Ve^{X/2^n}$. Thus $\hat{Z} - Z = \Omega^{-1}(\int_0^1 [e^{(X/2^n)(1-s)}, [e^{(X/2^n)s}, \hat{Z}]]\,ds)$, so

$$\text{(3.15)} \qquad \begin{aligned} \|\hat{Z} - Z\| &\leqq \|\Omega^{-1}\| \left\| \int_0^1 [e^{(X/2^n)(1-s)}, [e^{(X/2^n)s}, \hat{Z}]]\,ds \right\| \\ &\leqq \|\Omega^{-1}\| \frac{2}{3} \left\| \frac{X}{2^n} \right\|^2 e^{\mu(X/2^n)} \|\hat{Z}\| \end{aligned}$$

as in the proof of Lemma 3.3. We now show that for $\omega \equiv \|I - e^{X/2^n}\| < 1$,

$$\text{(3.16)} \qquad \|\Omega^{-1}\| \leqq \frac{1}{2}\frac{1}{1-\omega},$$

$$\text{(3.17)} \qquad \left\| \frac{X}{2^n} \right\| \leqq \log\left(\frac{1}{1-\omega}\right),$$

$$\text{(3.18)} \qquad e^{\mu(X/2^n)} \leqq \frac{1}{1-\omega}.$$

When combined with (3.15) and $\hat{Z} = L^{-1}(W, X)$, $Z = L_n^{-1}(W, X)$, we shall have (3.10), thus completing the proof.

To show (3.16), let $Q = \Omega(V)$ so that $V = \Omega^{-1}(Q)$. Now, $2V = Q + YV + VY$ where $Y \equiv I - e^{X/2^n}$ and $\|Y\| = \omega < 1$. Thus, $2\|V\| \leqq \|Q\| + 2\omega\|V\|$, so $\|V\| \leqq \|Q\|/2(1 - \omega)$. Inequality (3.16) follows immediately since $\|V\| = \|\Omega^{-1}(Q)\|$.

To get (3.17), use $X/2^n = \log e^{X/2^n} = \log(I - Y)$, and

$$\|\log(I - Y)\| \leqq \sum_{m=1}^{\infty} \frac{\|Y\|^m}{m} = |\log(1 - \|Y\|)| = \log\left(\frac{1}{1-\omega}\right).$$

This also gives (3.18) because $e^{\mu(X/2^n)} \leqq e^{\|X/2^n\|} \leqq \exp(\log(1/(1-\omega))) = 1/(1-\omega)$. □

From Theorem 3.4, we can easily obtain a bound on the logarithmic condition number, $\|L^{-1}(\cdot, X)\|$ in terms of the norm, $\|L_n^{-1}(\cdot, X)\|$ of the inverse trapezoid approximant.

COROLLARY 3.5. *Let $\omega \equiv \|I - e^{X/2^n}\| < 1$ and define*

$$\omega_1 \equiv \tfrac{1}{3}(1/(1-\omega)\log(1/(1-\omega)))^2.$$

If n is large enough so that $\omega_1 < 1$, then

$$\|L_n^{-1}(\cdot, X)\|/(1 + \omega_1) \leqq \|L^{-1}(\cdot, X)\| \leqq \|L_n^{-1}(\cdot, X)\|/(1 - \omega_1).$$

Proof. Use (3.10) for the proof. □

As an example, if $\|I - e^{X/2^n}\| \leqq \frac{1}{4}$, then $0.953\|L_n^{-1}(\cdot, X)\| \leqq \|L^{-1}(\cdot, X)\| \leqq 1.052\|L_n^{-1}(\cdot, X)\|$. To obtain bounds on the exponential condition number $\|L(\cdot, X)\|$, we now return to the problem of how well $L_n(\cdot, X)$ approximates $L(\cdot, X)$.

THEOREM 3.6. *Let $\omega_2 \equiv \omega_1/(1 - \omega_1)$ for ω and ω_1 as in Corollary 3.5. Assume that n is large enough so that ω, ω_1, and ω_2 are less than one. Then for any $Z \in \mathbb{R}^{p \times p}$, we have*

$$\|L(Z, X) - L_n(Z, X)\| \leqq \omega_2 \|L(\cdot, X)\| \, \|Z\|.$$

Proof. Let Z be given and let $W \equiv L_n(Z, X)$, so that $L_n^{-1}(W, X) = Z$. Let $\hat{Z} \equiv L^{-1}(W, X)$ so that $L(\hat{Z}, X) = W = L_n(Z, X)$. Then by Theorem 3.4, $\|\hat{Z} - Z\| \leqq$

$\omega_1\|\hat{Z}\|$. But by Corollary 3.5, $\|\hat{Z}\| \leqq \|Z\|/(1-\omega_1)$ so $\|\hat{Z}-Z\| \leqq \omega_1\|Z\|/(1-\omega_1) \equiv \omega_2\|Z\|$. Thus,

$$\begin{aligned}\|L(Z,X)-L_n(Z,X)\| &= \|L(Z,X)-L(\hat{Z},X)\|\\ &= \|L(Z-\hat{Z},X)\|\\ &\leqq \|L(\cdot,X)\|\,\|Z-\hat{Z}\|\\ &\leqq \|L(\cdot,X)\|\omega_2\|Z\|. \qquad \square\end{aligned}$$

COROLLARY 3.7. *Under the assumptions of Theorem* 3.6,

$$\|L_n(\cdot,X)\|/(1+\omega_2) \leqq \|L(\cdot,X)\| \leqq \|L_n(\cdot,X)\|/(1-\omega_2).$$

Proof. Use standard norm arguments and Theorem 3.6 for the proof. □

As an example, if $\|I-e^{X/2^n}\| \leqq \frac{1}{4}$, then $0.950\|L_n(\cdot,X)\| \leqq \|L(\cdot,X)\| \leqq 1.055\|L_n(\cdot,X)\|$.

To illustrate the trapezoid approximation method and Theorem 3.6, let $X = [\begin{smallmatrix}0&1\\0&0\end{smallmatrix}]$. Then

$$e^X = \begin{bmatrix}1 & 1\\ 0 & 1\end{bmatrix} \quad \text{and} \quad e^{X/2^n} = \begin{bmatrix}1 & 1/2^n\\ 0 & 1\end{bmatrix}.$$

If we impose a scaling condition of $\|I-e^{X/2^n}\| < \frac{1}{4}$, then we may take $n = 3$, in which case $\|I-e^{X/8}\| = \frac{1}{8}$. Let $Z = e^X e^{X^T} + e^{X^T}e^X = [\begin{smallmatrix}3&2\\2&3\end{smallmatrix}]$. (For reasons explained in the next section, this choice of Z can be expected to have a large component in the matrix direction which maximally perturbs e^X.) Using (3.3) and (3.4), we find (to four significant figures), $L_3(Z,X) = [\begin{smallmatrix}4&5.328\\2&4\end{smallmatrix}]$. To compare this with $L(Z,X)$, note that X is nilpotent with $X^2 = 0$. Thus from (2.2),

$$L(Z,X) = \sum_{n=1}^{\infty}\frac{1}{n!}\sum_{k=0}^{n-1}X^kZX^{n-1-k} = Z + \frac{XZ+ZX}{2} + \frac{XZX}{6} = \begin{bmatrix}4 & 5.333\\ 2 & 4\end{bmatrix}.$$

This gives, as in Theorem 3.6, $0.005 = \|L(Z,X)-L_3(Z,X)\| \leqq \|L(\cdot,X)\|\omega_2\|Z\| = 0.064$, where $\|L(\cdot,X)\| = 1.609$ was determined by finding the largest singular value of the associated Kronecker matrix D in (2.5). It is interesting to note that one power method cycle, using L_3 and this Z, gives an estimate of 1.592 for the norm of $L(\cdot,X)$.

4. Numerical results. In this section, we discuss some of the details of testing the trapezoid approximation and finite-difference condition estimation procedures for the exponential and logarithmic matrix functions.

From § 3, we have implemented the trapezoid approximation method (3.3), (3.4) for the matrix exponential in conjunction with the subroutine MATEXP of Ward [32]. To avoid analytical complications in the sensitivity estimate resulting from the use of the balancing transformation BALANX (which is a modified version of the EISPACK subroutine BALANC [30]), we have implemented (3.3), (3.4) after the back substitution BALINV in MATEXP. For one cycle of the power method, this results in a condition estimate, which costs $4n+4$ matrix multiplications, where the scaling parameter n is chosen so that $\|X/2^n\|_1 \leqq \log(5/4) \cong 0.223$. This ensures that $\|I-e^{X/2^n}\| \leqq \frac{1}{4}$, as seen by the following lemma. (By Corollary 3.7, this scaling condition forces the norm of the trapezoid approximation, $\|L_n(\cdot,X)\|$ to be within six percent of the exponential condition number, $\|L(\cdot,X)\|$.)

LEMMA 4.1. *If* $\|Z\| \leqq \log(1+\omega)$, *then* $\|I-e^Z\| \leqq \omega$.

Proof. $\|I-e^Z\| \leqq e^{\|Z\|}-1 \leqq e^{\log(1+\omega)}-1 = \omega$. □

The subroutine MATEXP needs about $8 + n$ matrix multiplications to evaluate e^X (to a relative precision of about 10^{-16}), so the sensitivity estimate via (3.3), (3.4) is about 1.9 times as expensive as evaluating e^X when $n = 6$, which was the average value of n for the examples we considered.

We also implemented the inverse trapezoid approximation (3.6), (3.7) to estimate the condition of the logarithm, subject to the scaling condition $\|I - A^{1/2^n}\| \leqq \frac{1}{4}$. The square root of a matrix can be obtained in a stable manner by using the Schur algorithm described in [5]. This involves finding the real Schur form of A: $A = QTQ^T$ where Q is orthogonal and T is quasi-upper-triangular. Once this is done, $A^{1/2} = QT^{1/2}Q^T$, where $T^{1/2}$ is found by a simple linear recursion involving the entries of T and the square roots of the main diagonal entries of T (including the 2×2 blocks corresponding to the complex conjugate eigenvalues of T; see [22]). Moreover, the jth square root satisfies $A^{1/2^j} = QT^{1/2^j}Q^T$, which means that the Schur decomposition need only be done once in the process of generating $A^{1/2^n}$. This is important because the Schur decomposition of a matrix of order p requires about $8p^3$ floating-point operations (flops), whereas the square root of a quasi-upper-triangular matrix of order p requires only about $p^3/6$ flops. The logarithm of $A^{1/2^n}$ can be approximated by truncating the slowly convergent Taylor series, $\log(I - Y) = -\sum_{m=1}^{\infty} Y^m/m$, but rational Padé approximants are generally superior. For example, it is shown in [20] that if $\|Y\| = \|I - A^{1/2^n}\| \leqq \frac{1}{4}$, then the eighth-order main diagonal Padé approximant $R_{88}(Y) \equiv P_{88}(Y)Q_{88}^{-1}(Y)$ differs from $\log(I - Y)$ by less than 10^{-18}, whereas the sixteenth-order Taylor approximant, which requires about the same amount of work, can be in error by as much as 5×10^{-12}. In the above,

$$P_{88}(Y) \equiv -Y + \frac{7}{2}Y^2 - \frac{73}{15}Y^3 + \frac{41}{12}Y^4 - \frac{743}{585}Y^5 + \frac{31}{130}Y^6 - \frac{111}{5775}Y^7 + \frac{761}{1801800}Y^8,$$

$$Q_{88}(Y) \equiv 1 - 4Y + \frac{98}{15}Y^2 - \frac{28}{5}Y^3 + \frac{35}{13}Y^4 - \frac{28}{39}Y^5 + \frac{14}{143}Y^6 - \frac{4}{715}Y^7 + \frac{1}{12870}Y^8.$$

Moreover, when $\|I - A^{1/2^n}\| \leqq \frac{1}{4}$, the Padé denominator matrix $Q_{88}(Y)$ is very well-conditioned with $K(Q_{88}(Y)) \equiv \|Q_{88}\| \, \|Q_{88}^{-1}\| \leqq 7.59$ (see [20]).

The inverse scaling and squaring procedure for evaluating the logarithm of a matrix takes about $11 + n/6$ matrix multiplications, whereas the first cycle of the power method of estimating the condition number $\|L^{-1}(\cdot, X)\|$ takes about $2 + 13/6n$ matrix multiplications. Thus the condition estimate takes about 1.2 times the effort needed to evaluate the logarithm when $n = 6$.

We have also implemented the "finite-difference" power method for the exponential and logarithmic functions. Given Z_0 define $\tilde{W}_0 \equiv (F(X + \delta Z_0) - F(X))/\delta$, $W_0 \equiv \tilde{W}_0/\|\tilde{W}_0\|$, and $Z_1 \equiv (F(X^T + \delta W_0) - F(X^T))/\delta$. Then $\|Z_1\|$ provides a condition estimate of F at X, at a cost of two function evaluations beyond $F(X)$, when we use the fact that $F(X^T) = (F(X))^T$.

A common problem, for both the trapezoid and finite-difference approximation methods, is the choice of the initial matrix Z_0. The complex nature of both methods makes it difficult to use "look-ahead" procedures such as those described in [8] and [6]. Instead, we have tried two different methods of choosing Z_0. The first consists of letting Z_0 have random entries in the interval $[-1, 1]$. This practically guarantees that Z_0 has a nontrivial component in the matrix direction that maximizes $\|L(Z, X)\|$. Consequently, one power method cycle usually provides an estimate of $\|L(\cdot, X)\|$ that is sufficient for the purposes of condition estimation [8]. We found that this was the case for the problems that we tested and that for most of the examples, one cycle of the finite-difference power method with a random Z_0 produced a condition estimate that was within 90 percent of the true condition number while none of the one cycle estimates was less than 25 percent of the true value.

The second method of choosing Z_0, for the exponential function, consists of setting $Z_0 \equiv (e^{X^T}e^X + e^X e^{X^T})/2$. The rationale behind this choice is that since

$$L(Z,X) = \int_0^1 e^{X(1-s)} Z e^{Xs}\, ds,$$

if we set $Z = I$, then $L(I, X) = e^X$. The adjoint step in the power method then gives

$$L(e^X, X^T) = \int_0^1 e^{X^T(1-s)} e^X e^{X^T s}\, ds \cong \frac{e^{X^T}e^X + e^X e^{X^T}}{2} = Z_0.$$

Thus one cycle of the power method with Z_0 as above has approximately the effect of two cycles and the resulting condition estimate should be much nearer the true condition number. We found that this was indeed the case and the resulting condition estimates were always better than those obtained with random matrices. A similar procedure was used for the logarithmic problem.

To determine the true condition numbers for our problem set, the trapezoid power method was iterated until the estimates from one iteration to the next had a relative difference of less than 10^{-8}. (The resulting values were cross-checked by iterating the finite-difference method.)

In Tables 1 and 2, we give the following relative condition numbers:

$$K_{\text{TRAP}} \equiv \frac{\|X\|}{\|F(X)\|} \| L(\cdot, X) \|_{\text{TRAP}}, \tag{4.1}$$

$$K_{\text{FD}} \equiv \frac{\|X\|}{\|F(X)\|} \| L(\cdot, X) \|_{\text{FD}}, \tag{4.2}$$

$$K_{\text{EXACT}} \equiv \frac{\|X\|}{\|F(X)\|} \| L(\cdot, X) \|, \tag{4.3}$$

for the exponential and logarithmic functions where $\| L(\cdot, X) \|_{\text{TRAP}}$ and $\| L(\cdot, X) \|_{\text{FD}}$ refer to the one-cycle power method estimates of $\| L(\cdot, X) \|$ obtained by using the trapezoid and finite-difference approximation methods, respectively.

The problems tested included eight examples from the standard collection of matrices [16], four examples of Ward [32]; 10 examples arising from state space models [1] in control theory [3], [7], [25], [26]; and 1,000 randomly generated matrices of orders between two and 16. For brevity, we discuss only a representative subsample consisting of six problems.

TABLE 1
Condition estimates for $F(X) = e^X$.

Problem number	K_{TRAP} (from 4.1)	K_{FD} (from 4.2)	K_{EXACT} (from 4.3)
1	7.49	7.50	7.50
2	53.9	53.9	53.9
3	2×10^4	2×10^4	2×10^4
4	1.59	1.59	1.68
5	2×10^{11}	2×10^{11}	2×10^{11}
6	3×10^3	3×10^3	3×10^3

TABLE 2
Condition estimates for $F(X) = \log X$.

Problem number	K_{TRAP} (from 4.1)	K_{FD} (from 4.2)	K_{EXACT} (from 4.3)
1	5.15	5.17	5.25
2	9×10^6	9×10^6	9×10^6
3	6×10^9	6×10^9	6×10^9
4	3.76	3.76	4.03
5	3×10^{11}	3×10^{11}	3×10^{11}
6	6×10^6	6×10^6	6×10^6

The first four problems of Tables 1 and 2 were taken from [32]. Of these, Examples 3 and 4 are interesting because they show, as noted by Ward [32], that the condition estimation scheme used in the subroutine MATEXP can give very conservative bounds. For problem 3, MATEXP predicted that not more than 12 digits of accuracy would be lost in the computation of e^X, whereas one cycle of the power method for the Fréchet derivative (see K_{TRAP} and K_{FD}, Table 1, problem 3) predicted four digits would be lost. In fact, the computed result had lost exactly four digits of accuracy. Similarly, for problem 4, MATEXP predicted a loss of at most nine digits, the power method predicted a loss of one digit, and the computed result had lost one digit of accuracy. This illustrates that condition estimates based on the norm of the Fréchet derivative have the virtue of reliability. For the fifth problem, $X = [\begin{smallmatrix} 0 & \xi \\ 0 & 0 \end{smallmatrix}]$ with $\xi = 10^6$. This value of ξ was chosen because the exponential condition number is then very large. The excellent agreement between K_{TRAP}, K_{FD}, and K_{EXACT} in Tables 1 and 2 is reminiscent of the fact that inverse power method estimates of $\|A^{-1}\|$ become more accurate as A becomes more singular.

An interesting feature of this problem is the strong dependence of K_{FD} on δ, as illustrated in Table 3. For example, $K_{\text{FD}} = 7 \times 10^{36}$ when $\delta/\|X\| = 5 \times 10^{-9}$. This seems rather conservative since $K_{\text{EXACT}} \equiv K(F, X) = 2 \times 10^{11}$. However, the given values of K_{FD} are correct and appropriate, as the following two points will make clear. First, for a given value of δ, K_{FD} is a lower bound on K_δ in (1.3):

$$K_{\text{FD}} = (\|e^{X+\delta Z} - e^X\|/\delta)(\|X\|/\|e^X\|) \leqq K_\delta.$$

Thus K_{FD} estimates K_δ rather than $K(F, X) \equiv \lim_{\delta \to 0} K_\delta$. Normally, when $\delta = 5 \times 10^{-9}\|X\|$, the difference between K_δ and $K(F, X)$ is small. However, and this is the second point, for this example, K_δ grows dramatically with δ. To see this, let $Z = [\begin{smallmatrix} 0 & 0 \\ 1 & 0 \end{smallmatrix}]$, then (after some algebra),

$$\frac{\|e^{X+\delta Z} - e^X\|}{\delta} \frac{\|X\|}{\|e^X\|} \cong \frac{\sqrt{\delta\xi}\, e^{\sqrt{\delta\xi}}}{2\delta^2} \leqq K_\delta.$$

TABLE 3
Perturbation estimates for Problem 5 with $\|X\| = 10^6$.

$\delta/\|X\|$	K_{FD}
5×10^{-9}	7×10^{36}
1×10^{-9}	9×10^{20}
5×10^{-10}	3×10^{17}
1×10^{-10}	1×10^{13}
5×10^{-11}	2×10^{12}
1×10^{-11}	3×10^{11}
5×10^{-12}	2×10^{11}

For example, if $\xi = 10^6$ and $\delta = 5 \times 10^{-9}\xi = 5 \times 10^{-3}$, then $\sqrt{\delta\xi}e^{\sqrt{\delta\xi}}/2\delta^2 = 7.24 \times 10^{36}$. In fact, for this problem, K_{FD} provides a reasonably good estimate of nonlinear, "large-scale" perturbation effects.

This points the way to choosing the right value of δ to use with K_{FD}: $\delta/\|X\|$ should be on the order of the uncertainty in the data, X, or if X is known exactly, $\delta/\|X\|$ should be near the machine epsilon, since this is the size of the error induced by machine representation. After extensive numerical testing, we found that good results were consistently obtained by taking $\delta = \varepsilon 10^3\|X\|$ where ε is the machine epsilon ($\cong 2.8 \times 10^{17}$ for double precision on a VAX 11/780). This value of δ is small enough so that $(F(X + \delta Z) - F(X))/\delta$ provides a good approximation to $L(Z, X)$, but not so small as to generate the truncation effects which occur when $\delta/\|X\|$ is at or below the machine epsilon. For extremely ill-conditioned problems (for example, problem 5 with $\xi \geqq 10^8$) even $\delta = \varepsilon 10^3\|X\|$ is too large to give a good estimate for $\|L(\cdot, X)\|$. In cases of this type, the trapezoid method provides a reliable means of estimating $\|L(\cdot, X)\|$ (see Corollary 3.5) since it does not depend on δ.

The last problem (#6) in Tables 1 and 2 is taken from [26] and illustrates the fact that condition estimates based on upper triangular canonical forms can be extremely conservative. For this problem,

$$X = \begin{bmatrix} 48 & -49 & 50 & 49 \\ 0 & -2 & 100 & 0 \\ 0 & -1 & -2 & 1 \\ -50 & 50 & 50 & -52 \end{bmatrix}.$$

Let $X = SJS^{-1}$ where J is the Jordan form of X. Petkov, Christov, and Konstantinov [26] show that the Jordan decomposition bound,

$$\frac{\|e^{X+\delta Z} - e^X\|_2}{\delta}\frac{\|X\|_2}{\|e^X\|_2} \leqq 16\delta\|S\|_2^2\|S^{-1}\|_2^2 e^{4\delta\|S\|_2\|S^{-1}\|_2}\ \frac{\|X\|_2}{\|e^X\|_2},$$

gives

$$K_\delta(F, X) \leqq 4 \times 10^{104}$$

for $\delta = 4 \times 10^{-3}$. However, Lyapunov arguments can be given to show that $K_\delta(F, X) \leqq 2 \times 10^6$; a lower bound for $K_\delta(F, X)$ is given by $K_{\mathrm{FD}} = 3.4 \times 10^3$ for $\delta = 4 \times 10^{-3}$.

5. Conclusion. The natural connection between the Fréchet derivatives of matrix functions and sensitivity allows us to develop a very general condition estimation procedure based on finite-difference approximations. This procedure is computationally reasonable since it only requires two extra function evaluations. As seen in the section on numerical tests, the ability to manipulate the "stepsize" δ in the finite-difference method can lead to sensitivity estimates even when the size of the perturbation is relatively large (see Table 3, § 4). This area needs further research, as does the related problem of condition estimation for perturbations that are restricted in some way, as in the theory of structured singular values.

We have also presented an alternative sensitivity estimation procedure for the matrix exponential and logarithmic functions. This method is based on a trapezoid approximation of the integral representation of the Fréchet derivative of the exponential function.

Because of its form, this method dovetails nicely with the "scaling and squaring" method of evaluating the matrix exponential and the "inverse scaling and squaring" method of evaluating the logarithm of a matrix. Both the finite-difference and trapezoid approaches require almost the same effort computationally. However, the trapezoid method has an advantage in that it does not depend on the stepsize δ, and consequently is a more reliable method for estimating the norm of the Fréchet derivative when the matrix function is very ill-conditioned, as in Example 5.

Appendix A. The square root and logarithm of a matrix. In this Appendix, we show that any real matrix $A \in \mathbb{R}^{p\times p}$, with no eigenvalues on the negative real axis including zero, has a unique real square root and a unique real logarithm. We also justify the inverse scaling and squaring formula $\log A = 2^n \log A^{1/2^n}$.

LEMMA A1. *Let $A \in \mathbb{R}^{p\times p}$ with no eigenvalues on the negative real axis, including zero. Then there exists a unique real matrix X, such that we have the following*:

(A1) (1) $X^2 = A$,

(A2) (2) *The eigenvalues of X are restricted to the sector* $-\pi/2 < \arg(z) < \pi/2$.

Proof. The existence of such a matrix X follows from the Cauchy integral formula for operators. This method was used by DePrima and Johnson in [11], in which this lemma was proved under the added condition that X satisfies: (3) $XS = SX$ whenever $AS = SA$. However, this condition can be shown to be a consequence of (A1) and (A2). □

LEMMA A2. *Let $A \in \mathbb{R}^{p\times p}$ with no eigenvalues on the negative real axis, including zero. Then there exists a unique real matrix X, such that we have the following*:

(A3) (1) $e^X = A$,

(A4) (2) *The eigenvalues of X lie in the strip* $-\pi < \mathrm{Im}(z) < \pi$.

Proof. The proof is similar to that of Lemma 6.1. □

LEMMA A3. *Let $A \in \mathbb{R}^{p\times p}$ with no eigenvalues on the negative real axis. Let $A^{1/2}$ and* $\log A$ *denote the unique real square root and logarithm of A as in Lemmas* A1 *and* A2, *respectively. Then*

(A5) $$A^{1/2} = e^{1/2 \log A}$$

and

(A6) $$\log A = 2 \log A^{1/2}.$$

Proof. Let $X = \log A$ satisfy (6.3) and (6.4). Then $e^{X/2}$ satisfies $e^{X/2}e^{X/2} = e^X = A$ and the eigenvalues of $e^{X/2}$ lie in the sector $-\pi/2 < \arg(z) < \pi/2$. This means that the real matrix $e^{X/2}$ satisfies (A1) and (A2) and so by Lemma A1, $A^{1/2} = e^{X/2} = e^{1/2 \log A}$, which proves (A5).

Now suppose that $\hat{X} = A^{1/2}$ satisfies (A1) and (A2). Using the Cauchy integral operator representation of the logarithm of $A^{1/2}$, we see that the eigenvalue condition (A2) implies that the eigenvalues of $\log A^{1/2}$ lie in the strip $-\pi/2 < \mathrm{Im}(z) < \pi/2$. Thus the matrix $X \equiv 2 \log A^{1/2}$ satisfies (A3) and (A4) and must be equal to $\log A$ by Lemma A2. This proves (A6). □

COROLLARY A4. *For A as in Lemma* A3, *the "inverse scaling and squaring" formula* $\log A = 2^n \log A^{1/2^n}$ *is valid.*

Proof. By Lemma A3, $\log A = 2 \log A^{1/2} = 4 \log A^{1/4} = \cdots = 2^n \log A^{1/2^n}$. □

Appendix B. Examples of Fréchet derivatives. The following lemma enables us to find the derivatives of the square root and logarithmic functions.

LEMMA B1. *Let F be diffeomorphic at X, that is, let F be invertible in a neighborhood of $Y \equiv F(X)$ and let the derivative, $L_F(\cdot, X)$, of F at X be nonsingular. Then the derivative, $L_{F^{-1}}(\cdot, Y)$ of F^{-1} at Y exists and is given by the inverse of the derivative of F at X*:

$$L_{F^{-1}}(\cdot, F(X)) = L_F^{-1}(\cdot, X).$$

Proof. Although the proof of this lemma is not hard, we omit it for the sake of brevity. □

Using this lemma, we may find the derivatives of the inverse of the functions considered in Examples 1 and 2 of the introductory section.

Example 3. Let $A \in \mathbb{R}^{p\times p}$ with no eigenvalues on the negative real axis including zero, and let $A^{1/2}$ denote the square root of A as in Lemma A1. Condition (A2) on the eigenvalues of $A^{1/2}$ ensures that the Sylvester operator $L(Z) = A^{1/2}Z + ZA^{1/2}$ is invertible [21]. Hence, the derivative, $L_{1/2}$ of the square root function $A \rightarrow A^{1/2}$ is the inverse of L in Example 1 of the Introduction: $L_{1/2}(W, A) = Z$, where $L(Z, A^{1/2}) = W$.

Example 4. Let $A \in \mathbb{R}^{p\times p}$ with no eigenvalues on the negative real axis including zero, and let $\log A$ denote the logarithm of A as in Lemma A2. Condition (A4) ensures that the exponential derivative operator, L, defined by (1.6) is invertible. Hence, the derivative, $L_{\log}$, of the logarithmic function $A \rightarrow \log A$ is the inverse of L: $L_{\log}(W, A) = Z$ where $L(Z, \log A) = W$. (See § 3 for more details.)

Example 5. Let $X \in \mathbb{R}^{p\times p}$ be invertible. Then

$$(X+\delta Z)^{-1} = X^{-1} - \delta X^{-1}ZX^{-1} + O(\delta^2),$$

so the derivative of the inverse function is given by $L(Z, X) = -X^{-1}ZX^{-1}$. It is interesting to note that the inverse function is invariant under the inversion operation and

$$L^{-1}(\cdot, X) = L(\cdot, X^{-1}).$$

Since the squaring and exponential functions are related via the identity $e^X = (e^{X/2})^2$, it is not surprising that there exists a chain rule relationship between their derivatives:

$$L_{\exp}(Z,X) = \tfrac{1}{2}L_s(L_{\exp}(Z,X/2), e^{X/2})$$

where L_s and $L_{\exp}$ denote the derivatives of the squaring and exponential functions, respectively. This relationship is a consequence of the following lemma.

LEMMA B2. *Let $F(X) \equiv g(f(X))$ where we assume that the derivatives of f and g exist at X and $Y = f(X)$, respectively. Then the derivative of F at X exists and is given by*

$$L_F(Z,X) = L_g(L_f(Z,X), Y)$$

where L_f, L_g, and L_F denote the derivatives of f, g, and F, respectively.

Proof. The proof follows rather easily from (1.5). □

Example 6. Let $A \in \mathbb{R}^{p\times p}$ with no eigenvalues on the negative real axis including zero. Then we may define a real qth power of A, say $X = A^q$ by setting $X = e^{q \log A}$. We may write $A^q = h(g(f(A)))$ where $f(A) = \log A$, $g(B) = qB$ and $h(C) = e^C$. Then the derivative, L_q, of the map $A \rightarrow A^q$ is given by $L_q(Z, A) = L_{\exp}(qL_{\log}(Z, A), q \log A) = qL_{\exp}(L_{\log}(Z, A), q \log A)$.

Note added in proof. We wish to thank N. J. Higham for pointing out to us that our method for computing a matrix square root based on [5], while arrived at independently, is essentially identical to that given in [19]. The latter's much more thorough analysis should be consulted for details.

REFERENCES

[1] B. D. O. ANDERSON AND J. B. MOORE, *Linear Optimal Control*, Prentice-Hall, Englewood Cliffs, NJ, 1971.

[2] P. ANSELONE, *Collectively Compact Operator Approximation Theory*, Prentice-Hall, Englewood Cliffs, NJ, 1971.

[3] W. F. ARNOLD, *Numerical solution of algebraic Riccati equations*, Ph.D. thesis, University of Southern California, Los Angeles, CA, December 1983.

[4] J. BELINFANTE AND B. KOLMAN, *A Survey of Lie Groups and Lie Algebras with Applications and Computational Methods*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1972.

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

[6] R. BYERS, A LINPACK-style condition estimator for the equation $AX - XB^T = C$, IEEE Trans. Automat. Control, 29 (1984), pp. 926-928.

[7] J. C. CHUNG AND E. Y. SHAPIRO, *Constrained eigenvalue/eigenvector assignment—application to flight control systems*, in Proc. Conference on Information and Systems, Department of Electrical Engineering and Computer Science, Princeton University, Princeton, NJ, 1982.

[8] A. CLINE, C. B. MOLER, G. W. STEWART, AND J. H. WILKINSON, *An estimate for the condition number of a matrix*, SIAM J. Numer. Anal., 16 (1979), pp. 368-375.

[9] W. CULVER, *On the existence and uniqueness of the real logarithm of a matrix*, Proc. Amer. Math. Soc., 17 (1966), pp. 1146-1151.

[10] G. DAHLQUIST AND A. BJÖRCK, *Numerical Methods*, Prentice-Hall, Englewood Cliffs, NJ, 1974.

[11] C. DEPRIMA AND C. JOHNSON, *The range of* $A^{-1}A^*$ *in* GL $(n, \mathbb{C})$, Linear Algebra Appl., 9 (1974), pp. 202-222.

[12] N. DUNFORD AND J. SCHWARTZ, *Linear Operators, Part* I: *General Theory*, 4th ed., John Wiley, New York, 1967.

[13] F. GANTMACHER, *The Theory of Matrices, Vol.* I, Chelsea, New York, 1959.

[14] G. GOLUB AND C. VAN LOAN, *Matrix Computations*, Johns Hopkins University Press, Baltimore, MD, 1983.

[15] A. GRAHAM, *Kronecker Products and Matrix Calculus with Applications*, John Wiley, New York, 1981.

[16] R. GREGORY AND D. KARNEY, *A Collection of Matrices for Testing Computational Algorithms*, John Wiley, New York, 1969.

[17] F. HAUSDORFF, *Die Symbolische Exponential formel in der Gruppentheorie*, Berichte der Sächsischen Akademie der Wissenschaften (Math. Phys. Klasse), Leipzig, Vol. 58, 1906, pp. 19-48.

[18] B. HELTON, *Logarithms of matrices*, Proc. Amer. Math. Soc., 19 (1968), pp. 733-738.

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

[20] C. KENNEY AND A. J. LAUB, *Padé error estimates for the logarithm of a matrix*, Internat. J. Control, to appear, 1989.

[21] P. LANCASTER, *Explicit solutions of linear matrix equations*, SIAM Rev., 12 (1970), pp. 554-566.

[22] B. LEVINGER, *The square root of a* 2×2 *matrix*, Math. Mag., 53 (1980), pp. 222-224.

[23] W. MAGNUS, *On the exponential solution of differential equations for a linear operator*, Comm. Pure and Appl. Math., 7 (1954), pp. 649-673.

[24] C. B. MOLER AND C. VAN LOAN, *Nineteen dubious ways to compute the exponential of a matrix*, SIAM Rev., 20 (1978), pp. 801-836.

[25] B. C. MOORE AND A. J. LAUB, *Computation of supremal* (A, B)*-invariant and controllability subspaces*, IEEE Trans. Automat. Control, 23 (1978), pp. 783-792.

[26] P. PETKOV, N. D. CHRISTOV, AND M. M. KONSTANTINOV, *Computational methods for linear control systems—some open questions*, in Proc. 26th Conference on Decision and Control, Los Angeles, CA, 1987, pp. 818-823.

[27] S. PUTHENPURA AND N. SINHA, *Transformation of continuous-time model of a linear multivariable system from its discrete-time model*, Electronics Letters, 20 (1984), pp. 737-738.

[28] J. RICE, *A theory of condition*, SIAM J. Numer. Anal., 3 (1966), pp. 287-310.

[29] N. SINHA AND G. LASTMAN, *Transformation algorithm for identification of continuous-time multivariable systems from discrete data*, Electronics Letters, 17 (1981), pp. 779-780.

[30] B. T. SMITH, J. M. BOYLE, J. J. DONGARRA, B. S. GARBOW, Y. IKEBE, V. C. KLEMA, AND C. B. MOLER, *Matrix Eigensystem Routines*—EISPACK *Guide*, 2nd ed., Lecture Notes in Computer Science 6, Springer-Verlag, New York, 1976.

[31] C. VAN LOAN, *The sensitivity of the matrix exponential*, SIAM J. Numer. Anal., 14 (1977), pp. 971-981.

[32] R. C. WARD, *Numerical computation of the matrix exponential with accuracy estimate*, SIAM J. Numer. Anal., 14 (1977), pp. 600-610.

[33] A. WOUK, *Integral representation of the logarithm of matrices and operators*, J. Math. Anal. Appl., 11 (1965), pp. 131-138.