ON THE LIMITATION AND APPLICATION OF PADÉ APPROXIMATION TO THE MATRIX EXPONENTIAL

Charles Van Loan

Non-normality in the matrix A and its effect on Padé approximation of the matrix exponential is discussed. Against this background we remark upon the selection and efficient evaluation of the appropriate approximant. An application from control theory serves to illustrate some of the principles discussed.

1 The Limitations of Padé Approximation

We shall consider the problem of computing the exponential e^A of an n-by-n matrix A. Here n is small enough such that the explicit formation of e^A is feasible. One approach to this problem is to compute an eigenvalue decomposition $A = XBX^{-1}$ and then invoke the formula $e^A = Xe^BX^{-1}$. Parlett [5] has detailed a technique of this type based upon the Schur decomposition

$$(1.1) \quad Q^*A\,Q = \mathrm{diag}(\lambda_1,\ldots,\lambda_n) + N$$

Here, Q is unitary, $N = (n_{ij})$ is strictly upper triangular ($n_{ij} = 0$, $i \geq j$), and $\lambda(A) = \{\lambda_1,\ldots,\lambda_n\}$ is the spectrum of A. If A is normal ($A^*A = AA^*$), then it is easy to verify that $N = 0$. In this case the algorithm is particularly easy to implement and the computed e^A extremely accurate. However, as with all eigenvalue methods for computing the exponential, serious numerical difficulties can arise when A is a non-normal matrix having confluent or nearly confluent eigenvalues [4].

A desire to avoid these difficulties accounts, in part, for the attractiveness of using Padé approximants $R_{pq}(A) = [D_{pq}(A)]^{-1}N_{pq}(A)$ where

$$N_{pq}(z) = \sum_{j=0}^{p} \frac{(p+q-j)!p!}{(p+q)!j!(p-j)!} z^j$$

$$D_{pq}(z) = \sum_{j=0}^{q} \frac{(p+q-j)!q!}{(p+q)!j!(q-j)!} (-z)^j$$

However, even though no eigenvalue computations are required when these appróximants are used, the effects of non-normality and confluence can be present. If

$$A = \begin{bmatrix} 0 & 6 & 0 & 0 \\ 0 & 0 & 6 & 0 \\ 0 & 0 & 0 & 6 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

then $\| R_{11}(A) - e^A \| = 18$ even though $R_{11}(z)$ approximates the exponential exactly on the spectrum of A. (Here, as elsewhere, $\| \cdot \|$ denotes the 2-norm.)

In general, the bounds on the accuracy of an approximation $f(A)$ to e^A deteriorate with loss of normality and not just eigenvalue confluence. We refer the reader to the analysis of Wragg and Davies [11]. which makes this point clear. Their results are derived through exploitation of the Jordan canonical form. In [7] the author derived bounds on $\| f(A) - e^A \|$ which involved powers of $\| N \|$, where N is the matrix of (1.1). The size of N is related to Henrici's "departure from normality" [3]. Of course, if A is normal and $f(z)$ is defined on $\lambda(A)$, then it is easy to verify from (1.1) that

$$\| f(A) - e^A \| = \max_{z \in \lambda(A)} | f(z) - e^z |$$

This illustrates that for normal A, the accuracy of $f(A)$ depends solely upon the behavior of $f(z)$ on the spectrum of A.

All these remarks raise the possibility that the problem of computing e^A is inherently difficult for certain non-normal matrices. In [8] we examined the sensitivity of the map $A \to e^{At}$ in hopes of answering this question. Various upper bounds to the relative perturbation

$$\phi(t) = \frac{\| e^{(A+E)t} - e^{At} \|}{\| e^{At} \|}$$

were derived. For example, it was shown that

$$\phi(t) \leq t \| E \| M_S(t)^2 e^{M_S(t) \| E \| t}$$

where

$$M_S(t) = \sum_{k=0}^{n-1} \| Nt \|^k / k!$$

and N is the matrix of (1.1). This bound is typical in that it "deteriorates" as A departs from normality.

More light is shed on the sensitivity problem through the formulation of the exponential "condition number"

$$\nu(A,t) = \max_{\|E\| = 1} \left\| \int_0^t e^{A(t-s)} E e^{As} ds \right\| \frac{\| A \|}{\| e^{At} \|} .$$

This quantity amounts to a normalized Frechet derivative of the map $A \to e^{At}$. One can show that for a

a given $t \geq 0$ there exists a perturbation $\hat{E}$ such that

$$\phi(t) \cong \frac{\|\hat{E}\|}{\|A\|} \nu(A,t) .$$

This indicates that if $\nu(A,t)$ is large, as it tends to be when A is non-normal, then small relative changes in A can induce relatively large changes in e^{At}. We refer the interested reader to [8].

These observations must be borne in mind when assessing algorithms for computing e^A. In general, rounding errors of the order $\varepsilon \nu(A,1)$ can be expected in the computed version of e^A no matter what algorithm is used and, hence, no technique should be faulted for producing errors of this magnitude. Here, ε is the machine precision. The situation is analogous to that of solving linear systems $Ax = b$. In this setting even a "stable" algorithm such as Gaussian elimination with pivoting need not produce an accurate solution when A is ill-conditioned with respect to inversion [2].

Are methods involving Padé approximation to e^A stable in the sense of Gaussian elimination or do they introduce errors greater than the inherent sensitivity of the problem warrants? We know of no rigorous analysis which answers this question. However, it is clear that the success of a particular Padé technique depends a great deal upon the details of the implementation. To illustrate this point let us consider the following family of approximations to e^A :

$$F(A,p,q,j) = \left[R_{pq}(A/2^j)\right]^{2^j}$$

Usually, $\|A\|/2^j \approx 1$. If A is non-normal and $\lambda(A)$

is in the open left-half-plane, then it is possible for $\|e^A\| \approx \|F(A,p,q,j)\|$ to very small while

$$\left\| e^{A/2^{j-k}} \right\| \approx \left\| \left[R_{pq}\left(\frac{A}{2^j}\right) \right]^{2^k} \right\|$$

is very big for some intermediate value of k, $1 \leq k < j$. This is known as the "hump" phenomena [4]. Since the rounding errors in the computed $F(A,p,q,j)$ may be of the order

$$\varepsilon \max_{1 \leq k \leq j} \left\| \left[R_{pq}\left(\frac{A}{2^j}\right) \right]^{2^k} \right\|$$

where ε is the machine precision, the resulting approximate to e^A may have large relative error.

If instead we estimate e^A with an approximation of the form $e^{-\lambda}F(A+\lambda I,p,q,j)$ where $\lambda \geq \|A\|$, then the effect of the "hump" is somewhat dissipated. This is because the smallness in the approximation to e^A is achieved through the "stable" scalar-matrix multiplication $e^{-\lambda}F(A+\lambda I,p,q,j)$ rather than through the subtractive cancellation which may transpire during the repeated squaring of $R_{pq}(A/2^j)$. (Subtractive cancellation can cause a severe loss of numerical accuracy.)

Ward [10] advocates translation by λ, not to minimize cancellation, but to draw the eigenvalues of $A + \lambda I$ closer to the origin. For this reason he sets $\lambda = -\text{trace}(A)/n$. This choice of λ, or any other, undoubtedly effects the accuracy of the computed approximant to e^A. Yet, it is interesting to note that Ward's rigorous and very useful error analysis does not in any way explicitly exploit the fact that a translation has been done. This reminds us once again that there is no rigorous analysis relating the accur-

acy of the computed version of $e^{-\lambda}F(A+\lambda I,p,q,j)$ to $\nu(A,1)$, non-normality, the "hump", etc. Until such an analysis is done, the precise limitations which these factors impose on Padé approximation will remain unclear.

2 The Selection and Evaluation of F(A,p,q,j)

Let us now turn our attention to some of the practical problems which arise in connection with F(A,p,q,j). (Assume A is translated if necessary.) In [4] it was shown that if

$$\frac{\|A\|}{2^j} \leq \frac{1}{2}$$

then

$$F(A,p,q,j) = e^{A+E}$$

where

$$\frac{\|E\|}{\|A\|} \leq \frac{8\,\|A\|^{p+q}}{2^{j(p+q)}}\,\frac{p!\,q!}{(p+q)!(p+q+1)!} \equiv \varepsilon(p,q,j). \tag{2.1}$$

F(A,p,q,j) costs approximately w(p,q,j) multiplicative operations where

$$w(p,q,j) = n^3\,[j + \max\{p,q\} + \tfrac{1}{3} - \tfrac{4}{3}\,\delta_{oq}]$$

This follows since one must form $A^2,\ldots,A^{\max\{p,q\}}$; solve the linear system $[D_{pq}(A/2^j)]X = N_{pq}(A/2^j)$; and compute X^{2^j}. (We assume that the linear system is solved using Gaussian elimination. Of course, there is no linear system if $q = 0$.)

Now consider the problem of choosing p,q, and j such that for a given $\varepsilon > 0$,

$$\frac{\|E\|}{\|A\|} \leq \varepsilon(p,q,j) \leq \varepsilon$$

Clearly, there are many p,q, and j for which this inequality holds, but it makes economic sense to select these parameters in such a way that $w(p,q,j)$ is minimized. It turns out that for practical values of $\|A\|$ and ε we need only consider diagonal approximants $(p=q)$. The simplified form of $\varepsilon(q,q,j)$ and $w(q,q,j)$ makes it easy to determine an "optimum" q and j [4]. For example, if $\|A\| = 100$ and $\varepsilon = 10^{-6}$, then $F(A,3,3,8)$ is the most efficient approximant under this method of assessing work.

We briefly mention that the operation count $w(q,q,j) = n^3[j + q + \frac{1}{3}]$ can be somewhat reduced if the special structure of $R_{qq}(B)$ is exploited. $(B = A/2^j.)$ To illustrate, suppose $q = 2r+1$ where r is a positive integer and note

$$R_{qq}(B) = \frac{\sum_{k=0}^{r} c_{2k}(B^2)^k + B\sum_{k=0}^{r} c_{2k+1}(B^2)^k}{\sum_{k=0}^{r} c_{2k}(B^2)^k - B\sum_{k=0}^{r} c_{2k+1}(B^2)^k}$$

where $c_k = (2q-k)!k!/(2q)!k!(q-k)!$. It is obvious that to evaluate $R_{qq}(B)$, one need only compute B^2, $(B^2)^2,\ldots,(B^2)^r$; do a matrix multiplication by B; and solve a linear system. Thus, $F(A,q,q,j)$ can be evaluated in approximately $n^3[j + \frac{q}{2} + \frac{4}{3}]$ operations rather than $n^3[j + q + \frac{1}{3}]$ operations as predicted above. We refer the reader to [6] for apt remarks concerning the efficient evaluation of matrix polynomials. Of course, the modified work functions which arise from "fast" schemes for evaluating the numerator and denominator can be used to determine "optimum" q and j just as in [4].

3 Approximating Integrals Involving e^A

In the course of solving the optimal linear regulator problem with step input, one is led to the following integrals:

$$(3.1)\quad H(\Delta) = \int_0^\Delta e^{As} B\, ds$$

$$(3.2)\quad Q(\Delta) = \int_0^\Delta e^{A^T s} Q_c e^{As}\, ds$$

$$(3.3)\quad M(\Delta) = \int_0^\Delta e^{A^T s} Q_c H(s)\, ds$$

$$(3.4)\quad W(\Delta) = \int_0^\Delta H(s)^T Q_c H(s)\, ds$$

Here, $\Delta > 0$, Q_c is a symmetric, positive definite nxn matrix, and B is an nxp matrix with $n \geq p$.

These integrals can be expressed as power series in Δ with computable matrix coefficients. One method for approximating (3.1)-(3.4) is to evaluate truncated versions of these series. See [1] and the references therein. In [9] we noted that if

$$C = \begin{bmatrix} -A^T & I & 0 & 0 \\ 0 & -A^T & Q_c & 0 \\ 0 & 0 & A & B \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

and

$$e^{C\Delta} = \begin{bmatrix} F_1(\Delta) & G_1(\Delta) & H_1(\Delta) & K_1(\Delta) \\ 0 & F_2(\Delta) & G_2(\Delta) & H_2(\Delta) \\ 0 & 0 & F_3(\Delta) & G_3(\Delta) \\ 0 & 0 & 0 & F_4(\Delta) \end{bmatrix}$$

then

$$H(\Delta) = G_3(\Delta)$$
$$Q(\Delta) = F_3(\Delta)^T G_2(\Delta)$$
$$M(\Delta) = F_3(\Delta)^T H_2(\Delta)$$
$$W(\Delta) = [B^T F_3(\Delta)^T K_1(\Delta)] + [B^T F_3(\Delta)^T K_1(\Delta)]^T$$

These results follow by noting that $F_3(\Delta) = e^{A\Delta}$ and that the submatrices which "make up" the upper triangle of $e^{C\Delta}$ are various convolutions. For example,

$$G_3(\Delta) = \int_0^\Delta e^{-A^T(\Delta-s)} Q_c e^{As}\, ds$$

These observations suggest that approximations H, Q, M, and W to $H(\Delta)$, $Q(\Delta)$, $M(\Delta)$, and $W(\Delta)$ can be obtained by approximating $e^{C\Delta}$ by $F(C\Delta,q,q,j)$. The computational process involves selecting q and j (say by the methods of §2), forming $F(C\Delta,q,q,j)$, and then combining its various submatrices. The errors in the resulting approximations are easy to bound by using (2.1).

If the problem is of low enough dimension, then an easy course of action is to just input $C\Delta$ to any matrix exponential program such as Ward's [10]. If efficiency is of interest then the special structure of $C\Delta$ had better be exploited. Some of the ways this can be done are detailed in [9] .

References

1 Armstrong, E.S. and A.K. Caglayan, An algorithm for the weighting matrices in the sampled-data optimal linear regulator problem, NASA Technical Note, TN D-8372, 1976.

2 Forsythe, G.E. and C.B. Moler, Computer Solution of Linear Algebraic Systems, Prentice Hall, Englewood Cliffs, New Jersey, 1967.

3 Henrici, P., Bounds for iterates, inverses, spectral variation and fields of values of non-normal matrices, Numerische Math., 4(1962), 24-40.

4 Moler, C.B. and C. Van Loan, Nineteen ways to compute the exponential of a matrix, Cornell Computer Science Technical Report TR-76-283, 1976.(To appear SIAM Review.)

5 Parlett, B.N., Computation of functions of triangular matrices, Memo No. ERL-M481, Electronics Research Laboratory, University of California, Berkeley, 1974.

6 Paterson, M.S. and L.J. Stockmeyer, On the number of nonscalar multiplications necessary to evaluate polynomials, SIAM J.Comp., 2(1973), 60-66.

7 Van Loan, C, A Study of the matrix exponential, University of Manchester Numerical Analysis Report 7, 1974.

8 Van Loan, C, The sensitivity of the matrix exponential, Cornell Computer Science Technical Report TR 76-270, 1976.(To appear in SIAM J.Num. Analysis.)

9 Van Loan, C., Computing integrals involving the matrix exponential, Cornell Computer Science Technical Report TR 76-298, 1976.

10 Ward, R.C., Numerical Computation of the matrix exponential with accuracy estimate, Union Carbide Corp. Nuclear Division Technical Report UCCND CSD 24, Knoxville Tennessee, 1975.

11 Wragg, A., and C. Davies, Computation of the exponential of a matrix I: theoretical considerations, JIMA, 11(1973), 369-375

C.Van Loan*
Dept.Computer Science
Cornell, Ithaca, NY, 14853

*Supported by NSF grant MCS76-08686.