How large is the exponential of a banded matrix?

Arieh Iserles[1]

Abstract

Let A be a banded matrix of bandwidth $s \geq 1$. The exponential of A is usually dense. Yet, its elements decay very rapidly away from the diagonal and, in practical computation, can be set to zero away from some bandwidth $r > s$ which depends on the required accuracy threshold. In this paper we investigate this phenomenon and present sharp, computable bounds on the rate of decay.

1 Sparsity and matrix exponents

How large is a matrix exponential? And why should it matter in practical numerical calculations?

On the face of it, the first question is sheer nonsense. Let $A \in \mathrm{M}_n(\mathbb{R})$, the set of $n \times n$ real matrices. By an assiduous choice of A, the entries of

$$\mathrm{e}^A = \sum_{m=0}^{\infty} \frac{1}{m!} A^m$$

can be made as large (or as small) as we wish them to be (Horn & Johnson 1985). In particular, even if A is sparse, e^A is likely to be a dense matrix. Yet, it is the contention of the present paper that this is a simplistic point of view and that, provided that A is a banded matrix, e^A is itself within an exceedingly small distance from a banded matrix.

In Figure 1 we have performed the following exercise. 10^5 tridiagonal matrices $A_1, A_2, \ldots, A_{10^5} \in \mathrm{M}_{50}(\mathbb{R})$ were chosen at random, with nonzero entries distributed uniformly in $[-1, 1]$. The entries of the matrix E, plotted at the upper left corner of Figure 1, are the maximum across the relevant entries of e^{A_k}, $k = 1, 2, \ldots, 10^5$. (Throughout this paper, 'plotting a matrix' means that, using the `MATLAB` function `mesh`, we display a matrix as a three-dimensional surface: this allows, at a glance, to take heed of the size of its elements.) It can be observed at once that the entries of E decay rapidly away from the diagonal. This is confirmed in the figure at the upper right corner, which displays $\log_{10} |E_{k,l}|$, $k, l = 1, 2, \ldots, 50$. The decay away from the diagonal is evidently faster than exponential. Although E is formally dense, its entries are *very* small away from a narrow band along the diagonal. Further affirmation is provided by the bottom two figures, which display the vector $\{E_{25+k,25-k} : k = -24, \ldots, 24\}$ to an absolute (on the left) and base-10 logarithmic (on the right) scale. The reason for base-10 logarithms is that they demonstrate, at a single glance, the number of decimal digits. Thus, $|E_{1,49}| \approx 10^{-80}$ and $|E_{20,30}| \approx 10^{-8}$. Given that E represents a (statistical) upper bound on the size of the exponential and bearing in mind that numerical calculations are invariably required to finite accuracy, almost

[1]Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Silver Street, Cambridge CB3 9EW, England, email `ai@damtp.cam.ac.uk`.

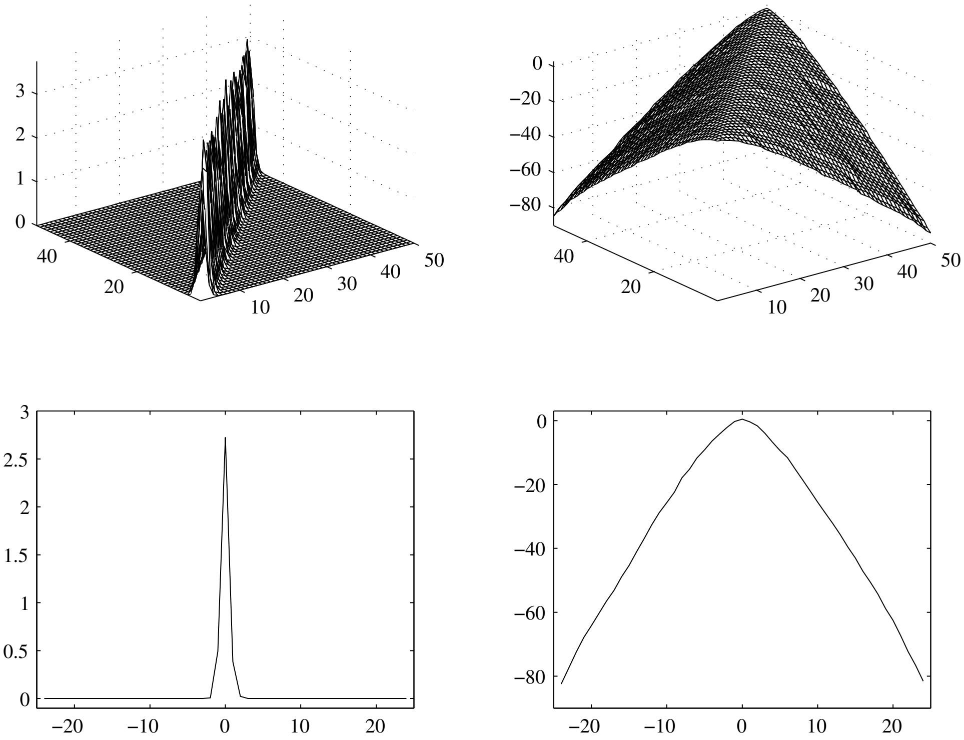


Figure 1: Exponentials of a tridiagonal matrix.

always far coarser than 10^{-80}, it follows that each e^{A_k} can be approximated very well by a banded matrix!

This is perhaps not as surprising as it may appear at first, since the growth in the (k, l) entry of e^A is governed by a competition between the growth in the entries of A^m and the decay of $1/m!$. Since factorials ultimately overtake powers, one can expect decay to occur. Yet, the precise rate of decay is of interest and its magnitude is, we believe, surprising.

Lest there is an impression that this behaviour is specific to tridiagonal matrices, Figure 2 provides similar data for quindiagonal (i.e., with five nonzero diagonals) matrices. The decay away from the diagonal is less rapid than in the tridiagonal case, yet fast enough for our observations to retain their validity.

Of course, there is a great deal of difference between experiments with random matrices and rigourous mathematical statements. In this paper we prove that the behaviour indicated in Figures 1 and 2 is indicative of exponentials of *all* banded matrices. Moreover, tight upper bounds on the worst-possible rate of decay away from the exponential can be established for every bandwidth. In other words, given an arbitrary matrix $A \in \mathrm{M}_n(\mathbb{R})$ of bandwidth s and such that $\rho = \max_{k,l=1,2,\ldots,n} |A_{k,l}|$, for every *threshold* $\varepsilon > 0$ we can determine (as a function of s and ρ but, interestingly

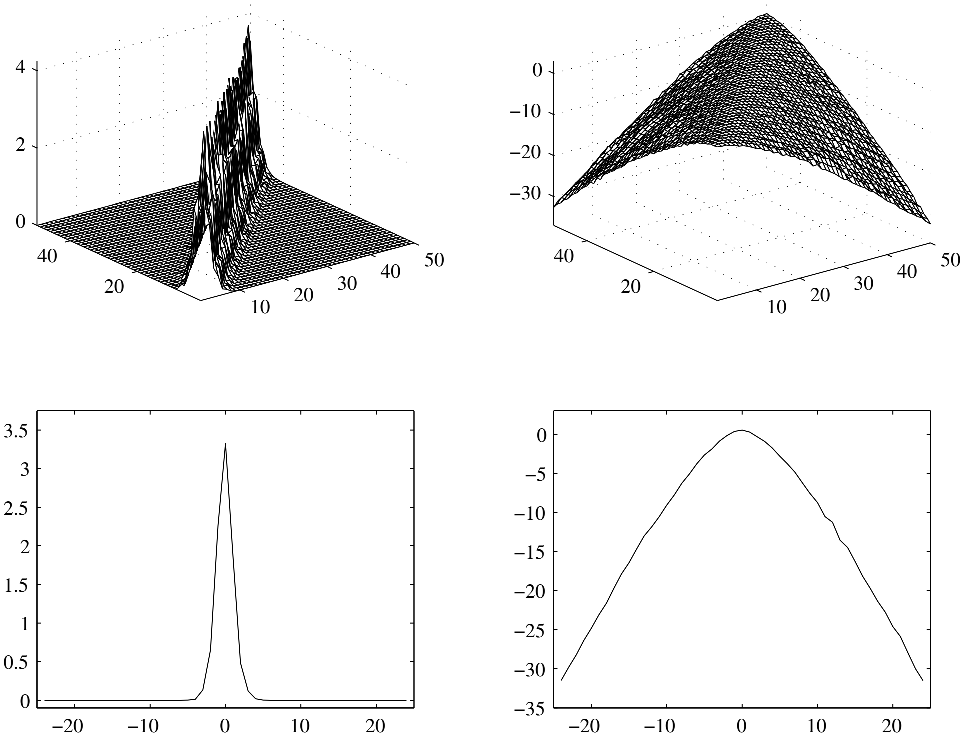


Figure 2: Exponentials of a quindiagonal matrix.

enough, not of n) an integer r such that $|E_{k,l}| < \varepsilon$ for all $|k-l| > r$, where $E = \mathrm{e}^A$.

Why should all this matter, except for an intrinsic mathematical interest? Classical methods for the evaluation of a matrix exponential cannot take advantage of the above phenomenon. Such methods typically can be classified into three categories:

1. **Rational approximants:** The exponential is replaced by a rational function, $\mathrm{e}^z \approx r(z) := p(z)/q(z)$, where $p \in \mathbb{P}_\alpha$, $q \in \mathbb{P}_\beta$, $q(0) = 1$ and $\mathrm{e}^z - p(z)/q(z)$ is in some sense 'small'. Typical measure of smallness is the order of approximation at the origin, which results in *Padé approximants* and their modifications (Baker 1975, Moler & Van Loan 1983). Other criteria of smallness can be applied, e.g. minimisation with respect to the L_∞ norm in some subset of the complex plane or interpolation away from the origin (Iserles & Nørsett 1991). Thus, instead of evaluating e^A, one computes two matrix-valued polynomials, $p(A)$ and $q(A)$, and inverts the latter to obtain $r(A)$.

   Suppose that A is banded. Then, using the approach of this paper, it is possible to show that $r(A)$ is near a banded matrix, although less near than in the exponential case. Yet, it is not clear at all how this behaviour can be exploited in the design and implementation of effective numerical methods for the exponentiation of banded matrices.

2. **Krylov subspace methods:** The underlying idea is to seek an approximant to $\mathrm{e}^A \boldsymbol{v}$, where $A \in \mathrm{M}_n(\mathbb{R})$ and $\boldsymbol{v} \in \mathbb{R}^n$, from the r-dimensional *Krylov subspace* $K_{n,r} = \mathrm{Span}\{\boldsymbol{v}, A\boldsymbol{v}, \ldots, A^{r-1}\boldsymbol{v}\}$ (Hochbruck & Lubich 1997, Tal Ezer & Kosloff 1984). It is possible to show that surprisingly good and robust approximants can be obtained for relatively modest values of r.

   It is not transparent at all how, for banded A, the nearness of e^A to a banded matrix can be exploited for an improved Krylov-subspace approximation of the exponential. The striking efficiency of Krylov subspace methods hinges on the fact that $K_{n,r}$ retains, in a well-understood sense, much of the spectral information of the matrix A, whilst cutting the dimension. This has no apparent connection to the size of the entries of e^A.

3. **Methods of numerical linear algebra:** The conceptually simplest means of evaluating e^A is by spectral decomposition: if $A = VDV^{-1}$, where D is diagonal, then $\mathrm{e}^A = V\mathrm{e}^D V^{-1}$. This, however, is not a viable technique for general matrices. In place of a spectral decomposition, one can factorize A in a different form, e.g. into a Schur decomposition (Golub & Van Loan 1989, Moler & Van Loan 1983). Regardless of the merits of this approach, it is quite clear that the nearness of e^A to a banded matrix is of no use whatsoever in this framework.

To recap, our observation does not help in making classical methods more effective. Intriguingly, this is not the case for a new generation of methods for the approximation of exponentials, that have been recently introduced within the context of *Lie-group methods.* Let $G \subseteq \mathrm{M}_n(\mathbb{R})$ be a matrix Lie group and $\mathfrak{g}$ be the corresponding Lie algebra. (We refer the reader to (Carter, Segal & Macdonald 1995) and (Varadarajan 1984) for a good exposition of Lie groups and Lie algebras.) Given $A \in \mathfrak{g}$, it is true that $\mathrm{e}^A \in G$ and this forms vital part of many recent numerical methods for differential equations evolving on Lie groups: Runge–Kutta/Munthe-Kaas schemes (Munthe-Kaas 1998), Magnus expansions (Iserles & Nørsett 1999) and Fer expansions (Zanna 1997). Not every exponential approximant takes a matrix from a Lie algebra to a Lie group (Celledoni & Iserles 1998). In the case of *quadratic* Lie groups, e.g. $\mathrm{O}_n(\mathbb{R})$, $\mathrm{U}_n(\mathbb{R})$ and $\mathrm{Sp}_n(\mathbb{R})$, it suffices to use diagonal Padé approximants, but it can be proved that the only analytic function f such that $f(z) = 1 + z + \mathcal{O}(z^2)$ and $f : \mathfrak{sl}_n(\mathbb{R}) \to \mathrm{SL}_n(\mathbb{R})$ for all $n \geq 2$ is the exponential function itself!

The goal of designing exponential approximants that take $A \in \mathfrak{g}$ into its Lie group has motivated recent interest in a new generation of algorithms (Celledoni & Iserles 1998, Celledoni, Iserles & Nørsett 1999). The common denominator is that A is split in the form

$$A = \sum_{k=1}^{s} A_k,$$

where $A_k \in \mathfrak{g}$, $k = 1, 2, \ldots, s$, has an exponential that can be easily evaluated exactly and so that

$$\mathrm{e}^{tA_1}\mathrm{e}^{tA_2}\cdots\mathrm{e}^{tA_s} = \mathrm{e}^{tA} + \mathcal{O}\big(t^{p+1}\big)$$

for sufficiently large value of $p \geq 1$. Recently, it has been demonstrated that the above approach can be improved by letting

$$A = \sum_{k=1}^{r} a_k Q_k,$$

where $r = \dim \mathfrak{g}$ and $\{Q_1, Q_2, \ldots, Q_r\}$ is a basis of the algebra, and approximating

$$\mathrm{e}^{tA} \approx \mathrm{e}^{g_1(t)Q_1} \mathrm{e}^{g_2(t)Q_2} \cdots \mathrm{e}^{g_s(t)Q_s},$$

where $g_1, g_2, \ldots, g_r$ are polynomials (Celledoni & Iserles 1999). The main design features of this method are the right choice of the basis and an exploitation of certain features of the underlying Lie algebra. An important byproduct is that, whenever it is known that e^{tA} is near to a banded matrix, it is possible to amend the method to produce a banded approximant. Moreover, this procedure brings radically down the computational cost and allows to take full advantage of sparsity. We do not delve into the details of the method from (Celledoni & Iserles 1999), merely using it as an example of an application of the results in this paper.

In Section 2 we analyse in detail the case of a tridiagonal matrix, demonstrating that the decay of the entries of e^A away from the diagonal is bounded from above by a modified Bessel function. The theory for matrices of general bandwidth is introduced in Section 3, where we use Fourier analysis to obtain upper bounds on the decay away from the origin. Finally, in Section 4, we sketch briefly our conclusions and examine how the results can be generalized from the exponential to other well-behaved functions.

2 Tridiagonal matrices

Let $A \in \mathrm{M}_n(\mathbb{R})$ be a tridiagonal matrix and

$$\rho = \max_{k,l=1,2,\ldots,n} |A_{k,l}|.$$

We denote the entries of A^m by $(A^m_{k,l})_{k,l=1}^n$, whence $A^0_{k,l} = \delta_{k,l}$ and

$$A^{m+1}_{k,l} = \sum_{j=\max\{1,l-1\}}^{\min\{n,l+1\}} A^m_{k,j} A_{j,l}, \qquad k,l = 1,2,\ldots,n, \quad m \geq 0. \tag{2.1}$$

Proposition 1 *For every $m \geq 0$, $k = 1,2,\ldots,n$ and $|l| \leq m$, it is true that*

$$|A^m_{k,k+l}| \leq c_{m,l} \rho^m,$$

where $c_{0,l} = \delta_{0,l}$ and

$$c_{m+1,l} = c_{m,l-1} + c_{m,l} + c_{m,l+1}, \qquad |l| \leq m+1, \tag{2.2}$$

while $c_{m+1,l} = 0$, $|l| \geq m+2$.

Proof A trivial consequence of (2.1). Specifically, for $l = m+1$ we have

$$A^{m+1}_{k,k+m+1} = A^m_{k,k+m} A_{k+m,k+m+1},$$

hence we can let $c_{m+1,m+1} = c_{m,m}$. When $l = m$, we obtain

$$A^{m+1}_{k,k+m} = A^m_{k,k+m-1} A_{k+m-1,k+m} + A^m_{k,k+m} A_{k+m,k+m}$$

and we can take $c_{m+1,m} = c_{m,m-1} + c_{m,m}$. In the case $l = 0, 1, \ldots, m-1$ the recursion leads to

$$A^{m+1}_{k,k+l} = A^m_{k,k+l-1} A_{k+l-1,k+l} + A^m_{k,k+l} A_{k+l,k+l} + A^m_{k,k+l+1} A_{k+l+1,k+l}$$

and again we can choose $c_{m,l}$ consistently with (2.2). An identical argument extends to negative values of l. $\square$

Letting

$$C_m(z) := \sum_{l=-m}^{m} c_{m,l} z^l, \qquad m \in \mathbb{Z}^+,$$

we deduce at once that

$$C_m(z) = \left(z + 1 + \frac{1}{z}\right)^m, \qquad m \in \mathbb{Z}^+. \tag{2.3}$$

We are interested in bounding the size of the entries of $E := \mathrm{e}^A$. To this end we note that, for every $k = 1, 2, \ldots, n$ and $k + l = 1, 2, \ldots, n$ it is true that

$$E_{k,k+l} = \sum_{m=|l|}^{\infty} \frac{1}{m!} A^m_{k,k+l},$$

therefore

$$|E_{k,k+l}| \le \sum_{m=|l|}^{\infty} \frac{1}{m!} c_{m,l} \rho^m = F_{|l|}(\rho),$$

where

$$F_r(\rho) := \sum_{m=r}^{\infty} \frac{1}{m!} c_{m,r} \rho^m, \qquad r \in \mathbb{Z}^+.$$

To investigate the size of functions F_r, we expand (2.3) in Laurent series. It follows easily that

$$\begin{aligned}
&\left(z + 1 + \frac{1}{z}\right)^m \\
&= \sum_{k=0}^{m} \binom{m}{k} \left(z + \frac{1}{z}\right)^k = \sum_{k=0}^{m} \binom{m}{k} \sum_{l=0}^{k} \binom{k}{l} z^{2l-k} \\
&= \sum_{k=0}^{\lfloor m/2 \rfloor} \binom{m}{2k} \sum_{l=0}^{2k} \binom{2k}{l} z^{2(l-k)} + \sum_{k=0}^{\lfloor (m-1)/2 \rfloor} \binom{m}{2k+1} \sum_{l=0}^{2k+1} \binom{2k+1}{l} z^{2(l-k)-1}
\end{aligned}$$

$$= \sum_{l=-\lfloor m/2\rfloor}^{\lfloor m/2\rfloor} \left[\sum_{k=|l|}^{\lfloor m/2\rfloor} \frac{m!}{(m-2k)!(k+l)!(k-l)!} \right] z^{2l}$$
$$+ \sum_{l=-\lfloor (m-1)/2\rfloor}^{\lfloor (m+1)/2\rfloor} \left[\sum_{k=\max\{-l,l+1\}}^{\lfloor (m+1)/2\rfloor} \frac{m!}{(m-2k-1)!(k+l)!(k-l)!} \right] z^{2l-1}.$$

We concern ourselves with nonnegative indices l, since it is trivial to prove $c_{m,-l} = c_{m,l}$. Therefore

$$c_{m,2l} = \sum_{k=l}^{\lfloor m/2\rfloor} \frac{m!}{(m-2k)!(k+l)!(k-l)!}$$
$$c_{m,2l+1} = \sum_{k=l}^{\lfloor (m-1)/2\rfloor} \frac{m!}{(m-2k-1)!(k+l+1)!(k-l)!}$$

for all relevant values of l. We thus deduce that

$$F_{2l}(\rho) = \sum_{m=l}^{\infty} \sum_{k=l}^{\lfloor m/2\rfloor} \frac{1}{(m-2k)!(k+l)!(k-l)!}\rho^m = \sum_{k=l}^{\infty} \frac{1}{(k-l)!(k+l)!} \sum_{m=2k}^{\infty} \frac{\rho^m}{(m-2k)!}$$
$$= \mathrm{e}^\rho \sum_{k=l}^{\infty} \frac{\rho^{2k}}{(k-l)!(k+l)!} = \mathrm{e}^\rho \sum_{k=0}^{\infty} \frac{\rho^{2(k+l)}}{k!(k+2l)!} = \mathrm{e}^\rho I_{2l}(2\rho),$$

where $I_\nu(z)$ is the *modified Bessel function* (Abramowitz & Stegun 1965, Rainville 1960) Likewise,

$$F_{2l+1} = \sum_{m=l}^{\infty} \sum_{k=l}^{\lfloor (m-1)\rfloor} \frac{1}{(m-2k-1)!(k-l)!(k+l+1)!}\rho^m$$
$$= \sum_{k=l}^{\infty} \frac{1}{(k-l)!(k+l+1)!} \sum_{m=2k+1}^{\infty} \frac{\rho^m}{(m-2k-1)!}$$
$$= \mathrm{e}^\rho \sum_{k=0}^{\infty} \frac{1}{k!(k+2l+1)!}\rho^{2(k+l)!} = \mathrm{e}^\rho I_{2l+1}(2\rho).$$

We thus deduce that

$$F_r(\rho) = \mathrm{e}^\rho I_r(2\rho), \qquad r \in \mathbb{Z}^+. \tag{2.4}$$

The proof of the following theorem follows at once from (2.4).

Theorem 2 *Let A be tridiagonal and the magnitude of its nonzero entries bounded by $\rho \geq 0$. Then*

$$|(\mathrm{e}^A)_{k,l}| \leq \mathrm{e}^\rho I_{|k-l|}(2\rho), \qquad k, l = 1, 2, \ldots, n.$$

An alternative interpretation of the bound (2.4) is that the worse possible case is obtained for a bi-infinite *Toeplitz* matrix with the symbol $\rho(z + 1 + 1/z)$, whence all the upper bounds become equalities and $E_{k,k+s} = F_{|s|}(\rho)$, $k, s \in \mathbb{Z}$.[2]

[2] We refer the reader to (Grenander & Szegő 1958) for the theory of Toeplitz matrices.

Table 1: Entries of e^A larger in magnitude than a given threshold ε and the bandwidth within which they reside for tridiagonal and quindiagonal matrices. 'nz' and 'bd' stand for the number of nonzero entries and 'their' bandwidth respectively.

ε	Tridiagonal				Quindiagonal			
	worst case		average case		worst case		average case	
	nz	bd	nz	bd	nz	bd	nz	bd
10^{-2}	2170	5	968.9	1.94	5218	13	1888.9	4.28
10^{-4}	2944	7	1608.3	3.56	7058	18	3247.1	7.89
10^{-6}	3710	9	2175.4	5.01	8494	22	4416.0	10.86
10^{-8}	4468	11	2699.2	6.37	9898	26	5479.0	13.70
10^{-10}	5218	13	3194.4	7.65	10930	29	6469.6	16.39
10^{-12}	5960	15	3666.5	8.89	12278	33	7405.3	18.96

Since

$$I_\nu(z) = \sum_{m=0}^{\infty} \frac{1}{m!\Gamma(m+\nu+1)} \left(\frac{z}{2}\right)^{2m+\nu},$$

it follows that

$$I_r(2\rho) \approx \frac{\rho^r}{r!} \approx \frac{1}{\sqrt{2\pi r}} \left(\frac{\mathrm{e}\rho}{r}\right)^r, \qquad r \gg 1.$$

In other words, scaling by the magnitude of entries along the diagonal,

$$\log \frac{F_r(\rho)}{F_0(\rho)} \approx r(\log\rho - \log r + 1) - \tfrac{1}{2}\log(2\pi r) - \log I_0(2\rho), \qquad r \gg 1,$$

and the upper bound decays hyper-exponentially away from the diagonal. This illustrates the behaviour along the right column of Figure 1.

In Table 1 we have displayed the growth in the number of nonzero entries greater than a threshold $\varepsilon > 0$ for tridiagonal and (to illustrate the next section) quindiagonal 200×200 matrices for decreasing ε. We have considered both the *worst case* of a Toeplitz matrix and the *average case,* a mean of 1000 randomly selected matrices. For each case we have listed two numbers: the overall number of nonzero entries (out of 40000) and the minimal bandwidth within which they reside. It is evident that the rate of decay is very rapid: even in the case $\varepsilon = 10^{-12}$, not that far from IEEE machine epsilon, sparsity remains significant. It is important to bear in mind that, in the worse case, a Toeplitz matrix, although not identical to the bi-infinite Toeplitz matrix T_∞ with the same symbol, has spectrum exponentially near to $\sigma(T_\infty)$ (Grenander & Szegő 1958). This explains why our bounds for the decay in the size of the elements of e^{T_∞} remain remarkably sharp for finite-dimensional Toeplitz matrices.

In general, let A be a tridiagonal matrix whose entries are bounded by $\rho > 0$ and choose a threshold ε. Suppose that r is the least integer such that $\mathrm{e}^\rho I_r(2\rho) \le \varepsilon$. Then we might set all the elements of e^A with $|k-l| > r$ equal to zero whilst committing an entry-wise error less than ε. Note that this bound does not depend on the dimension n. The computation of such r is assisted by the following observation about modified Bessel functions.

Proposition 3 *Let ψ be the* Digamma function *(Abramowitz & Stegun 1965, p. 258). For every $0 \le \rho < 2\mathrm{e}^{\psi(\nu+1)}$ the sequence $\{I_\eta(\rho)\}_{\eta\ge\nu}$ is strictly monotonically decreasing to zero as $\eta \to \infty$.*

Proof The derivative of the modified Bessel function I_ν, where $\nu \in \mathbb{R}$, with respect to its parameter is

$$\frac{\partial}{\partial\nu} I_\nu(\rho) = I_\nu(\rho)\log\frac{\rho}{2} - \left(\frac{\rho}{2}\right)^\nu \sum_{k=0}^{\infty} \frac{\psi(k+\nu+1)}{k!\Gamma(k+\nu+1)}\left(\frac{\rho}{2}\right)^{2k},$$

where Γ is the Gamma function (Abramowitz & Stegun 1965, p. 377). Substituting the series expansion of I_ν, we have

$$\begin{aligned} I_\nu(\rho) &= \left(\frac{\rho}{2}\right)^\nu \sum_{k=0}^{\infty} \frac{1}{k!\Gamma(k+\nu+1)}\left(\frac{\rho}{2}\right)^{2k}\log\frac{\rho}{2} - \left(\frac{\rho}{2}\right)^\nu \sum_{k=0}^{\infty} \frac{\psi(k+\nu+1)}{k!\Gamma(k+\nu+1)}\left(\frac{\rho}{2}\right)^{2k} \\ &= -\left(\frac{\rho}{2}\right)^\nu \sum_{k=0}^{\infty}\left[\psi(k+\nu+1) - \log\frac{\rho}{2}\right]\frac{1}{k!\Gamma(k+\nu+1)}\left(\frac{\rho}{2}\right)^{2k}. \end{aligned}$$

As long as $0 \le \rho < 2\mathrm{e}^{\psi(\nu+1)}$, we have $\psi(k+\nu+1) - \log\frac{\rho}{2} > 0$, $k \ge 0$, and $\partial I_\nu(\rho)\partial\nu < 0$. This proves the proposition. $\square$

The condition of the proposition is always fulfilled for sufficiently large ν, since, according to the integral representation in (Abramowitz & Stegun 1965, p. 259),

$$\begin{aligned} \psi(t) = \frac{\Gamma'(t)}{\Gamma(t)} &= -\gamma + \int_0^\infty \frac{\mathrm{e}^{-\xi} - \mathrm{e}^{-t\xi}}{1-\mathrm{e}^{-\xi}}\mathrm{d}\xi \\ &= -\gamma + \int_0^\infty \frac{\mathrm{e}^{-\xi} - \mathrm{e}^{-2\xi} + \mathrm{e}^{-2\xi} - \mathrm{e}^{-t\xi}}{1-\mathrm{e}^{-\xi}}\mathrm{d}\xi \\ &= -\gamma + 1 + \int_0^\infty \frac{\mathrm{e}^{-2\xi} - \mathrm{e}^{-t\xi}}{1-\mathrm{e}^{-\xi}}\mathrm{d}\xi, \end{aligned}$$

where $\gamma \approx .57721$ is the Euler constant. Since the integrand is nonnegative for $t \ge 2$, we deduce that ψ is positive in that regime. Moreover, $\psi(t) \approx \log t$ for $t \gg 1$, hence it becomes unbounded for $t \to \infty$ (Abramowitz & Stegun 1965, p. 259).

3 General banded matrices

Let $A \in \mathrm{M}_n(\mathbb{R})$ be a banded matrix of *bandwidth* $s \ge 1$: thus, $A_{k,l} = 0$ for $|k-l| \ge s+1$. We let

$$\rho = \max_{k,l=1,2,\ldots,n} |A_{k,l}|.$$

It is important to investigate how much of the analysis of the tridiagonal case $s = 1$ survives in the present setting. Proposition 1 is a case in point and it requires a trivial amendment. As before, we denote by $A^m_{k,l}$ the entries of the mth power of A, $m \ge 0$.

Proposition 4 *It is true that*

$$|A^m_{k,k+l}| \le c_{m,l}\rho^m, \qquad m \ge 0, \quad k = 1,2,\ldots,n, \quad |l| \le m,$$

where $c_{0,l} = \delta_{0,l}$,

$$c_{m+1,l} = \sum_{j=-l}^{l} c_{m,s+j}, \qquad |l| \le m+1, \tag{3.1}$$

and $c_{m+1,l} = 0$ *for* $|l| \ge m+2$.

Proof Follows similarly to Proposition 1, except that, in place of (2.1), we use the recursion

$$A^{m+1}_{k,l} = \sum_{j=\max\{1,l-s\}}^{\min\{n,l+s\}} A^m_{k,j} A_{j,l}.$$

□

Retaining the definition of C_m, it follows at once from (3.1) that

$$C_m(z) = \sum_{l=-sm}^{sm} c_{m,l} z^l = \left(\sum_{l=-s}^{s} z^l\right)^m, \qquad m \in \mathbb{Z}^+.$$

Letting again $E = \mathrm{e}^A$, we deduce identically to Section 2 that

$$|E_{k,l}| \le F_{|k-l|}(\rho), \qquad k,l = 1,2,\ldots,n. \tag{3.2}$$

This is as far as the technique of Section 2 takes us. At least in the quindiagonal case $s = 2$ it is possible to express $F_r(\rho)$, with substantial effort, as a linear combination of modified Bessel functions,

$$\begin{aligned} F_r(\rho) &= \mathrm{e}^\rho \sum_{j=-\infty}^{\infty} I_{|j|}(2\rho) I_{r+2j}(2\rho) \\ &= \mathrm{e}^\rho \left\{ I_0(2\rho) I_r(2\rho) + \sum_{j=1}^{\infty} I_j(2\rho)[I_{r+2j}(2\rho) + I_{r-2j}(2\rho)] \right\}, \end{aligned}$$

except that this expression is of limited transparency and utility. We need a different approach, and it is provided by *Fourier analysis.*

We observe that

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} \mathrm{e}^{-\mathrm{i}r\theta} C_m(\mathrm{e}^{\mathrm{i}\theta}) \mathrm{d}\theta = \sum_{k=-sm}^{sm} c_{m,l} \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathrm{e}^{\mathrm{i}(r-l)\theta} \mathrm{d}\theta = c_{m,r}.$$

Thus, we deduce from (3.2) that for every $r \in \mathbb{Z}^+$

$$F_r(\rho) = \sum_{m=r}^{\infty} \frac{\rho^m}{m!} \frac{1}{2\pi} \int_{-\pi}^{\pi} \mathrm{e}^{-\mathrm{i}r\theta} \left(\sum_{k=-s}^{s} \mathrm{e}^{\mathrm{i}k\theta}\right)^m \mathrm{d}\theta$$

$$= \frac{1}{2\pi} \sum_{m=r}^{\infty} \mathrm{e}^{-\mathrm{i}r\theta} \sum_{m=r}^{\infty} \frac{1}{m!} \left[\rho \left(\sum_{k=-s}^{s} \mathrm{e}^{\mathrm{i}k\theta} \right) \right]^m \mathrm{d}\theta$$

$$= \frac{1}{2\pi} \int_{-\pi}^{\pi} \mathrm{e}^{-\mathrm{i}r\theta} \left\{ \exp\left(\rho \sum_{k=-s}^{s} \mathrm{e}^{\mathrm{i}k\theta} \right) - \sum_{m=0}^{r-1} \frac{1}{m!} \left[\rho \left(\sum_{k=-s}^{s} \mathrm{e}^{\mathrm{i}k\theta} \right) \right]^m \right\} \mathrm{d}\theta.$$

In other words, for each $r \in \mathbb{Z}^+$ we are interested in the rate of decay of the rth Fourier component of the function

$$g_r(z) := \exp\left(\rho \sum_{k=-s}^{s} z^k \right) - \sum_{m=0}^{r-1} \frac{1}{m!} \left[\rho \left(\sum_{k=-s}^{s} z^k \right) \right]^m.$$

Note that the integral representation of F_r can be somewhat simplified,

$$F_r(\rho)$$
$$= \frac{1}{\pi} \int_0^{\pi} \cos r\theta \left(\exp\left\{ \rho \left[1 + 2 \sum_{k=1}^{s} \cos(k\theta) \right] \right\} - \sum_{m=0}^{r-1} \frac{1}{m!} \{ \rho [1 + 2 \sum_{k=1}^{s} \cos(k\theta)] \}^m \right) \mathrm{d}\theta.$$

This can be readily used to produce an alternative proof of (2.4) by exploiting the identity

$$I_r(z) = \frac{1}{\pi} \int_0^{\pi} \mathrm{e}^{z \cos\theta} \cos(r\theta) \mathrm{d}\theta$$

(Abramowitz & Stegun 1965, p. 376). For more general analysis we require a classical estimate on the rate of decay of Fourier components (or, equivalently, of the coefficients of Laurent series) of analytic functions which can be found, for example, in (Henrici 1974, Vol. I, p. 221).

Theorem 5 *Let f be analytic in the annulus $\{z \in \mathbb{C} : \alpha < |z - z_0| < \beta\}$ and for $\sigma \in (\alpha, \beta)$ let*

$$\mu(\sigma) := \max_{-\pi \le \theta \le \pi} |f(z_0 + \sigma \mathrm{e}^{\mathrm{i}\theta})|.$$

The coefficients $\{f_n\}_{n=-\infty}^{\infty}$ of the Laurent series of f satisfy the inequality

$$|f_n| \le \frac{\mu(\sigma)}{\sigma^n}, \qquad n \in \mathbb{Z}. \tag{3.3}$$

In our case, each g_r is analytic in the punctured complex plane $\mathbb{C} \setminus \{0\}$, therefore $\alpha = 0$, $\beta = \infty$ and $z_0 = 0$. Moreover, for each $r \in \mathbb{Z}^+$,

$$\mu(\sigma) = \mu_r(\sigma) = \max_{-\pi \le \theta \le \pi} |f_r(\sigma \mathrm{e}^{\mathrm{i}\theta})| = \max_{-\pi \le \theta \le \pi} \left| \sum_{m=r}^{\infty} \frac{1}{m!} \left(\rho \sum_{k=-s}^{s} \sigma^k \mathrm{e}^{\mathrm{i}k\theta} \right)^m \right|$$
$$\le \sum_{m=r}^{\infty} \frac{1}{m!} \left(\rho \sum_{k=-s}^{s} \sigma^k \right)^m = g_r(\sigma).$$

Therefore, for every $\sigma > 1$ we obtain the upper bound

$$|F_r(\rho)| \le \sigma^{-r} g_r(\sigma) := \varphi_r(\sigma), \qquad r \in \mathbb{Z}^+. \tag{3.4}$$

The problem with the bound (3.4), though, is which $\sigma > 1$ to choose. Clearly, for different values of r we can choose different values of $\sigma = \sigma_r$, to make the bound (3.4) as low as possible.

As before, the tridiagonal case provides us with a clue. Letting $s = 1$, we have

$$\varphi_r(\sigma) = \sigma^{-r}\left\{ e^{\rho(\sigma+1+\sigma^{-1})} - \sum_{m=0}^{r-1} \frac{1}{m!}[\rho(\sigma+1+\sigma^{-1})]^m \right\}.$$

Therefore,

$$\varphi_r'(\sigma) - \frac{r}{\sigma}\varphi_r(\sigma) + \frac{\rho}{\sigma}(1-\sigma^{-2})\varphi_{r-1}(\sigma).$$

Setting the derivative to zero we have

$$\frac{r}{\rho} = (1-\sigma^{-1})\frac{\varphi_{r-1}(\sigma)}{\varphi_r(\sigma)}.$$

Since

$$\sigma\varphi_r(\sigma) - \varphi_{r-1}(\sigma) = -\frac{[\rho(1+\sigma^{-1}+\sigma^{-2})]^{r-1}}{(r-1)!} \approx 0, \qquad r \gg 1,$$

we have $\varphi_{r-1}(\sigma)/\varphi_r(\sigma) \approx \sigma$, therefore

$$\frac{r}{\rho} \approx \sigma - \sigma^{-1}$$

with the positive root

$$\sigma_r = \frac{r + \sqrt{r^2+4\rho^2}}{2\rho} \approx \frac{r}{\rho}.$$

Therefore, for sufficiently large r,

$$\begin{aligned} |F_r(\rho)| &\le \left(\frac{2\rho}{r+\sqrt{r^2+4\rho^2}}\right)^r \left[e^{\rho+\sqrt{r^2+4\rho}} - \sum_{m=0}^{r-1} \frac{(\rho+\sqrt{r^2+4\rho^2})^m}{m!} \right] \\ &\approx \left(\frac{\rho}{r}\right)^r \left(e^r - \sum_{m=0}^{r-1}\frac{r^m}{m!}\right) \approx \frac{\rho^r}{r!}. \end{aligned}$$

The above analysis remains virtually intact for general $s \ge 1$ and the equation for optimal σ is approximately

$$\rho \sum_{k=-s}^{s} k\sigma^k = r.$$

In general, this equation cannot be solved in a closed form. Suppose, however, that $r \gg 1$ and let $\varepsilon := \rho/r$, whence $0 < \varepsilon \ll 1$. We seek a solution of the form

$$\sigma_r = \left(\frac{c}{\varepsilon}\right)^\beta [1 + o(1)], \qquad \varepsilon \downarrow 0,$$

where $c>0$ and β are unknown constants. Substituting into the equation, it is trivial to verify that $c=\beta=1/s$, whence $[r/(\rho s)]^{1/s}$ is a very good approximation to the optimal value. After some easy algebra this leads to an upper bound

$$|F_r(\rho)| \le \left(\frac{\rho s}{r}\right)^{r/s}\left[\mathrm{e}^{r/s}-\sum_{m=0}^{r-1}\frac{(r/s)^m}{m!}\right], \qquad r\gg 1. \tag{3.5}$$

Theorem 6 *Let $E=\mathrm{e}^A$, where A is a banded matrix of bandwidth $s\ge 1$ and set $\rho=\max_{k,l=1,2,\ldots,n}|A_{k,l}|$. Then*

$$|E_{k,l}| \le \left(\frac{\rho s}{|k-l|}\right)^{|k-l|/s}\left[\mathrm{e}^{|k-l|/s}-\sum_{m=0}^{|k-l|-1}\frac{(|k-l|/s)^m}{m!}\right], \qquad |k-l|\gg 1. \tag{3.6}$$

It is easy to use the theorem to bound the bandwidth r within which one can confine all the entries of E which exceed $\varepsilon>0$ in magnitude. However, in that case it is more reasonable (and computationally straightforward) to compute numerically σ_r that minimises φ_r and choose an integer r such that $\varphi_r(\sigma_r)<\varepsilon$.

Figure 3 displays the functions $\varphi_r(\sigma)$ in a base-10 logarithmic scale. The minimum of the curve corresponds to the best value σ_r. It is evident from the figure that good estimates of the minimum are crucial to avoid an overly pessimistic estimate of the rate of decay. This, indeed, is a justification for numerical computation of the minimum, in preference to the estimate of Theorem 6.

4 Generalizations and conclusions

Let

$$h(z)=\sum_{m=0}^{\infty}h_m z^m$$

be an arbitrary function, analytic in the disc $|z|<\gamma$. How fast – if at all – decay the elements of $h(A)$ away from the diagonal if A is a banded matrix?

In Figure 4 we have plotted $h(\rho T)$, where $T\in \mathrm{M}_{100}(\mathbb{R})$ is Toeplitz, tridiagonal, with the symbol $z+1+z^{-1}$, while $h(z)=(1-\frac{1}{2}z)^{-1}(1+\frac{1}{2})$, for two different values of ρ. It is evident that, unlike in the exponential case, a very minor amendment in the size of the coefficients leads to radically different behaviour.

Retaining notation from Section 4, we briefly comment on the validity of our results in the more general setting. Again, A is of bandwidth $s\ge 1$ and its nonzero elements are bounded in magnitude by $\rho>0$. We have

$$F_r(\rho)=\frac{1}{2\pi}\int_{-\pi}^{\pi}\mathrm{e}^{-\mathrm{i}r\theta}\sum_{m=r}^{\infty}h_m\left(\rho\sum_{k=-s}^{s}\mathrm{e}^{\mathrm{i}k\theta}\right)^m \mathrm{d}\theta,$$

where $|F_{|k-l|}(\rho)|$ is a bound on the magnitude of the (k,l) entry of $h(A)$. Using Theorem 5 we deduce that, for every $\sigma>1$

$$|F_r(\rho)|\le \sigma^{-r}\max_{-\pi\le\theta\le\pi}\left|\sum_{m=r}^{\infty}h_m\left(\rho\sum_{k=-s}^{s}\sigma^k\mathrm{e}^{\mathrm{i}k\theta}\right)^m\right|$$

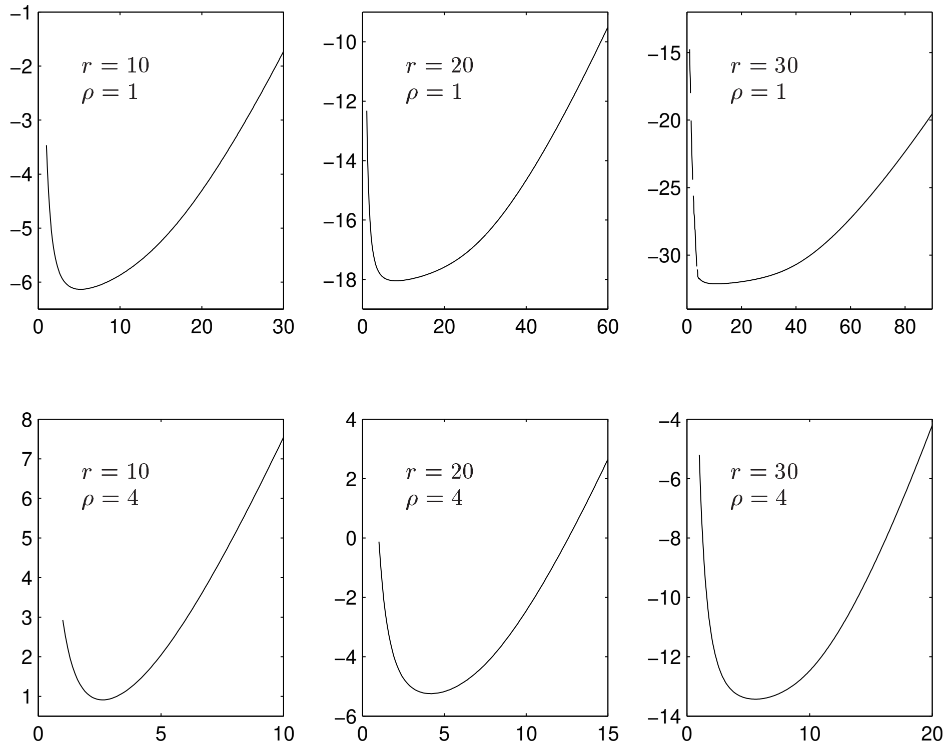


Figure 3: The function $\log_{10} \varphi_r(\sigma)$ for a range of values of r and different numbers ρ.

$$\le \sigma^{-r} \sum_{m=r}^{\infty} |h_m| \left(\rho \sum_{k=-s}^{s} \sigma^k \right)^m := \sigma^{-r} \tilde{h}_r \left(\rho \sum_{k=-s}^{s} \sigma^k \right). \tag{4.1}$$

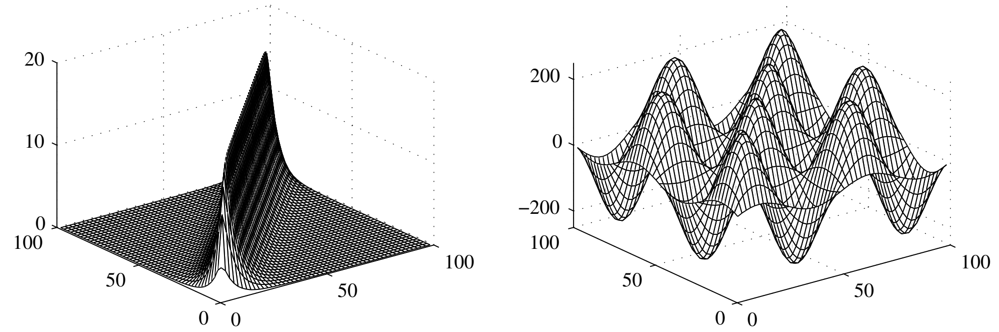
Of course, (4.1) makes sense only as long as the argument of $\tilde{h}_r$ is within the disc of convergence. Since, e.g. by the Cauchy criterion, $\tilde{h}_r$ and h share the same radius of convergence, we require that $\sigma > 1$ obeys

$$\rho \sum_{k=-s}^{s} \sigma^k < \gamma. \tag{4.2}$$

This means that we must restrict the range of ρ to

$$\rho < \frac{\gamma}{2s+1},$$

otherwise (4.2) fails for every $\sigma > 1$. This explains the difference between the two matrices in Figure 4: while $\rho = \frac{66}{100} < \frac{1}{3}$ lies on the 'safe' side, $\rho = \frac{67}{100}$ is marginally too large.



$\rho = \frac{66}{100}$ $\rho = \frac{67}{10}$

Figure 4: The matrix $h(A)$ for the [1/1] Padé approximant to the exponential, $h(z) = (1 - \frac{1}{2}z)^{-1}(1 + \frac{1}{2})$, and a tridiagonal Toeplitz matrix $A = \rho T \in \mathrm{M}_{100}(\mathbb{R})$.

To obtain the least upper bound in (4.1), we need to minimise $\sigma^{-r}\tilde{h}_r(\rho\sum_{k=-s}^{s}\sigma^k)$ for all $\sigma > 1$ that satisfy (4.2). Returning to the example from Figure 4, namely the [1/1] diagonal Padé approximant to the exponential, $h(z) = (1 - \frac{1}{2}z)^{-1}(1 + \frac{1}{2}z)$, we have $h_0 = 1$, $h_m = 2^{-m+1}$ for $m \in \mathbb{N}$ and

$$\tilde{h}_r(y) = \frac{2}{1 - \frac{1}{2}y}(\tfrac{1}{2}y)^r, \qquad 0 < y < 2,$$

in the tridiagonal case $s = 1$. Letting $\xi = \frac{1}{2}\rho(\sigma + 1 + \sigma^{-1}) < 1$, we have $\sigma = \frac{1}{2}(-1 + 2\xi/\rho + \sqrt{4\xi^2/\rho^2 - 4\xi/\rho - 3})$, therefore need to minimise

$$\sigma^{-r}\tilde{h}_r(\rho(\sigma + 1 + \sigma^{-1})) = \frac{2}{1 - \xi}\left[\frac{\xi}{2}\left(\frac{2\xi}{\rho} - 1 - \sqrt{4\frac{\xi^2}{\rho^2} - 4\frac{\xi}{\rho} - 3}\right)\right]^r$$

with respect to ξ. This can be done numerically.

To recap, we have demonstrated in this paper that the entries in the exponential of a banded matrix rapidly decay away from the diagonal. Moreover, by providing estimates on the rate of decay, we have shown that is it possible to predict, given any threshold $\varepsilon > 0$, the bandwidth outside of which the elements of e^A are smaller than ε in modulus. Similar analysis can be extended to analytic functions. In the case of analytic functions with finite radius of convergence, though, we need to restrict the size of the entries of A, otherwise decay away from the diagonal is not assured.

Computer experiments demonstrate that similar phenomenon takes place for matrices with more elaborate sparsity patterns. Upon exponentiation, having thrown away small entries, the surviving sparsity pattern, although degraded in comparison with the original matrix, is often substantial enough to be of interest in practical computation.

Acknowledgements

The author is grateful to Per Christian Moan for his helpful comments on an earlier version of the manuscript.

References

Abramowitz, M. & Stegun, I. (1965), *Handbook of Mathematical Functions*, Dover, New York.

Baker, G. A. (1975), *Essentials of Padé Approximants*, Academic Press, New York.

Carter, R., Segal, G. & Macdonald, I. (1995), *Lectures on Lie Groups and Lie Algebras*, LMS Student Texts, Cambridge University Press, Cambridge.

Celledoni, E. & Iserles, A. (1998), Approximating the exponential from a Lie algebra to a Lie group, Technical Report 1998/NA03, University of Cambridge.

Celledoni, E. & Iserles, A. (1999), Numerical calculation of the matrix exponential based on Wei–Norman equations, Technical Report (in preparation), University of Cambridge.

Celledoni, E., Iserles, A. & Nørsett, S. P. (1999), Complexity of Lie-algebraic discretization methods, Technical Report (in preparation), University of Cambridge.

Golub, G. H. & Van Loan, C. F. (1989), *Matrix Computations*, 2nd edn, The Johns Hopkins Press, Baltimore.

Grenander, U. & Szegő, G. (1958), *Toeplitz Forms and Their Applications*, Chelsea, New York.

Henrici, P. (1974), *Applied and Computational Complex Analysis*, Vol. I, John Wiley & Sons, New York.

Hochbruck, M. & Lubich, C. (1997), 'On Krylov subspace approximations to the matrix exponential operator', *SIAM J. Num. Anal.* **34**, 1911–1925.

Horn, R. A. & Johnson, C. R. (1985), *Matrix Analysis*, Cambridge University Press, Cambridge.

Iserles, A. & Nørsett, S. P. (1991), *Order Stars*, Chapman and Hall, London.

Iserles, A. & Nørsett, S. P. (1999), 'On the solution of linear differential equations in Lie groups', *Philosophical Trans. Royal Soc. A.* To appear.

Moler, C. B. & Van Loan, C. F. (1983), 'Nineteen dubious ways to compute the exponential of a matrix', *SIAM Review* **20**, 801–836.

Munthe-Kaas, H. (1998), 'Runge–Kutta methods on Lie groups', *BIT* **38**, 92–111.

Rainville, E. D. (1960), *Special Functions*, Macmillan, New York.

Tal Ezer, H. & Kosloff, R. (1984), 'An accurate and efficient scheme for propagating the time dependent Schrödinger equation', *J. Chem. Phys.* **81**, 3967–3970.

Varadarajan, V. S. (1984), *Lie Groups, Lie Algebras, and Their Representations*, number 102 *in* 'Graduate Texts in Mathematics', Springer-Verlag.

Zanna, A. (1997), Collocation and relaxed collocation for the Fer and the Magnus expansions, Technical Report 1997/NA17, Department of Applied Mathematics and Theoretical Physics, University of Cambridge, England.