**Numerische Mathematik manuscript No.**
(will be inserted by the editor)

# A Gautschi-type method for oscillatory second-order differential equations

**Marlis Hochbruck, Christian Lubich**

Mathematisches Institut, Universität Tübingen, D-72076 Tübingen, Germany,
e-mail: marlis@na.uni-tuebingen.de, lubich@na.uni-tuebingen.de

January 1998

**Summary** We study a numerical method for second-order differential equations in which high-frequency oscillations are generated by a linear part. For example, semilinear wave equations are of this type. The numerical scheme is based on the requirement that it solves linear problems with constant inhomogeneity exactly. We prove that the method admits second-order error bounds which are independent of the product of the step size with the frequencies. Methods with this property are called long-time-step methods in [3]. Our analysis also sheds new light on the mollified impulse method of [3]. We include results of numerical experiments with the sine-Gordon equation.

**Key words** Numerical integrator, oscillatory solutions, error bounds, stability, nonlinear wave equations.

*Mathematics Subject Classification (1991):* 65L05, 65L70, 65M12, 65M20.

## 1 Introduction

In this paper we study a numerical method for the solution of systems of second-order differential equations

$$y'' = -Ay + g(y) \,, \qquad y(0) = y_0 \,, \; y'(0) = y_0' \,, \tag{1}$$

where $A$ is a symmetric and positive semi-definite real matrix of arbitrarily large norm. We are interested in using step sizes that are not restricted by the frequencies of $A$, neither for stability nor for accuracy.

García-Archilla, Sanz-Serna and Skeel [3] recently proposed and analyzed a method for oscillatory differential equations, which they called the *mollified impulse method*. They obtained error bounds for numerical solutions of (1) which do not deteriorate when the product of the step size with the frequencies becomes large or, what is potentially worse, is close to multiples of $\pi$. Their method is based on the splitting $u'' = -Au$, $v'' = g(v)$.

Here we study a method which is instead based on the requirement that it reduces to an *exact solver for linear equations* (1) *with constant inhomogeneity* $g$. Such a method, which is simple to construct, can be traced back to an old paper of Gautschi [5]. More recently, in [7] we found methods of this type numerically promising in combination with Krylov subspace techniques for approximating the product of the matrix exponential, or related matrix functions, with a vector. Our positive numerical experience called for a rigorous error analysis of such methods.

The error analysis developed here gives very detailed information about the structure of the error. The error is of second order uniformly in the frequencies. It turns out to be largely determined by a scalar function of two variables which accounts for the mixing of frequencies by the numerical method. As a practical consequence, this can be used for the construction of a suitable filter function which appears in the scheme. Our error and stability analysis provides also new insight for the mollified impulse method.

The methods considered in this paper require, in every time step, the computation of the product $\varphi(h^2A)v$ of analytic functions $\varphi$ of the matrix $A$ scaled by the square of the step size $h$, with a vector. This is easy if the eigendecomposition of $A$ is available, most notably in pseudospectral methods for nonlinear wave equations. Otherwise (or possibly in combination with a partial eigendecomposition), such matrix-function vector products can be computed with Krylov subspace methods [2, 6]. A further alternative, which appears however less favourable in the present context, is to solve in every time step a linear initial value problem, which is associated with the matrix function in question, by a standard numerical integrator with smaller step sizes.

The paper is organized as follows: In Section 2 we present the numerical method and some of its variants, and an extension to more general equations $y'' = f(y) + g(y)$. Section 3 develops the error analysis for Eq. (1), with the main result stated in Theorem 1. A major technical difficulty in this paper is to bound the Schur multiplier norm of matrices composed of values of the error function. Such bounds are derived in Section 4. They depart from optimality only by logarithmic terms. Section 5 deals with the fixed-stepsize stability of the method for linear problems (1) with $g(y) = -By$ for positive semi-definite $B$. Section 6 gives some suitable filter functions. In Section 7 we discuss relationships and differences to the mollified impulse

method. Section 8 concludes the paper with numerical experiments on the sine-Gordon equation.

For a recent survey article on existing numerical approaches to oscillatory differential equations we refer to [8].

## 2 The integration scheme

Our starting point is the variation-of-constants formula for the solution of (1),

$$y(t+\tau) = \cos\tau\Omega \cdot y(t) + \Omega^{-1}\sin\tau\Omega \cdot y'(t) + \int_0^\tau \Omega^{-1}\sin(\tau-s)\Omega \cdot g(y(t+s))\,ds\ . \qquad (2)$$

Here and in the following we write

$$\Omega = A^{1/2}\ .$$

For an equation (1) with *constant* inhomogeneity $g$, (2) shows that

$$y(t+h) - 2y(t) + y(t-h) = h^2\sigma(h^2A)(-Ay(t)+g)\ , \qquad (3)$$

where the function $\sigma$ is given by

$$\sigma(x^2) = \left(\frac{\sin\frac{1}{2}x}{\frac{1}{2}x}\right)^2 = 2\,\frac{1-\cos x}{x^2} = 2\int_0^1 x^{-1}\sin(1-\theta)x\,d\theta\ . \qquad (4)$$

In the general case of (1), formula (3) suggests to replace $g(y(t))$ by a suitable constant vector $g_n$ over a time step, and to consider the numerical integration scheme with step size $h$,

$$y_{n+1} - 2y_n + y_{n-1} = h^2\sigma(h^2A)(-Ay_n + g_n)\ , \qquad (5)$$

where $y_n$ is an approximation to $y(t_n)$ at time $t_n = nh$. The obvious choice would be to set $g_n = g(y_n)$, in which case (5) can be considered as belonging to a class of methods introduced by Gautschi [5, p. 392f.]. However, like in [3], it turns out to be favourable to take instead a modified argument in $g$:

$$g_n = g(\phi(h^2A)y_n)\ , \qquad (6)$$

where the *filter function* $\phi$ is a suitably chosen real function whose purpose is to filter out resonant frequencies. We assume

$$\phi(0) = 1\ , \qquad \phi(k^2\pi^2) = 0\ , \quad k = 1,2,3,\ldots \qquad (7)$$

We assume throughout, without further mention, that $\phi$ and its first two derivatives are bounded on the positive half-line. It is reasonable to assume also

$$|\phi(x)| \le 1 , \qquad x \ge 0 . \tag{8}$$

Examples for possible choices of $\phi$ will be given in Section 6.

To obtain a second starting value for the recursion (5), we set

$$y_1 = \cos h\Omega \cdot y_0 + \Omega^{-1} \sin h\Omega \cdot y_0' + \tfrac{1}{2} h^2 \sigma(h^2 A)\, g_0 . \tag{9}$$

Like for the Störmer/Verlet/leapfrog method, there is a one-step version of the scheme (5):

$$\begin{aligned} v_{n+1/2} &= v_n + \tfrac{1}{2} h \sigma(h^2 A)\, (-A y_n + g_n) \\ y_{n+1} &= y_n + h\, v_{n+1/2} \\ v_{n+1} &= v_{n+1/2} + \tfrac{1}{2} h \sigma(h^2 A)\, (-A y_{n+1} + g_{n+1}) . \end{aligned} \tag{10}$$

This scheme yields $v_n = (y_{n+1} - y_{n-1})/(2h)$, which can be interpreted as an approximation to an *averaged* velocity

$$\overline{v}(t) = \frac{1}{2h} \int_{-h}^{h} y'(t+\tau)\, d\tau .$$

The method (10) is mathematically equivalent to (5) with (9) if $v_0$ is taken as

$$v_0 = \psi(h^2 A)\, y_0' , \tag{11}$$

where $\psi(x^2) = \sin x / x$. The interpretation of this expression as an approximated time average comes once more from (2). In case that approximations to the velocities themselves are of interest, they can be obtained by post-processing via

$$y_{n+1}' = y_{n-1}' + 2h\psi(h^2 A)(-A y_n + g_n) . \tag{12}$$

These values would again be exact when $g$ is constant. This can be seen by differentiating (2) with respect to $\tau$.

The above method can be viewed as a special case, for $f(y) = -Ay$, of a method for more general differential equations

$$y'' = f(y) + g(y) .$$

Given $y_n$ and $y_n'$, one computes a suitable averaged value $\overline{y}_n$ and the solution of

$$u'' = f(u) + g(\overline{y}_n) , \qquad u(0) = y_n,\ u'(0) = y_n' . \tag{13}$$

Then, $y_{n+1}$ and $y_{n+1}'$ are computed from

$$\begin{aligned} y_{n+1} - 2y_n + y_{n-1} &= u(h) - 2u(0) + u(-h) , \\ y_{n+1}' - y_{n-1}' &= u'(h) - u'(-h) , \end{aligned} \tag{14}$$

or from the averaged-velocity version that corresponds to (10). When (13) is solved approximately by a numerical method with smaller time steps, then this becomes a symmetric multiple-time-stepping scheme.

## 3 Finite-time error analysis

We make no smoothness assumption about the (highly oscillatory) solution and impose instead, as in [3], a finite-energy condition:

$$\tfrac{1}{2}y'(t)^T y'(t) + \tfrac{1}{2}y(t)^T A y(t) \le \tfrac{1}{2}K^2 \ . \tag{15}$$

The following result shows second-order convergence of $y_n$ in the Euclidean norm and first-order convergence in the energy norm. The Euclidean norm and its induced matrix norm are both denoted by $\|\cdot\|$ throughout the paper.

**Theorem 1** *In Eq.* (1)*, let $A$ be a symmetric and positive semi-definite $N \times N$ matrix, and assume that $g, g_y, g_{yy}$ are bounded in the Euclidean norm or its induced norms by $M_0, M_1, M_2$, respectively. Let the solution satisfy the finite-energy condition* (15) *for $0 \le t \le T$. Then, the error of the numerical method of Section 2 is bounded for $0 \le nh \le T$ by*

$$\|y_n - y(t_n)\| \le h^2 \cdot Ce^{Lt_n}(M_1 K t_n + M_2 K^2 t_n^2 + M_1 M_0 t_n^2)\, \ell(n,N) \ ,$$

*where $C$ is a constant which depends only on the filter function $\phi$, $L = \sqrt{M_1}$, and $\ell(n,N) \le \log(n+1)\ \log(N+1)$ and also $\ell(n,N) \le \sqrt{N}$. A bound of the same type holds for $h\|\Omega(y_n - y(t_n))\| + h\|v_n - \overline{v}(t_n)\| + h\|y_n' - y'(t_n)\|$.*

The proof provides much more detailed information about the structure of the error. This will be made explicit at the end of this section. The logarithmic term $\ell(n,N)$ comes from our technique of estimating the entrywise product of the Jacobian $g_y$ with certain matrices depending on the numerical scheme and the frequencies of $A$. We conjecture that this logarithmic term can be omitted in the estimate.

We note that condition (15) implies

$$\|\Omega y(t)\| \le K \ , \quad \|y'(t)\| \le K \ ,$$

which are the conditions we will actually work with. In the case of higher regularity $\|\Omega^2 y(t)\| \le K$, $\|\Omega y'(t)\| \le K$, our analysis would yield second-order bounds also for $\|y_n' - y'(t_n)\|$.

The proof of Theorem 1 proceeds via a series of lemmas. In the following, $C$ always denotes a constant which depends only on the choice of the filter function $\phi$, and which takes on different values on different occurrences.

**Lemma 1** *The truncation error*

$$d_n = y(t_{n+1}) - 2y(t_n) + y(t_{n-1}) - h^2\sigma(h^2A)(-Ay(t_n) + g(\phi(h^2A)y(t_n)))$$

*is of the form*

$$d_n = h^3 L_n\, \Omega y(t_n) + h^4 z_n \ ,$$

*where the matrix $L_n$, given by* (16) *below, is bounded by $\|L_n\| \le CM_1$, and $\|z_n\| \le CM_2K^2$.*

*Proof* By the variation-of-constants formula (2) for $y(t_n \pm h)$, we obtain

$$d_n = \int_0^h \Omega^{-1} \sin(h-\tau)\Omega \cdot \Big( g(y(t_n+\tau)) - 2g(\phi(h^2A)y(t_n)) + g(y(t_n-\tau)) \Big)\, d\tau \ .$$

By assumption (15), we have

$$\|y(t_n \pm \tau) - y(t_n)\| \le \int_0^\tau \|y'(t_n \pm s)\|\, ds \le K\tau \ .$$

This gives us, with $G_n = g_y(y(t_n))$,

$$g(y(t_n \pm \tau)) - g(y(t_n)) = G_n(y(t_n \pm \tau) - y(t_n)) + r_n^{\pm} \ , \quad \|r_n^{\pm}\| \le M_2K^2\tau^2 .$$

Since $(1-\phi(x^2))/x$ is bounded for $x > 0$, we have

$$\|(I - \phi(h^2A))y(t_n)\| \le h\, \|(I - \phi(h^2\Omega^2))\,(h\Omega)^{-1}\| \cdot \|\Omega y(t_n)\| \le h\, CK \ ,$$

using again (15) in the last inequality. This yields

$$g(y(t_n)) - g(\phi(h^2A)y(t_n)) = G_n\,(I - \phi(h^2A))y(t_n) + s_n \ , \qquad \|s_n\| \le M_2C^2K^2h^2 \ .$$

Using the variation-of-constants formula (2) for $y(t_n \pm \tau)$ and defining

$$L_n = 2\int_0^1 (h\Omega)^{-1} \sin(1-\theta)h\Omega \cdot G_n \cdot (\cos\theta h\Omega - \phi(h^2\Omega^2))(h\Omega)^{-1}\, d\theta \ , \tag{16}$$

we thus obtain the desired result. □

**Lemma 2** *The errors $e_n = y_n - y(t_n)$ satisfy*

$$e_{n+1} = -W_{n-1}e_0 + W_n e_1 + \sum_{j=1}^{n} W_{n-j}(h^2F_j e_j - d_j)$$

*with $W_n = (\sin(n+1)h\Omega)\,(\sin h\Omega)^{-1}$, and with matrices $F_j$ bounded by $\|F_j\| \le M_1$.*

*Proof* By definition of the truncation error, we have

$$e_{n+1} - 2e_n + e_{n-1} = \\ h^2\sigma(h^2A)(-Ae_n + g(\phi(h^2A)y_n) - g(\phi(h^2A)y(t_n))) - d_n \ .$$

Since $\sigma(h^2A)(g(\phi(h^2A)y_n) - g(\phi(h^2A)y(t_n))) = F_n e_n$ with the matrix

$$F_n = \sigma(h^2A) \int_0^1 g_y(\phi(h^2A)(y(t_n) + \theta e_n))\, d\theta \cdot \phi(h^2A) \ ,$$

which is bounded by $M_1$, and since $2 - h^2\sigma(h^2A)A = 2\cos h\Omega$, the error equation becomes

$$e_{n+1} - 2\cos h\Omega\, e_n + e_{n-1} = h^2 F_n e_n - d_n \ ,$$

or in one-step form,

$$\begin{pmatrix} e_{n+1} \\ e_n \end{pmatrix} = R \begin{pmatrix} e_n \\ e_{n-1} \end{pmatrix} + \begin{pmatrix} h^2 F_n e_n - d_n \\ 0 \end{pmatrix} \ ,$$

with

$$R = \begin{pmatrix} 2\cos h\Omega & -I \\ I & 0 \end{pmatrix} \ .$$

Clearly then,

$$\begin{pmatrix} e_{n+1} \\ e_n \end{pmatrix} = R^n \begin{pmatrix} e_1 \\ e_0 \end{pmatrix} + \sum_{j=1}^{n} R^{n-j} \begin{pmatrix} h^2 F_j e_j - d_j \\ 0 \end{pmatrix} \ .$$

The result now follows from verifying that $(R^n)_{11} = W_n$ and $(R^n)_{12} = -W_{n-1}$. For example, this can be done using the block Schur decomposition

$$R = U \begin{pmatrix} e^{ih\Omega} & X \\ 0 & e^{-ih\Omega} \end{pmatrix} U^* \ , \qquad U = \tfrac{1}{\sqrt{2}} \begin{pmatrix} e^{ih\Omega} & -I \\ I & e^{-ih\Omega} \end{pmatrix}$$

with $X = -2e^{-ih\Omega}\cos h\Omega$, noting that

$$R^n = U \begin{pmatrix} e^{inh\Omega} & W_{n-1}X \\ 0 & e^{-inh\Omega} \end{pmatrix} U^* \ .$$

□

**Lemma 3** *We have*

$$\Big\| \sum_{j=1}^{n} W_{n-j}\, d_j \Big\| \le h^2 \cdot C(M_1 K t_n + M_2 K^2 t_n^2 + M_1 M_0 t_n^2)\, \ell(n, N)$$

*with* $\ell(n, N) \le \min(\log(n+1)\log(N+1), \sqrt{N})$.

*Proof* In view of Lemma 1 and the variation-of-constants formula (2) for $t = 0$ and $\tau = t_j$, we write

$$\sum_{j=1}^{n} W_{n-j}\, d_j = h^2 (a_n + b_n + c_n)$$

with

$$\begin{aligned}
a_n &= h \sum_{j=1}^{n} W_{n-j}\, L_j \left(\cos t_j \Omega \cdot \Omega y_0 + \sin t_j \Omega \cdot y_0'\right) \\
b_n &= h \sum_{j=1}^{n} W_{n-j}\, L_j \int_0^{t_j} \sin(t_j - s)\Omega \cdot g(y(s))\, ds \\
c_n &= h^2 \sum_{j=1}^{n} W_{n-j}\, z_j \ .
\end{aligned}$$

We study $a_n, b_n, c_n$ in parts (a),(b),(c) of the proof, respectively.

(a) Let $\omega_k$ be the $k$th eigenvalue of $\Omega$, and let $Q$ be the orthogonal matrix of eigenvectors, so that $Q^T \Omega Q = \operatorname{diag}(\omega_k)$. We write

$$a_n = t_n (U_n\, \Omega y_0 + V_n\, y_0')$$

and denote the matrix entries in the eigenbasis representation as

$$(\mu_n^{k\ell}) = Q^T U_n Q \ , \quad (\nu_n^{k\ell}) = Q^T V_n Q \ , \quad (\gamma_n^{k\ell}) = Q^T G_n Q \ .$$

For fixed $k, \ell$, we omit the superscripts in the matrix entries and write $\alpha = h\omega_k$, $\beta = h\omega_\ell$. We have

$$\begin{pmatrix} \mu_n \\ \nu_n \end{pmatrix} = \frac{1}{n} \sum_{j=0}^{n-1} \delta_j(\alpha,\beta)\, \gamma_{n-j} \begin{pmatrix} \cos(n-j)\beta \\ \sin(n-j)\beta \end{pmatrix} \ ,$$

where

$$\delta_j(\alpha,\beta) = 2\, \frac{\sin(j+1)\alpha}{\sin\alpha} \int_0^1 \frac{\sin(1-\theta)\alpha}{\alpha} (\cos\theta\beta - \phi(\beta^2))\, \frac{d\theta}{\beta} \, . \qquad (17)$$

To estimate the above sum, we use partial summation. Let

$$\varepsilon_n(\alpha,\beta) = \frac{1}{n} \sum_{j=0}^{n-1} \delta_j(\alpha,\beta)\, e^{-ij\beta} \, . \qquad (18)$$

We then have

$$\begin{aligned}
\mu_n + i\nu_n &= \sum_{j=0}^{n-1} \frac{1}{n}\, \delta_j(\alpha,\beta)\, e^{-ij\beta} \cdot \gamma_{n-j} \cdot e^{in\beta} \\
&= \Big( \varepsilon_n(\alpha,\beta)\, \gamma_0 + \sum_{j=0}^{n-1} \frac{j+1}{n}\, \varepsilon_{j+1}(\alpha,\beta)\, (\gamma_{n-j} - \gamma_{n-j-1}) \Big) e^{in\beta} \ .
\end{aligned}$$

Recall that $\gamma_j$ is the $(k,\ell)$ component of $\widehat{G}_j = Q^T G_j Q$, where $G_j = g_y(y(t_j))$. Letting $E_n = (\varepsilon_n(h\omega_j, h\omega_k))_{j,k=1}^N$ and $D_n = \mathrm{diag}\,(e^{inh\omega_k})$, we thus have

$$a_n = t_n \,\mathrm{Re}\, Q\Big(E_n \bullet \widehat{G}_0 + \sum_{j=0}^{n-1} \frac{j+1}{n}\, E_{j+1} \bullet (\widehat{G}_{n-j} - \widehat{G}_{n-j-1})\Big) \cdot D_n Q^T (\Omega y_0 - i y_0') \;,$$

where $\bullet$ denotes the entrywise product of matrices. Since

$$\|\widehat{G}_j\| \le M_1 \;, \quad \|\widehat{G}_j - \widehat{G}_{j-1}\| \le M_2 K h \;, \tag{19}$$

Lemma 5 below gives us that

$$\|a_n\| \le t_n \, C\ell(n,N)\,(M_1 + M_2 K t_n)\, 2K \;.$$

(b) We set

$$r_n = \int_0^{t_n} e^{-is\Omega} g(y(s))\, ds \;.$$

In terms of the eigencomponents $(b_n^k) = Q^T b_n$ and $(r_n^k) = Q^T r_n$ we then have

$$b_n^k = h \;\mathrm{Im} \sum_{j=0}^{n-1} \sum_{\ell} \delta_j(h\omega_k, h\omega_\ell)\, e^{-ijh\omega_\ell} \cdot \gamma_{n-j}^{k\ell} r_{n-j}^{\ell} \cdot e^{inh\omega_\ell} \;.$$

Partial summation gives us (note that $r_0^\ell = 0$)

$$b_n^k = t_n \;\mathrm{Im} \sum_{j=0}^{n-1} \sum_{\ell} \frac{j+1}{n} \varepsilon_{j+1}(h\omega_k, h\omega_\ell) \cdot \Big(\gamma_{n-j}^{k\ell} r_{n-j}^{\ell} - \gamma_{n-j-1}^{k\ell} r_{n-j-1}^{\ell}\Big) \cdot e^{inh\omega_\ell} \;,$$

and with Lemma 5 we conclude

$$\|b_n\| \le t_n C\ell(n,N) \sum_{j=0}^{n-1} \Big( \|\widehat{G}_{n-j} - \widehat{G}_{n-j-1}\| \cdot \|r_{n-j}\| + \|\widehat{G}_{n-j-1}\| \cdot \|r_{n-j} - r_{n-j-1}\|\Big) \;.$$

From the variation-of-constants formula (2) and its differentiated version we obtain with (15) $\|r_j\| = \|e^{it_j\Omega} r_j\| \le 4K$. Together with $\|r_j - r_{j-1}\| \le M_0 h$ and (19) we therefore obtain

$$\|b_n\| \le t_n^2 \, C\, \ell(n,N)\,(4M_2K^2 + M_1 M_0) \;.$$

(c) Finally, Lemma 1 and the bound $\|W_n\| \le n+1$ give us

$$\|c_n\| \le C M_2 K^2 t_n^2 \;. \qquad \square$$

*Proof of Theorem 1.* For the errors in the starting values we have $e_0 = 0$ and by (9)

$$\begin{aligned}\|e_1\| &= \Big\| \int_0^h \Omega^{-1} \sin(h-\tau)\Omega \cdot (g(y(\tau)) - g(\phi(h^2\Omega^2)y_0))\, d\tau \Big\| \\ &\le C M_1 K\, h^3 \ .\end{aligned}$$

Moreover, for the matrices in Lemma 2 we have $\|W_n\| \le n+1$. With the estimate of Lemma 3, the stated bound for $\|e_n\|$ now follows from a discrete Gronwall inequality [3, Lemma 2] applied to the recursion of Lemma 2.

The error bound for $h(v_n - \overline{v}(t_n)) = e_n - e_{n-1}$ is then immediate, and the bound for $h\Omega\, e_n$ follows with Lemma 2, since also $\|h\Omega F_j\| \le 2M_1$ and $\|h\Omega\, e_1\| \le C M_1 K h^3$, and because we get $\| \sum_{j=1}^n W_{n-j} h\Omega d_j \| = O(h^2)$ as in Lemma 3. Finally, to obtain the bound for $e_n' = y_n' - y'(t_n)$ we note that (12) implies

$$e_{n+1}' = e_{n-1}' - 2\Omega \sin h\Omega \cdot e_n + O(h^2) \ .$$

Since $\| \sin h\Omega \cdot W_n \| \le 1$ and $\|e_n\| = O(h^2)$, we see from Lemma 2 that, on a fixed time interval,

$$\sin h\Omega \cdot e_{n+1} = h^2 \sin h\Omega\, (a_n + b_n) + O(h^3) \ , \tag{20}$$

where $a_n$ and $b_n$ are those of the proof of Lemma 3, and the $O(h^3)$ remainder term, $s_n$ say, is such that $\Omega s_n = O(h^2)$. Inserting this formula in the recursion for $e_n'$, it can be shown as in the proof of Lemma 3 that this implies $\|e_n'\| = O(h)$, where the constant in the $O$-symbol is of the same type as before. We omit the details for this last estimate. □

Formula (20) makes explicit the dominant error term for the eigencomponents corresponding to those frequencies for which $h\omega_k$ is bounded away from an integer multiple of $\pi$. Recall that $a_n$ and $b_n$ are determined by the error function $\varepsilon_n(\alpha,\beta)$, which is studied in Section 4.

## 4 Properties of the error function

**Lemma 4** *The error functions $\varepsilon_n(\alpha,\beta)$ defined by* (17), (18) *are uniformly bounded for all $\alpha, \beta \ge 0$ and $n \ge 0$, and $\lim_{n\to\infty} \varepsilon_n(\alpha,\beta) = 0$ if $\alpha \pm \beta \ne 2k\pi$ and $\alpha \ne k\pi$ with integer $k$.*

*Proof* The tools of this proof are trigonometric identities and repeatedly the mean value theorem. It is in this proof that condition (7) comes into play. Let

$$e^{in\beta} \varepsilon_n(\alpha,\beta) = \frac{1}{n\beta} S_n(\alpha,\beta) I(\alpha,\beta) \ ,$$

where

$$S_n(\alpha,\beta) = 2\sum_{j=0}^{n-1} \frac{\sin(j+1)\alpha}{\sin\alpha}\, e^{i(n-j)\beta}$$

and

$$\begin{aligned} I(\alpha,\beta) &= \int_0^1 \frac{\sin(1-\theta)\alpha}{\alpha}(\cos\theta\beta - \phi(\beta^2))d\theta \\ &= -\left(\frac{\cos\beta-\cos\alpha}{\beta^2-\alpha^2} + \tfrac{1}{2}\sigma(\alpha^2)\,\phi(\beta^2)\right). \end{aligned}$$

With the bounds $\sin(1-\theta)\alpha/\alpha \le 1-\theta$ and (8) we have

$$|I(\alpha,\beta)| \le 1, \qquad \text{for all } \alpha,\beta \ge 0. \tag{21}$$

From $\cos\theta\beta - \phi(\beta^2) = O(\beta^2)$, when $\beta$ tends to zero, we conclude that there is a constant $C_1$ such that

$$\frac{1}{\beta}|I(\alpha,\beta)| \le C_1\beta, \qquad \text{for all } \beta \ge 0. \tag{22}$$

Next we consider the real part of $S_n(\alpha,\beta)$, which by trigonometric identities turns out to be

$$\begin{aligned} \operatorname{Re} S_n(\alpha,\beta) = \frac{1}{\cos\beta-\cos\alpha}\,\frac{1}{\sin\alpha}\,(&-\sin n\alpha\cos\alpha\,(\cos\beta-\cos\alpha) \\ &-\sin\alpha\,(\sin n\beta\sin\beta - \sin n\alpha\sin\alpha) \\ &+\sin\alpha\cos\beta\,(\cos n\beta - \cos n\alpha)). \end{aligned} \tag{23}$$

$\operatorname{Re} S_n(\alpha,\beta)$ is a continuous, $2\pi$-periodic function in $\alpha,\beta$, hence we set

$$\alpha = 2k\pi + a, \qquad \beta = 2m\pi + b, \qquad 0 \le |a|,|b| \le \pi.$$

By continuity, it is sufficient to consider $\alpha,\beta$ with $0 < |a|,|b| < \pi$ and $|a| \ne |b|$. Moreover, $\operatorname{Re} S_n(\alpha,\beta) = \operatorname{Re} S_n(a,b)$ is an even function in $a,b$. Hence we can restrict ourselves to the case $a,b > 0$.

We consider the three terms in (23) separately. The first term is bounded by $n$. For the second term, by the generalized mean-value theorem for a fraction of differentiable functions, there is a $\theta$ between $a$ and $b$ such that

$$\frac{\sin nb\sin b - \sin na\sin a}{\cos b - \cos a} = -\cos\theta\,\frac{\sin n\theta}{\sin\theta} - n\cos n\theta,$$

and hence this expression is bounded by $2n$ for all $a,b \ge 0$. From the above bounds we conclude that the product of $I(\alpha,\beta)/(n\beta)$ with the first two terms in (23) is uniformly bounded for all $\alpha,\beta \ge 0$ and $n \ge 0$.

For the last term in (23), things are more complicated, because this term grows like $O(n^2)$ for $\alpha \to \beta$ and $\beta \to k\pi$. However, we will show that the

product of the third term with the integral $I(\alpha,\beta)/(n\beta)$ is bounded. The mean-value theorem guarantees the existence of $\theta$ between $a$ and $b$ such that

$$\frac{\cos nb - \cos na}{\cos b - \cos a} = n\,\frac{\sin n\theta}{\sin\theta}.$$

Hence, there is a constanct $C_2$ such that for $0 < \delta < \frac{1}{2}\pi$

$$\left|\frac{I(\alpha,\beta)}{n\beta}\right| \left|\frac{\cos nb - \cos na}{\cos b - \cos a}\right| \le \frac{C_2}{\delta} \qquad \text{for } \delta \le a, b \le \pi - \delta.$$

We now consider the case $\beta = b \to 0$. From $|\cos nb - \cos na| \le n|b-a|$ and $\cos b - \cos a = \frac{1}{2}(a-b)(a+b)(1+O(a^2+b^2))$ we obtain

$$\left|\frac{\cos nb - \cos na}{\cos b - \cos a}\right| \le \frac{4n}{a+b} \qquad \text{for } 0 < b < \tfrac{1}{2}\pi\ ,\ \ 0 < a < \pi\ .$$

For $\beta \to 0$ we therefore conclude with (22)

$$\left|\frac{I(\alpha,\beta)}{n\beta}\right| \left|\frac{\cos n\beta - \cos n\alpha}{\cos\beta - \cos\alpha}\right| \le 4C_1 \qquad \text{for } 0 < \beta < \tfrac{1}{2}\pi\ ,\ \ \alpha > 0\ .$$

For $\beta > \frac{1}{2}\pi$, we have for the product with the first term of $I(\alpha,\beta)$

$$\frac{1}{n\beta}\left|\frac{\cos n\beta - \cos n\alpha}{\beta^2 - \alpha^2}\right| \le \frac{1}{\beta(\beta+\alpha)} \le \frac{4}{\pi^2} \qquad \text{for } \beta > \tfrac{1}{2}\pi\ ,\ \ \alpha > 0\ .$$

Next we consider the product with the second term in $I(\alpha,\beta)$ for $\beta$ near $\pi$. Here we have similarly to the above

$$\left|\frac{\cos nb - \cos na}{\cos b - \cos a}\right| \le \frac{4n}{|\pi - a| + |\pi - b|}$$

for $\frac{1}{2}\pi < b < \frac{3}{2}\pi$, $0 < a < 2\pi$. By condition (7), we have $|\phi(\beta^2)| \le C_3|\pi - \beta|$ for $\beta$ near $\pi$, and hence

$$\left|\frac{\sigma(\alpha^2)\,\phi(\beta^2)}{n\beta}\right| \cdot \left|\frac{\cos n\beta - \cos n\alpha}{\cos\beta - \cos\alpha}\right| \le \frac{4C_3}{\frac{1}{2}\pi}$$

for $\frac{1}{2}\pi < \beta < \frac{3}{2}\pi$, $a > 0$. The same argument applies for $\beta$ near arbitrary integer multiples of $\pi$.

Combining these estimates, we see that $\mathrm{Re}\, e^{in\beta}\varepsilon_n(\alpha,\beta)$ is bounded independently of $\alpha, \beta$ and $n$. Similarly, we can show that such a uniform bound exists for the imaginary part, and hence $\varepsilon_n(\alpha,\beta)$ is bounded uniformly.

In the nonresonance case, where $|\alpha \pm \beta - 2k\pi| \ge \delta > 0$ and $|\alpha - k\pi| \ge \delta$ for all integers $k$, we have $|S_n(\alpha,\beta)| \le C/\delta$, which by (21) and (22) implies

$$|\varepsilon_n(\alpha,\beta)| \le \frac{C}{n\delta}\ . \tag{24}$$

This proves the second assertion of the lemma. □

The logarithmic term in Theorem 1 results from the following bound.

**Lemma 5** *Let $E_n = (\varepsilon_n(\alpha_j, \alpha_k))_{j,k=1}^N$, where the $\alpha_j$ are arbitrary non-negative real numbers. In the matrix norm induced by the Euclidean norm, the entrywise product of $E_n$ with an arbitrary $N \times N$ matrix $G$ is then bounded by*

$$\|E_n \bullet G\| \le C \, \log(n+1) \, \log(N+1) \, \|G\| \, .$$

*The constant $C$ depends only on the choice of the filter function $\phi$.*

*Remark.* We have immediately

$$\|E_n \bullet G\| \le C_0 \, \| \, |G| \, \| \le C_0 \sqrt{N} \, \|G\|$$

with $C_0 = \sup_{j,\alpha,\beta} |\varepsilon_j(\alpha, \beta)|$, which is finite by Lemma 4.

*Proof* The proof proceeds by splitting the matrix $E_n$ into a sum of matrices and estimating them separately. We may assume $\alpha_1 \le \alpha_2 \le \ldots \le \alpha_N$.

(a) Consider first the triangle $\Delta : 0 \le \alpha \le \beta < \pi$ and let $E_n^\Delta$ be the submatrix of $E_n$ defined by

$$E_n^\Delta = (e_{jk}) \qquad \text{with} \quad e_{jk} = \begin{cases} \varepsilon_n(\alpha_j, \beta_k) & \text{if } (\alpha_j, \beta_k) \in \Delta \\ 0 & \text{else.} \end{cases} \tag{25}$$

Here we write $\beta_k = \alpha_k$ in the second argument for notational clarity. We split $E_n^\Delta$ further into a part $E_n^V$ whose entry arguments are near the vertical edge $\alpha = 0$ of $\Delta$, into a part $E_n^D$ near the diagonal edge $\alpha = \beta$, and a part $E_n^C$ close to the corner $(0,0)$. For this, let $\varphi$ be a smooth cutting function with $\varphi(x) = 1$ for $x \le \frac{1}{3}$, and $\varphi(x) = 0$ for $x \ge \frac{2}{3}$. Further, let $\chi_n$ be the characteristic function of the interval $[0, 1/n]$. We have

$$E_n^\Delta = E_n^V + E_n^D + E_n^C$$

with

$$\begin{aligned} E_n^V &= (e_{jk}^V) = (\varphi(\alpha_j/\beta_k)(1 - \chi_n(\beta_k)) \, e_{jk}) \\ E_n^D &= (e_{jk}^D) = ((1 - \varphi(\alpha_j/\beta_k))(1 - \chi_n(\beta_k)) \, e_{jk}) \\ E_n^C &= (e_{jk}^C) = (\chi_n(\beta_k) \, e_{jk}) \, . \end{aligned}$$

(b) We now show for the part near the vertical edge that

$$\|E_n^V \bullet G\| \le C \log(n+1) \, \|G\| \, . \tag{26}$$

Let $G = (g_{jk})$ and consider for arbitrary vectors $x = (x_j)$, $y = (y_k)$

$$x^*(E_n^V \bullet G)y = \sum_{j,k} \overline{x}_j \, g_{jk} e_{jk}^V \, y_k \, .$$

Partial summation in horizontal direction gives, with $d^V_{jk} = e^V_{j+1,k} - e^V_{jk}$,

$$x^*(E_n^V \bullet G)y = -\sum_{j,k} \Big( \sum_{i \le j} \overline{x}_i g_{ik} \Big) d^V_{jk} y_k = -\sum_j \sum_{i,k} (\overline{x}_i \, 1_{\{i \le j\}}) \, g_{ik} \, (d^V_{jk} y_k)$$

where $1_{\{i \le j\}} = 1$ if $i \le j$, and zero else. This implies

$$|x^*(E_n^V \bullet G)y| \le \sum_j \|x\| \cdot \|G\| \cdot \|y\| \cdot \max_k |d^V_{jk}| \ ,$$

and hence

$$\|E_n^V \bullet G\| \le \sum_j \max_k |d^V_{jk}| \cdot \|G\| \ . \tag{27}$$

We have

$$\begin{aligned} d^V_{jk} &= e^V_{j+1,k} - e^V_{jk} \\ &= \varphi(\alpha_{j+1}/\beta_k)(1 - \chi_n(\beta_k))(\varepsilon_n(\alpha_{j+1}, \beta_k) - \varepsilon_n(\alpha_j, \beta_k)) + \\ &\qquad (\varphi(\alpha_{j+1}/\beta_k) - \varphi(\alpha_j/\beta_k))(1 - \chi_n(\beta_k))\varepsilon_n(\alpha_j, \beta_k) \ . \end{aligned}$$

By Lemma 4 we have $|\varepsilon_n(\alpha, \beta)| \le C$, and from the formulas in the proof of Lemma 4 one obtains also

$$\left| \frac{\partial \varepsilon_n}{\partial \alpha}(\alpha, \beta) \right| \le C \min(n, 1/\alpha) \quad \text{for } (\alpha, \beta) \in \Delta \text{ with } \alpha/\beta \le 2/3, \tag{28}$$

i.e., for those $(\alpha, \beta)$ for which $\varphi(\alpha/\beta) \ne 0$. Note that $\varphi(\alpha_{j+1}/\beta_k) - \varphi(\alpha_j/\beta_k) \ne 0$ only if $\alpha_{j+1}/\beta_k \ge \frac{1}{3}$ and $\alpha_j/\beta_k \le \frac{2}{3}$, that is, only if $\beta_k \in [\frac{3}{2}\alpha_j, 3\alpha_{j+1}]$. Then we have

$$|\varphi(\alpha_{j+1}/\beta_k) - \varphi(\alpha_j/\beta_k)| \le C \, \frac{\alpha_{j+1} - \alpha_j}{\beta_k} \le C \, \frac{\alpha_{j+1} - \alpha_j}{\frac{3}{2}\alpha_j} \ .$$

On the other hand, we have $\beta_k \ge 1/n$ for all $k$ which give non-vanishing entries in $E_n^V$. Combining these estimates gives

$$|d^V_{jk}| \le C \min(n, 1/\alpha_j) \cdot (\alpha_{j+1} - \alpha_j)$$

and hence

$$\sum_j \max_k |d^V_{jk}| \le C \left( 1 + \int_{1/n}^1 \frac{d\alpha}{\alpha} \right) = C(1 + \log n) \ .$$

Therefore, (27) implies (26).

(c) For the part near the diagonal we show

$$\|E_n^D \bullet G\| \le C \log(n+1) \, \log(N+1) \, \|G\| \ . \tag{29}$$

We proceed similarly to part (b), but now use anti-diagonal partial summation. With $d^D_{jk} = e^D_{j+1,k} - e^D_{jk}$, we have for arbitrary vectors $x\,, y$

$$\begin{aligned} x^*(E^D_n \bullet G)y &= \sum_{j,k} \overline{x}_j\, g_{jk} e^D_{jk}\, y_k = \sum_{j,k} \overline{x}_{j+k}\, g_{j+k,k} e^D_{j+k,k}\, y_k \\ &= -\sum_j \sum_{i,k} \overline{x}_{i+k} 1_{\{i\le j\}} g_{i+k,k} d^D_{j+k,k}\, y_k \\ &= -\sum_j \sum_{i,k} \overline{x}_i \,(1_{\{i-k\le j\}} g_{ik})\,(d^D_{j+k,k} y_k)\ . \end{aligned}$$

(Here we may think of $E^D_n$ and $G$ as being embedded in higher-dimensional matrices by extending them by zero, so that we need not care about the range of summations above.)

The matrix $G^{(j)} = (1_{\{i-k\le j\}} g_{ik})_{i,k}$ is obtained from $G$ by truncating a triangular part. Theorem 1 in [1] (see also references therein for related earlier work) shows that

$$\|G^{(j)}\| \le C\, \log(N+1)\, \|G\|\,, \tag{30}$$

which explains how the factor $\log(N+1)$ comes about. This implies

$$\|E^D_n \bullet G\| \le \sum_j \max_k |d^D_{j+k,k}| \cdot C\, \log(N+1)\, \|G\|\,. \tag{31}$$

In place of (28) we now have

$$\left|\frac{\partial \varepsilon_n}{\partial \alpha}(\alpha,\beta)\right| \le C \min(n, 1/(\beta-\alpha)) \qquad \text{for } (\alpha,\beta) \in \Delta \text{ with } \alpha/\beta \ge 1/3.$$

In the same way as in part (b), this bound together with (31) yields (29).

(d) For the part near the corner we have

$$\|E^C_n \bullet G\| \le C \log(N+1)\, \|G\|\,.$$

This follows as above using partial summation, (30), and the bound

$$\left|\frac{\partial \varepsilon_n}{\partial \alpha}(\alpha,\beta)\right| \le Cn \qquad \text{for } (\alpha,\beta) \in \Delta\,.$$

(e) The same arguments apply also to the complementary triangle $0 \le \beta < \alpha < \pi$ (with vertical edge $\alpha = \pi$, diagonal $\alpha = \beta$, and corner $(\pi,\pi)$), and in fact to every triangle whose corners have successive integer multiples of $\pi$ as coordinates and whose diagonal or anti-diagonal edge lies on one of the lines $\alpha \pm \beta = 2k\pi$ with integer $k$.

Using the decay properties of the error functions for large arguments, see the formulas in the beginning of the proof of Lemma 4, we obtain for

every square $\square_{l,m} = [(l-1)\pi, l\pi) \times [(m-1)\pi, m\pi)$ with $l, m = 1, 2, 3, \ldots$ (each of which is composed of two of the above triangles) the bound

$$\|E_n^{l,m} \bullet G\| \le C \left( \frac{1}{(1+|l^2-m^2|)m} + \frac{1}{l^2 m^2} \right) \cdot \log(n+1)\, \log(N+1)\, \|G\| ,$$

where $E_n^{l,m}$ is defined like $E_n^{\Delta}$ in (25), but with $\square_{l,m}$ in place of $\Delta$. For every integer $k$, the block-diagonal matrix

$$E_n^k = \sum_m E_n^{k+m,m}$$

then satisfies the bound

$$\|E_n^k \bullet G\| \le \max_m \|E_n^{k+m,m} \bullet G\| \le \frac{C}{1+k^2}\, \log(n+1)\, \log(N+1)\, \|G\| ,$$

and consequently

$$\|E_n \bullet G\| \le \sum_k \|E_n^k \bullet G\| \le C\, \log(n+1)\, \log(N+1)\, \|G\| ,$$

which was to be proved. □

## 5 Linear stability

To gain a better understanding of the behaviour of the method and the influence of the filter function $\phi$, we study the long-time error propagation for the linear system

$$y'' = -Ay - By \tag{32}$$

where both $A$ and $B$ are assumed symmetric and positive semi-definite. The method applied to this equation reads

$$y_{n+1} - 2y_n + y_{n-1} = -h^2 \sigma(h^2 A)(A + B\phi(h^2 A)) y_n . \tag{33}$$

It turns out favourable for stability to have a filter function that is non-negative:

$$\phi(x) \ge 0 \qquad \text{for} \quad x \ge 0 . \tag{34}$$

In the following we assume that squares of integer multiples of $\pi$ are the only zeros of $\phi$, and that no eigenvalue of $h\Omega$ is precisely an integer multiple of $\pi$. Then, the matrices

$$S = \sigma(h^2 A)^{1/2} , \quad F = \phi(h^2 A)^{1/2}$$

are non-singular. We introduce transformed variables

$$q_n = FS^{-1} y_n , \quad p_n = FS^{-1} v_n . \tag{35}$$

By (7), we have for all eigencomponents $|q_n^k| \leq C|y_n^k|$, and if the squares of integer multiples of $\pi$ are the only zeros, of multiplicity exactly 2, then we have also an inverse inequality for those components for which $h\omega_k$ is bounded away from an odd multiple of $\pi$. Since $A$, $F$, and $S$ commute, the recursion for $q_n$ has a symmetric matrix:

$$q_{n+1} - 2q_n + q_{n-1} = -h^2(SAS + SFBFS)q_n \ . \tag{36}$$

Let

$$\mu(x^2) = \frac{\phi(x^2)\,\sigma(x^2)}{(\cos\frac{1}{2}x)^2} \ . \tag{37}$$

Note that $\mu(0) = 1$, and

$$\|\mu(h^2A)\| = \max_k \mu((h\omega_k)^2) \leq \sup_{x\geq 0} \mu(x^2) < \infty$$

for filter functions $\phi$ with (7) and (34), because $\phi$ then has at least a double zero at the square of every integer multiple of $\pi$. We have the following stability criterion.

**Theorem 2** *In the above situation, if*

$$\|\mu(h^2A)\| \cdot \|h^2B\| \leq 4 \ , \tag{38}$$

*then the recursion is stable in the sense that*

$$\|q_n\| \leq n\,(\|q_0\| + \|q_1\|) \ , \qquad n > 1 \ .$$

*Proof* By diagonalization of the matrix in (36), it is seen that the recursion is stable if and only if the eigenvalues of $h^2(SAS + SFBFS)$ lie in the interval $[0, 4]$. It is clear that these eigenvalues are non-negative, so it remains to find the upper bound. Let $C = \cos\frac{1}{2}h\Omega$, so that $h^2SAS = 4(\sin\frac{1}{2}h\Omega)^2 = 4I - 4C^2$. We then have

$$h^2(SAS + SFBFS) = 4I - C(4I - h^2C^{-1}SFBFSC^{-1})C \ .$$

Under condition (38), $C(4I - h^2C^{-1}SFBFSC^{-1})C$ is positive semi-definite, and the eigenvalues of $h^2(SAS + SFBFS)$ are then bounded by 4. □

The proof also shows that the condition (38) is *necessary* if the recursion is to be stable for all positive semi-definite matrices $B$ of a fixed norm. This necessity is already obvious in the scalar case.

The stability bound for $q_n$ can be further used to obtain a bound for $y_n$, also for those eigencomponents where the inverse of $FS^{-1}$ is not reasonably bounded. As in Lemma 2, we have

$$y_{n+1} = -W_{n-1}y_0 + W_n y_1 - \sum_{j=1}^{n} W_{n-j} S^2 h^2 B F^2 y_j \ .$$

Noting that $F^2 y_j = FSq_j$, we obtain, with $c = \|FS\|$,

$$\|y_{n+1}\| \le n\|y_0\| + (n+1)\|y_1\| + c\,\|h^2 B\| \sum_{j=1}^{n} (n-j+1)\|q_j\| \ .$$

## 6 Choice of the filter function

A first possible choice of a filter function satisfying (7) is

$$\phi(x^2) = \sin x / x \ . \tag{39}$$

The absolute value of its complex error function $\varepsilon_n(\alpha, \beta)$ defined by (18) is plotted in Figure 1. The figure was computed with $n = 50$, but nearly identical graphs are obtained for all sufficiently large $n$ ($n \ge 10$ or 20, say). A considerably reduced error function is obtained for

$$\phi(x^2) = \frac{\sin x}{x} \left(1 + \tfrac{1}{6}(1 - \cos x)\right) \ . \tag{40}$$

This filter function is chosen such that the integral term in the error function, see (17), becomes small for small $\alpha, \beta$. This requires $\phi(x^2) = 1 - x^2/12 + O(x^4)$ for $x \to 0$. The absolute value of the error function for (40) is plotted in Figure 2. Unfortunately, the filter function (40) becomes negative on intervals between the squares of odd and even multiples of $\pi$ and hence does not satisfy condition (34) required for linear stability in the sense of Section 4. A filter function which satisfies (7) and (34) and whose error function becomes small for small $\alpha, \beta$, is given by

$$\phi(x^2) = \left(\frac{\sin x}{x}\right)^2 \left(1 + \tfrac{1}{2}(1 - \cos x)\right) \ . \tag{41}$$

Its stability threshold function $\mu$, given by (37), satisfies $\mu(x^2) < 1.04$ for all $x \ge 0$. The absolute value of its error function is plotted in Figure 3.

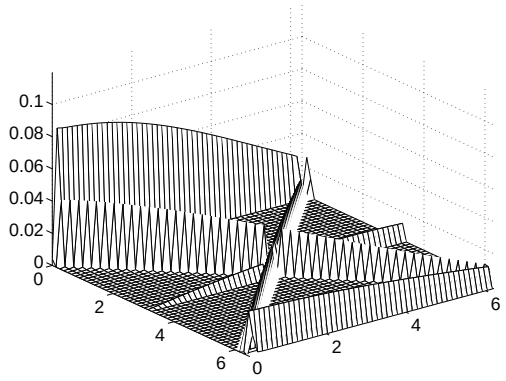



Figure 1. Error function for $\phi(x^2) = \sin x/x$.

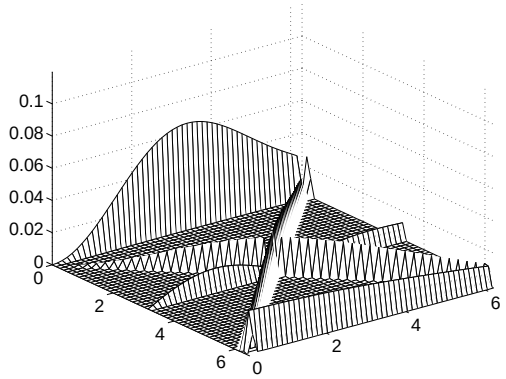



Figure 2. Error function for the filter (40).

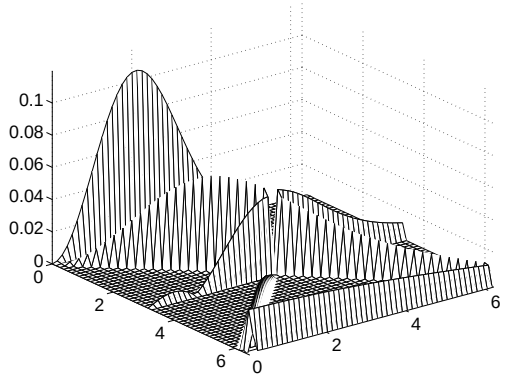



Figure 3. Error function for the filter (41).

## 7 Application to the mollified impulse method

We now show how the above analysis gives new insight into the mollified impulse method of García-Archilla, Sanz-Serna and Skeel [3, 4]. When ap-

plied to Eq. (1), their method reads

$$\begin{aligned} v_n^+ &= v_n + \tfrac{1}{2}h\phi(h^2A)\, g(\phi(h^2A)y_n) \\ \begin{pmatrix} y_{n+1} \\ v_{n+1}^- \end{pmatrix} &= \begin{pmatrix} \cos h\Omega & \Omega^{-1}\sin h\Omega \\ -\Omega \sin h\Omega & \cos h\Omega \end{pmatrix} \begin{pmatrix} y_n \\ v_n^+ \end{pmatrix} \\ v_{n+1} &= v_{n+1}^- + \tfrac{1}{2}h\phi(h^2A)\, g(\phi(h^2A)y_{n+1}) \ , \end{aligned} \tag{42}$$

with a filter function $\phi$ that vanishes at the squares of even multiples of $\pi$. They show second-order error bounds which are independent of the frequencies and of the dimension of the system.

Upon eliminating the (non-averaged) velocities, the scheme (42) becomes

$$y_{n+1} - 2y_n + y_{n-1} = h^2\sigma(h^2A)(-Ay_n + g_n) + h^2\delta(h^2A)\, g_n \ , \tag{43}$$

where $\delta = \psi\phi - \sigma$, with $\psi(x^2) = \sin x/x$. With minor modifications, the error analysis of Section 3 applies also to (43) and consequently to (42). The role of the error function is now taken by

$$\varepsilon_n^{\mathrm{MIM}}(\alpha,\beta) = \varepsilon_n(\alpha,\beta) - \delta(\alpha^2)\,\phi(\beta^2) \sum_{j=0}^{n-1} \frac{\sin(j+1)\alpha}{\sin\alpha}\,(e^{-ij\beta} - e^{-in\beta}) \ . \tag{44}$$

Figure 4 shows the absolute value of this error function (for $n = 50$) for the filter $\phi = \psi$, which is a favoured choice in [3] (the long-average method). In contrast to the situation in Section 6, it is now not possible to construct a filter function such that the error function (44) becomes arbitrarily small near $(0,0)$.

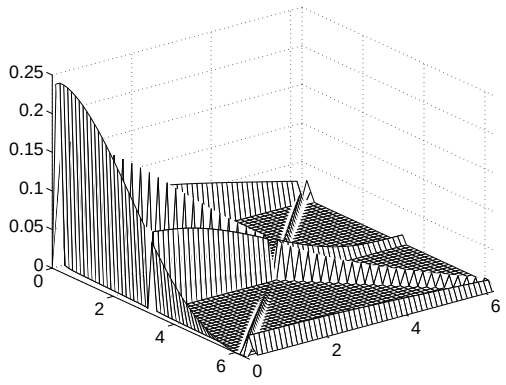



Figure 4. Error function (44) for $\phi(x^2) = \sin x/x$.

The error bounds of [3] applied to the equation $y'' = -Ay + \epsilon G(y - y_0)$, with $\epsilon \to 0$ and an arbitrary matrix $G$, can be shown to imply for the entrywise product of $E_n^{\mathrm{MIM}} = (\varepsilon_n^{\mathrm{MIM}}(h\omega_j, h\omega_k))_{j,k=1}^N$ with $G$ the bound $\|E_n^{\mathrm{MIM}} \bullet G\| \le C\,\|G\|$, without the logarithms that we did not succeed to eliminate in Lemma 5.

In addition to the error terms that were present also in Section 3, there is now an additional term in the error of the mollified implulse method which results from not solving equations with a constant inhomogeneity exactly. Consider the method (42) applied to a linear problem (1) with constant inhomogeneity $g$. Then, the error after the first step is $e_1 = \frac{1}{2}h^2\delta(h^2A)g$, and the defect in (43) is $d_n = -h^2\delta(h^2A)g$. By Lemma 2 and a trigonometric identity, we thus have for the error in the $(n+1)$st step

$$e_{n+1} = W_n e_1 - \sum_{j=1}^{n} W_{n-j}\, d_j = \tfrac{1}{2}h^2(I - \cos(n+1)h\Omega)\,\epsilon(h^2A)g$$

with $\epsilon(x^2) = \delta(x^2)/(1-\cos x)$. For $\phi = \psi$ this function is bounded in modulus by $\frac{1}{2}$, so that $\|e_{n+1}\| \le \frac{1}{2}h^2\|g\|$. Interestingly, the two-step scheme (43) with exact starting values ($e_0 = e_1 = 0$) does not give an $O(h^2)$ error bound uniformly in the frequencies. It produces an $O(nh^2)$ error term if, for some frequencies, $h\omega_k$ is close to an odd multiple of $\pi$.

The stability result of Theorem 2 does not extend unchanged to the mollified impulse method (42). In fact, the analysis of 2-dimensional linear systems in [3] shows that there exists no positive constant $c$ such that $\|h^2B\| \le c$ implies stability without restrictions on $h^2A$, unless $\phi(x^2)$ vanishes for all $x$ where $\psi(x^2)$ is negative. A straightforward adaptation of the proof of Theorem 2 shows that this latter condition on the filter function is also sufficient for the stability of (42) for equations (32) in arbitrary dimensions whenever $\|\mu(h^2A)\| \cdot \|h^2B\| \le 4$, where now $\mu(x^2) = \psi(x^2)\phi(x^2)^2/(\cos\frac{1}{2}x)^2$.

Both methods (10) and (42) are obviously time-reversible. An interesting property of (42) is its *symplecticness*, that is, the map $(y_n, v_n) \mapsto (y_{n+1}, v_{n+1})$ is symplectic when the method (42) is applied to (1) with $g(y) = -\nabla U(y)$ [3]. For the method (10), the one-step map is not symplectic in the variables $(y, v)$, but by comparison with the Störmer/Verlet method it is easily verified that it is symplectic in the transformed variables $(q, p)$ of (35). At present it is not clear, however, what is the significance of symplecticness of either method for the long-time behaviour of numerical solutions. There is no backward error analysis available which would, for example, guarantee long-time near-conservation of energy, unless $\|h^2A\| \ll 1$ which is not what these methods are meant for.

## 8 Numerical experiments

In this section we report on some numerical experiments with the sine-Gordon equation

$$u_{tt} = u_{xx} - \sin u \ ,$$

which we consider for $x \in [-1, 1]$ with periodic boundary conditions. Pseudospectral discretization in space with $N$ equidistant collocation points $x_j$ yields an approximation $U(t) = (U_j(t))_{j=1}^N$ with $U_j(t) \approx u(x_j, t)$. Its discrete Fourier transform

$$y(t) = \mathcal{F}_N U(t)$$

satisfies

$$y'' = -Ay - \mathcal{F}_N \sin(\mathcal{F}_N^{-1} y) \ ,$$

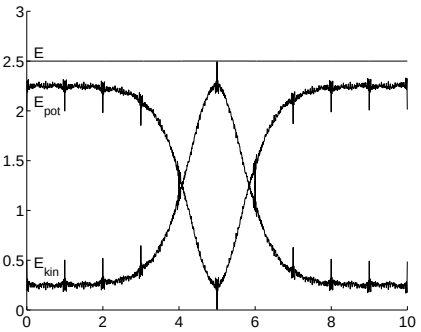
where $A = \mathrm{diag}(\omega_k^2)$ with

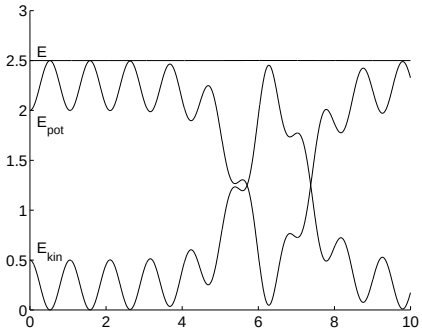
$$\omega_k = \begin{cases} k\pi & k = 0, \ldots, N/2 - 1 \\ (N-k)\pi & k = N/2, \ldots, N-1 \ . \end{cases}$$

We chose $N = 128$ and the initial position $U_j(0) = \pi$ for all $j$, and we considered two choices of initial velocities, corresponding to non-smooth and smooth solutions.

In the first case we chose $U'(0)$ as a vector of normally distributed random numbers scaled to Euclidean norm $\sqrt{N}$. (This is reproduced by the following Matlab 5 sequence: `randn('state',0); v=randn(N,1); v=v/norm(v)*sqrt(N)`.) Figure 5 shows the evolution of potential and kinetic energy in the time interval $[0, 10]$.

In the second case we chose $U'(0)$ as a scalar multiple of $(0.01 + \sin(2\pi j/N))_{j=1}^N$, again scaled to Euclidean norm $\sqrt{N}$. Potential and kinetic energy in the interval $[0, 10]$ are shown in Figure 6.

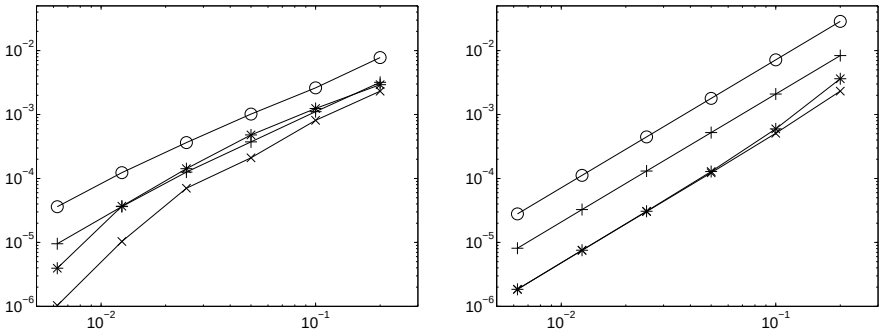



Figures 5 and 6. Kinetic and potential energies for two initial states.

For these two cases, Figures 7 and 8 plot the Euclidean norm (scaled by $1/\sqrt{N}$) of the error in the positions $U$ at $t = 10$ versus the step size. (Reference values were obtained by applying the methods with small step sizes.) The methods used are the mollified impulse method with the 'long-average' filter $\phi(x^2) = \sin x/x$ (shown with markers $\circ$), and the method (10) with the same filter (markers $+$) and with the filters (40) and (41) (with markers $\times$ and $*$, respectively). Taking no filter at all ($\phi \equiv 1$) in (10), which is not shown in the figures, gave errors more than an order of magnitude larger than for the most accurate filter (40) and a more erratic error curve in the nonsmooth example, and about the same errors as the 'long-average' filter (+) in the smooth example.

Very similar figures were obtained also for the errors in the velocities.




Figures 7 and 8. Errors versus step size.
$\circ$ : mollified impulse method with long-average filter (39)
$+, \times, *$ : Gautschi-type method with filters (39), (40), (41)

In experiments with different data, we did not always observe such a clear difference between the methods. For example, with initial positions $U_j(0) = \frac{1}{2}\pi$ and the same initial velocities as before, the error curves differed by less than a factor 2. The filters (40) and (less so) (41) were found advantageous throughout.

We tested also energy conservation on the interval $[0, 1000]$. We did not observe an energy drift for the methods and step sizes considered above.

*Acknowledgements* We are grateful to Gerhard Wanner for helpful comments on the presentation.

## References

1. J. R. Angelos, C. C. Cowen, and S. K. Narayan. Triangular truncation and finding the norm of a Hadamard multiplier. Linear Algebra Appl., 170:117–135, 1992.

2. V. L. Druskin and L. A. Knizhnerman. Krylov subspace approximations of eigenpairs and matrix functions in exact and computer arithmetic. Numer. Lin. Alg. Appl., 2:205–217, 1995.
3. B. García-Archilla, J. M. Sanz-Serna, and R. Skeel. Long-time-step methods for oscillatory differential equations. Applied Mathematics and Computation Reports 1996/7, Universidad de Valladolid, 1996. To appear in SIAM J. Sci. Comput.
4. B. García-Archilla, J. M. Sanz-Serna, and R. Skeel. The mollified impulse method for oscillatory differential equations. Applied Mathematics and Computation Reports 1997/5, Universidad de Valladolid, 1997. To appear in Proceedings of the 1997 Dundee Conference, D.F. Griffiths and G.A. Watson, eds.
5. W. Gautschi. Numerical integration of ordinary differential equations based on trigonometric polynomials. Numer. Math., 3:381–397, 1961.
6. M. Hochbruck and Ch. Lubich. On Krylov subspace approximations to the matrix exponential operator. SIAM J. Numer. Anal., 34(5):1911–1925, 1997.
7. M. Hochbruck, Ch. Lubich, and H. Selhofer. Exponential integrators for large systems of differential equations. SIAM J. Sci. Comput., 1998. To appear.
8. L. R. Petzold, L. O. Jay, and J. Yen. Numerical solution of highly oscillatory ordinary differential equations. Acta Numerica, 7:437–483, 1997.