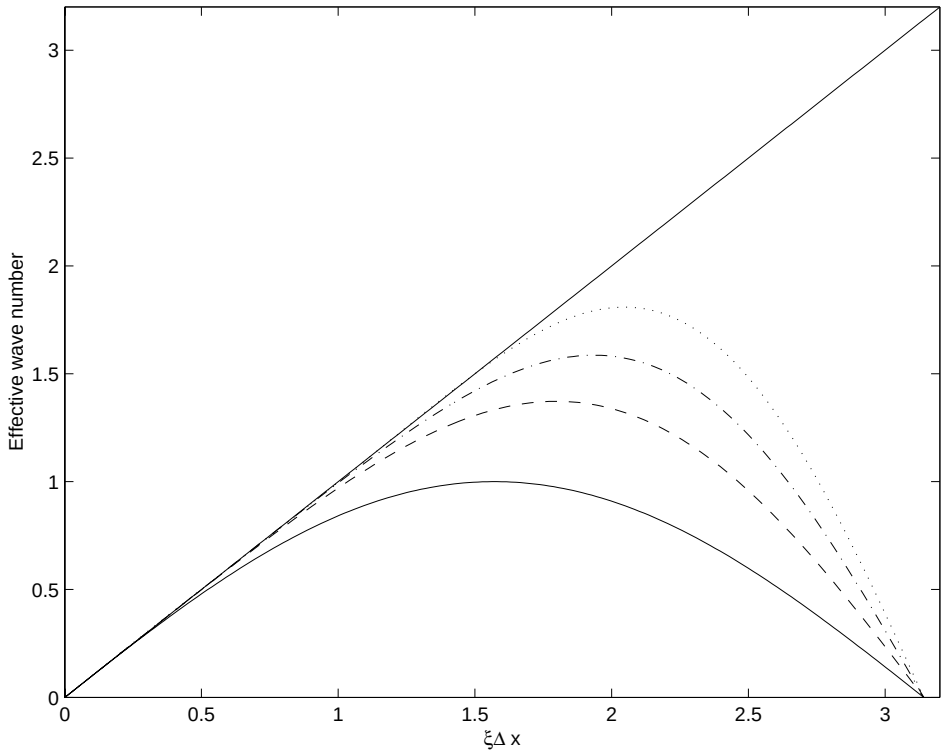
FOI-R–0407–SE
March 19, 2002
ISSN 1650-1942

Scientific report

Stefan Jakobsson

Frequency optimized computation methods



Division of Aeronautics, FFA
SE-172 90 STOCKHOLM

FOI – Swedish Defence Research Agency
Division of Aeronautics, FFA
SE-172 90 STOCKHOLM

FOI-R–0407–SE
March 19, 2002
ISSN 1650-1942

Scientific report

Stefan Jakobsson

Frequency optimized computation methods

Division of Aeronautics, FFA
SE-172 90 STOCKHOLM

Abstract

In this paper we develop an alternative method to derive finite difference approximations of derivatives. The purpose is to find schemes which work for a broader range of frequencies than the usual approximations based on polynomial fitting and Taylor's Theorem to the expense of less accuracy for low frequencies. The numerical schemes are obtained as solutions to constrained optimizations problems in a weighted L^2-norm in the frequency domain. We examine the accuracy of these schemes and compare them with the standard approximations. We also use the same approach to derive numerical schemes for time integration for differential equations with time independent operators. To test the accuracy of the different schemes, we study dispersion errors for a simple wave equation in one space dimension. We examine the number of points per wave length which is needed in order for the relative error in the phase velocity to be below a certain bound. A similar examination is carried out for the different time integration schemes.

1 Introduction

To solve a partial differential equation numerically one approximates the equation by a discrete equation and then solves this equation by a computer. In many applications this discrete equation contains a huge set of variables. This is in particular the case for radar cross section calculations for an aircraft. The radar cross section measures how visible the aircraft is for the radar. The radar transmits an electro-magnetic wave of a certain frequency (or rather a wave packet with a range of frequencies). The wave is scattered by the aircraft and the scattered wave is detected by the radar. The governing equations for this problem are the Maxwell equations and the typical range of frequencies is about 1-10 G Hz which corresponds to a wave length of about 3-30 cm. Many numerical methods require 10-20 grid points per wave length in order to resolve the wave accurately. For a typical aircraft this leads to a discrete equation with a huge number of variables. By choosing an efficient numerical method this number might be reduced. A common way to do this is to choose a higher order method. Such schemes are derived by using Taylor's theorem and polynomial fitting.

In this paper we propose an alternative method to derive the approximations which can reduce the number of variables. We consider the finite difference approximation in the frequency domain and choose our approximation such that it gives good result for a range of frequencies at the expense of less accuracy for low frequencies. The schemes are obtained as solutions to optimization problems in weighted L^2-norms in the frequency domain. The optimization may be subject to a linear constraint given by polynomial fitting. Similar ideas have been studied before by other authors, see for example Efraimsson [1], and Tam and Webb [9]. In [7], Tam and Kurbatskii studied extrapolation and interpolation using this approach.

In Section 2 we give some background to finite difference approximations and in Section 3 we derive the optimized schemes. We also use the same idea to derive time stepping methods for time dependent linear equations and this is carried out in Section 4 and 5. In Section 6 we discuss dispersion and numerical dispersion for a simple wave equation in one space dimension. We define a criterion in terms of the relative error in the phase velocity in order to measure how many points per wave length is needed for different finite difference approximations and levels of accuracy. A similar measure is defined for time integration schemes. These criteria are then applied in Section 7 to both standard and optimized schemes in order to evaluate there performance on wave problems.

2 Finite difference approximations of derivatives

One of the most basic problems in numerical analysis is to approximate the derivatives of a function at a point by using the function values at some neighbor points. In this section we begin by describing the standard way to do this by finite differences, that is, by using polynomial fitting and Taylor's Theorem. We continue by looking at finite difference approximations in the frequency domain. This will lead us to the main point of the paper, frequency optimized finite difference approximations which is described in the following section.

2.1 Polynomial fitting

Suppose we want to approximate the n-th derivative of a function f at the point x by using the function values f at the points $\{x_j\}_{j\in\mathbf{J}}$, where $\mathbf{J}$ is a set of indexes (the points x_j do not necessarily have to be equidistant). The linear nature of this problem suggest that a general finite difference approximation of $f^{(n)}(x)$ can be written as

$$f^{(n)}(x) \approx \sum_{j\in\mathbf{J}} a_j(x) f(x_j), \tag{1}$$

where the coefficients $a_j(x)$ depends on the point x. For example, the simplest first order approximation of the first derivative at x_1 by using the function values at the grid points x_1 and x_2 is

$$f'(x_1) \approx \frac{f(x_2) - f(x_1)}{x_2 - x_1}.$$

The standard way to find the coefficients $a_j(x)$, $j \in \mathbf{J}$, is to use polynomial fitting and Taylor's Theorem. For equidistant grids, these two methods coincide. To approximate the derivative by polynomial fitting we take the polynomial P_f of minimal degree which interpolates f at the points x_j: $f(x_j) = P_f(x_j)$, $j \in \mathbf{J}$. The approximation of the derivative is then

$$f^{(n)}(x) \approx P_f^{(n)}(x). \tag{2}$$

Since interpolation also is a linear problem this approximation is of the form (1). In fact, the coefficients $\mathbf{a}(x) = (a_j(x))_{j\in\mathbf{J}}$ can be found as the solution to the linear equation system

$$p_m^{(n)}(x) = \sum_{j\in\mathbf{J}} a_j(x) p_m(x_j), \qquad m = 0, 1, \ldots, \sharp\mathbf{J} - 1, \tag{3}$$

where $p_m(x) = x^m$ and $\sharp\mathbf{J}$ is the total number of indexes in $\mathbf{J}$. The matrix $\mathbf{P} = (p_m(x_j))_{j,m\in\mathbf{J}}$ is called the *Vandermonde matrix* for the points $\{x_j\}_{j\in\mathbf{J}}$. It is easy to check that the approximation of the first derivative above can be derived in this way.

2.1.1 Equidistant grids

All comparisons will consider central differences approximations on equidistant grids (although the derivation of the formulas will be on general grids). For an

equidistant grid $\{x_j\}_{j=-N}^{N}$, $x_j = x + j\Delta x$ with mesh size $\Delta x > 0$, we write

$$f^{(n)}(x) \approx \frac{1}{(\Delta x)^n} \sum_{j=-N}^{N} a_j f(x_j) \tag{4}$$

for a central difference approximation of the n-th derivative. The factor $\frac{1}{(\Delta x)^n}$ in front makes the coefficients $\{a_j\}_{j=-N}^{N}$ independent of the mesh size for schemes derived by polynomial fitting. We recall the following classical error estimate for finite difference approximations derived by polynomial fitting on equidistant grids

$$\left| f^{(n)}(x) - \frac{1}{(\Delta x)^n} \sum_{j=-N}^{N} a_j f(x_j) \right| \leq C(\Delta x)^{2N+1-n} \|f^{(2N+1)}\|_\infty,$$

for some constant C provided that $f \in C^{2N+1}([x - N\Delta x, x + N\Delta x])$. The exponent for Δx, here $2N + 1 - n$, is called the order of the approximation.

2.2 Finite difference approximations in the Fourier domain

The above result gives a good error estimate provided that f is sufficiently smooth. If this is not the case, it is natural to ask whether a lower order method is more accurate than a higher order for such functions. To answer this type of questions, at least partially, it is convenient to study finite difference approximations in the frequency domain and this will also be a natural starting point for our frequency optimized finite difference schemes. We continue by recalling some facts about Fourier transforms.

The Fourier transform of a function $f \in L^1(\mathbb{R})$ is given by the integral

$$\mathcal{F}[f](\xi) = \hat{f}(\xi) = \int_{-\infty}^{\infty} f(x) e^{-i\xi x}\, dx, \qquad \xi \in \mathbb{R}. \tag{5}$$

According to Fourier's inversion formula, f can be recovered from its Fourier transform via

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{f}(\xi) e^{i\xi x}\, d\xi, \qquad x \in \mathbb{R},$$

provided that $\hat{f} \in L^1(\mathbb{R})$ as well. The variable ξ is often called the *wave number*. For a monochromatic wave $e^{i\xi x}$, the wave number is related to the wave length as

$$\lambda = \frac{2\pi}{\xi}.$$

On the Fourier side, the finite difference approximation (1) becomes

$$\frac{1}{2\pi} \int_{-\infty}^{\infty} (i\xi)^n \hat{f}(\xi) e^{i\xi x}\, d\xi \approx \frac{1}{2\pi} \int_{-\infty}^{\infty} \sum_{j \in \mathbf{J}} a_j(x) e^{i\xi(x_j - x)} \hat{f}(\xi) e^{i\xi x}\, d\xi, \tag{6}$$

where we have used Fourier's inversion formula and the fact that a differentiation of f corresponds to a multiplication of $\hat{f}$ by $i\xi$ on the Fourier side. It is clear that $\sum_{j \in \mathbf{J}} a_j(x) f(x_j)$ approximates $f^{(n)}(x)$ well if the difference $(i\xi)^n -$

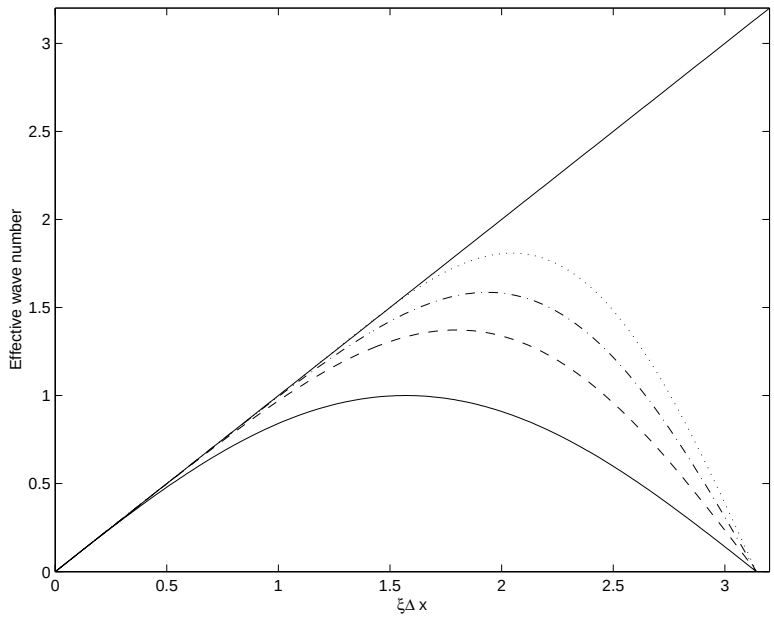


Figure 1. A comparison between Taylor methods of different orders: Taylor order 2, ——, Taylor order 4,– – ––,Taylor order 6, – · – · –, The DRP scheme for $\xi_{\rm opt} = 1.6$, · · · · · · ·.

$\sum_{j\in\mathbf{J}} a_j(x)e^{i\xi(x_j-x)}$ is small for all frequencies ξ in the support of $\hat{f}$. Even if $\hat{f}$ has not compact support but decays fast at infinity, as for smooth functions, the approximation might be good. In particular, if $f \in C^N(\mathbb{R})$ has compact support then

$$\hat{f}(\xi) = \frac{1}{(i\xi)^N}\frac{1}{2\pi}\int_{-\infty}^{\infty} f^{(N)}(x)e^{i\xi x}\,dx, \qquad \xi \neq 0$$

(this formula can be easily be proved by integration by parts), which shows that $\hat{f}$ decays as $1/\xi^N$ for large ξ. For the coefficients $a_j(x)$, on the other hand, one can prove that if they are determined by polynomial fitting then

$$\left((i\xi)^n - \sum_{j\in\mathbf{J}} a_j(x)e^{i\xi(x_j-x)}\right) = \mathcal{O}(|\xi|^{\sharp\mathbf{J}}) \tag{7}$$

for small ξ, where x is the point where the derivative is approximated. For example, for the two point set we considered above we have

$$i\xi - i\frac{(\exp(i\xi(x_2-x_1))-1)}{i(x_2-x_1)} = \mathcal{O}(|\xi|^2),$$

where

$$a_1 = -\frac{1}{x_2-x_1}, \qquad a_2 = \frac{1}{x_2-x_1}$$

and $x = x_1$. We define $\widetilde{\xi^n}(x,\xi)$ as the approximation of ξ^n induced by the finite difference scheme, see (7),

$$\widetilde{\xi^n}(x,\xi) = (-i)^n \sum_{j\in\mathbf{J}} a_j(x)e^{i\xi(x_j-x)}.$$

For $n = 1$, we have $\widetilde{\xi}(x, \xi)$ which we call the effective wave number. We will use these variables to measure the performance of the approximations. Since all our comparisons will be central differences on equidistant grids it is worthwhile to treat this case separately. Then we shall use the normalized wave number $\xi\Delta x$ and

$$\widetilde{(\xi\Delta x)^n}(\xi) = (-i)^n \sum_{j=-N}^{N} a_j e^{ij\xi\Delta x}$$

where we have used normalized coefficients as in (4). In Figure 1 we plot $\widetilde{\xi\Delta x}$ for Taylor approximations of the first derivative of order 2, 4 and 6 and the DRP scheme due to Tam and Webb for $\xi_{\mathrm{opt}} = 1.6$. The definition of ξ_{opt} will be given in the next section. The figure shows that a higher order approximation adapt better to $\xi\Delta x$ on a longer interval than a lower order. This shows that a higher order method might be better than a lower order even though f is not sufficiently smooth. For the optimized schemes which we construct below we give up some accuracy at the origin in the Fourier domain in order to have better accuracy within a specific chosen band of frequencies. This is exemplified in the figure by the DRP scheme. As is indicated by the figure, π is an upper limit for the normalized effective wave number to approximate $\xi\Delta x$ for any finite difference approximation. This means that we need at least two points per wave length in order resolve a wave.

3 Frequency optimization

In this section we introduce a method which we call frequency optimization to derive finite difference approximations of derivatives. The schemes are obtained as solutions to optimization problem in a weighted L^2-norm in the frequency domain. We also combine this method with polynomial fitting in order to derive schemes which are partly optimized but also have some order of accuracy at the origin. This method was used by Tam to derive what he called the *dispersion relation preserving* scheme [9]. In Section 7 we shall study both the normalized effective wave number and dispersion relations for these schemes and compare them with standard higher order schemes.

3.1 The optimization problem

Let ν be a function whose Fourier transform is positive on $\mathbb{R}$, is strictly positive on the interval $[-1,1]$ and belongs to $L^1(\mathbb{R})$. Furthermore, we assume that ν, and therefore also $\hat{\nu}$, are even functions. Examples of such functions will be given in Subsection 3.2. We define the Hilbert space $L^2_{\nu,\xi_{\text{opt}}}(\mathbb{R})$ as the set of functions g such that

$$\|g\|^2_{\nu,\xi_{\text{opt}}} = \frac{1}{2\pi\xi_{\text{opt}}} \int_{-\infty}^{\infty} |g(\xi)|^2 \hat{\nu}(\xi/\xi_{\text{opt}})\, d\xi < \infty, \tag{8}$$

where $\xi_{\text{opt}} > 0$ is a parameter. The corresponding inner product is

$$\langle g, h\rangle_{\nu,\xi_{\text{opt}}} = \frac{1}{2\pi\xi_{\text{opt}}} \int_{-\infty}^{\infty} g(\xi)\overline{h(\xi)}\hat{\nu}(\xi/\xi_{\text{opt}})\, d\xi.$$

The parameter ξ_{opt} should be thought of as the highest wave number which we take into account in our optimization problem. Consider an operator Q of the form

$$Q[f](x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} q(x,\xi)\hat{f}(\xi)e^{i\xi x}\, d\xi, \qquad x \in \mathbb{R}.$$

Operators of this type are called pseudodifferential operators and the function q is called the symbol of Q. It is clear that this includes all differential operators and many integral operators. In particular, if

$$Q[f](x) = \sum_{n=0}^{M} c_n(x) \frac{d^n}{dx^n} f(x).$$

then

$$q(x,\xi) = \sum_{n=0}^{M} c_n(x)(i\xi)^n.$$

As in the previous section we seek finite difference approximations of $Q[f](x)$ which use the function values of f at the points $\{x_j\}_{j\in\mathbf{J}}$

$$Q[f](x) \approx \sum_{j\in\mathbf{J}} a_j(x) f(x_j).$$

Clearly, the vector $\mathbf{a}(x) = (a_j(x))_{j\in\mathbf{J}}$ of coefficients determine the approximation completely. The purpose of the section is to develop a general method, which

includes polynomial fitting, to obtain such vectors. By using the inverse Fourier transform we can write the approximation as

$$Q[f](x) \approx \frac{1}{2\pi} \int_{-\infty}^{\infty} a_{\mathbf{J}}(x,\xi) \hat{f}(\xi) e^{i\xi x}\, d\xi, \tag{9}$$

where

$$a_{\mathbf{J}}(x,\xi) = \sum_{j \in \mathbf{J}} a_j(x) e^{i\xi(x_j - x)}. \tag{10}$$

is the symbol for the approximation. In order for the finite difference scheme to be a reasonable approximation of Q it is natural to require that the symbol $a_{\mathbf{J}}(x,\xi)$ approximates $q(x,\xi)$ well in some sense. For fixed weight ν, parameter ξ_{opt} and $x \in \mathbb{R}$ we will choose the symbol $a_{\mathbf{J}}(x,\cdot)$ as the best approximation of $q(x,\cdot)$ in the norm $\|\cdot\|_{\nu,\xi_{\text{opt}}}$ perhaps subject to a linear constraint given by polynomial fitting. This will determine the vector $\mathbf{a}(x)$ completely.

3.1.1 The linear constraint

We want the finite difference approximation to give correct result for all polynomials up to a certain degree, say L. Let $p_l(x) = x^l$ be the l-th monomial. The condition is

$$Q[p_l](x) = \sum_{j \in \mathbf{J}} a_j(x) p_l(x_j), \qquad l = 0, \ldots, L.$$

Of course, we must have $L \leq \sharp\mathbf{J} - 1$ otherwise the condition is over determined. If we let $\mathbf{q}(x) = (Q[p_0](x), \ldots, Q[p_L](x))$ and $\mathbf{P} = (P_{kl})_{k \in \mathbf{J}, 0 \leq L}$, $P_{kl} = p_l(x_k)$ (the first $L+1$ columns of the Vandermonde matrix), then the constraint is equivalent to the matrix equation

$$\mathbf{q}(x) = \mathbf{a}(x)\mathbf{P} \tag{11}$$

We can now formulate the optimization problem mathematically: Let the weight ν, parameter ξ_{opt} and $x \in \mathbb{R}$ be fixed. Find the extremal vector $\mathbf{a}(x) = (a_j(x))_{j \in \mathbf{J}}$ to

$$\inf_{\mathbf{q}(x) = \tilde{\mathbf{a}}(x)\mathbf{P}} \left\| q(x,\cdot) - \sum_{j \in \mathbf{J}} \tilde{a}_j(x) e^{i\cdot(x_j - x)} \right\|^2_{\nu,\xi_{\text{opt}}}. \tag{12}$$

If the coefficients are determined in this manner we say that the corresponding scheme is frequency optimized. For the optimization to make sense it is of course necessary that $q(x,\cdot) \in L^2_{\nu,\xi_{\text{opt}}}(\mathbb{R})$. The following theorem and the proceeding corollaries show how the coefficient for the frequency optimized finite difference scheme can be found.

THEOREM 3.1. *Let ν, ξ_{opt} and x be fixed. Then there exists a unique extremal vector $\mathbf{a}(x)$ to (12). The vector can be found as the solution to the linear equation system*

$$(\mathbf{a}\ \lambda) \begin{pmatrix} \mathbf{M} & \mathbf{P} \\ \mathbf{P}^* & 0 \end{pmatrix} = (\mathbf{b}\ \mathbf{q})$$

where the matrix $\mathbf{M} = (M_{kl})_{k,l\in\mathbf{J}}$ *and the row vector* $\mathbf{b} = (b_l)_{l\in\mathbf{J}}$ *have the components*

$$M_{kl} = \nu(\xi_{\text{opt}}(x_k - x_l)),$$

$$b_l(x) = \frac{1}{2\pi\xi_{\text{opt}}}\int_{-\infty}^{\infty} q(x,\xi)e^{i(x-x_l)\xi}\hat{\nu}(\xi/\xi_{\text{opt}})\,d\xi,$$

respectively. The matrix $\mathbf{P}$ *and the vector* $\mathbf{q}(x)$ *are as in equation (11) and* $\lambda(x) = (\lambda_0(x),\ldots,\lambda_L(x))$ *is the vector of Lagrangian multipliers for the problem. Moreover,*

$$\left\| q(x,\cdot) - \sum_{j\in\mathbf{J}} a_j(x)e^{i\cdot(x_j-x)} \right\|^2_{\nu,\xi_{\text{opt}}} = \|q(x,\cdot)\|^2_{\nu,\xi_{\text{opt}}} - 2\,\text{Re}\,(\mathbf{ba}^*)(x) + \mathbf{aMa}^*(x),$$

where $\mathbf{a}^*$ *is the Hermitian conjugate of* $\mathbf{a}$*.*

Proof. The theorem follows almost immediately from the results in Appendix A on constrained minimization in complex Hilbert spaces, Lemma A.1. We only have to verify that the matrix $\mathbf{M}$ and the vector $\mathbf{b}$ are given as above. According to Lemma A.1 $\mathbf{M}$ is the mass matrix for the basis $\{e^{i\cdot(x_j-x)}\}_{j\in\mathbf{J}}$ in $L^2_{\nu,\xi_{\text{opt}}}(\mathbb{R})$. We have

$$M_{kl} = \langle e^{i\cdot(x_k-x)}, e^{i\cdot(x_l-x)}\rangle_{\nu,\xi_{\text{opt}}} = \frac{1}{2\pi\xi_{\text{opt}}}\int_{-\infty}^{\infty} e^{i\xi(x_k-x_l)}\hat{\nu}(\xi/\xi_{\text{opt}})\,d\xi$$
$$= \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{i\xi_{\text{opt}}\xi(x_k-x_l)}\hat{\nu}(\xi)\,d\xi = \nu(\xi_{\text{opt}}(x_k-x_l))$$

which is the same formula as above. Here we have used the inversion formula in the last equality. In the same manner we verify the formula for the components of the vector $\mathbf{b}$

$$b_l(x) = \langle q(x,\cdot), e^{i\cdot(x_l-x)}\rangle_{\nu,\xi_{\text{opt}}} = \frac{1}{2\pi\xi_{\text{opt}}}\int_{-\infty}^{\infty} q(x,\xi)e^{i\xi(x-x_l)}\hat{\nu}(\xi/\xi_{\text{opt}})\,d\xi.$$

The value of the infimum problem follows from expanding the norm and the definition of the matrix $\mathbf{M}$ and the vectors $\mathbf{a}(x)$ and $\mathbf{b}(x)$. □

For unconstrained minimization we have the following simplification.

COROLLARY 3.2. *The coefficient vector for the optimal solution for (12) without any constraint is given as the solution to*

$$\mathbf{a}(x)\mathbf{M} = \mathbf{b}(x).$$

We remark that since $\mathbf{M}$ is a mass matrix it is strictly positive definite and therefore invertible.

The operators Q which we are primarily interested in are the the differential operators $\frac{d^n}{dx^n}$ which have the symbols $q_n(\xi) = (i\xi)^n$, $n = 0,1,2,\ldots$ ($n = 0$

corresponds to the identity operator). These operators have two important properties: they are translation invariant, which is reflected by that their symbols only depend on ξ, and their symbols are positively homogeneous of degree n, that is,

$$q(\alpha\xi) = \alpha^n q(\xi)$$

for all $\alpha > 0$. The next corollary shows how the the expressions for the coefficients $b_l(x)$ can be simplified for such symbols.

COROLLARY 3.3. *For a translation invariant operator Q with positively homogeneous symbol of order n we have*

$$b_l(x) = \xi_{\text{opt}}^n Q[\nu](\xi_{\text{opt}}(x - x_l)).$$

In particular, if Q is the n-th derivative then

$$b_l(x) = \xi_{\text{opt}}^n \nu^{(n)}(\xi_{\text{opt}}(x - x_l)). \tag{13}$$

Proof. The result follows from a linear change of variable in the formula for b_l

$$\begin{aligned} b_l(x) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} q(\xi_{\text{opt}}\xi) e^{i\xi_{\text{opt}}\xi(x-x_l)}\hat{\nu}(\xi)\,d\xi \\ &= \xi_{\text{opt}}^n \frac{1}{2\pi}\int_{-\infty}^{\infty} q(\xi) e^{i\xi_{\text{opt}}\xi(x-x_l)}\hat{\nu}(\xi)\,d\xi = \xi_{\text{opt}}^n Q[\nu](\xi_{\text{opt}}(x - x_l)). \end{aligned} \tag{14}$$

□

3.2 Examples of weight functions

We will now give three examples of functions ν which we will use in the numerical tests in Section 7. To simplify the calculation of the finite difference approximations, we must be able to find closed formulas for both the function and its derivatives.

Example 3.4 (The sinc function). *Let $\hat{\nu}$ be the characteristic function of the interval $[-1, 1]$ times π*

$$\hat{\nu}(\xi) = \begin{cases} \pi, & |\xi| \le 1, \\ 0, & |\xi| > 1. \end{cases}$$

A calculation shows that

$$\nu(x) = \frac{\pi}{2\pi}\int_{-1}^{1} e^{i\xi x}\,d\xi = \frac{\sin(x)}{x} = \operatorname{sinc}(x/\pi)$$

where $\operatorname{sinc}(t) = sin(\pi t)/(\pi t)$ is the function which appears in connection with Shannon's sampling Theorem in signal processing. The DRP-scheme due to Tam and Webb which we mentioned in the introduction is a seven point central difference approximation which is given as the optimal solution with respect to this weight and the side condition to give correct result for polynomils up to order 4. Tam and Shen suggsted to use $\xi_{\text{opt}} = 1.1$, see [6] and [8].

As an example of weight with non-compact support we choose the Gauss function.

Example 3.5 (The Gauss function). *Let*

$$\nu(x) = e^{-x^2/(2\pi^2)}.$$

The function ν is, apart from a constant factor, its own Fourier transform

$$\hat{\nu}(\xi) = \sqrt{2\pi}\pi e^{-\pi^2\xi^2/2}.$$

All derivatives of ν are of the form

$$\nu^{(n)}(x) = \pi^{-n} P_n(x/\pi) e^{-x^2/(2\pi^2)}.$$

where P_n is a polynomial of degree n. The polynomials satisfy the following recursion relation

$$P_{n+1}(x) = -xP_n(x) + P_n'(x).$$

In the last example we choose ν as the Bessel function of the first kind.

Example 3.6 (The Bessel function). *For our last example we define*

$$\hat{\nu}(\xi) = \frac{2}{\sqrt{1-\xi^2}}$$

for $|\xi| < 1$ and zero otherwise. If we apply the inverse Fourier transform to this function we obtain the Bessel function of the first kind and order zero

$$\nu(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{i\xi x}\hat{\nu}(\xi)\,d\xi = \frac{1}{\pi}\int_{-1}^{1} \frac{e^{i\xi x}}{\sqrt{1-\xi^2}}\,d\xi = \mathrm{BesselJ}(0,x).$$

The derivatives of the Bessel functions of the first kind satisfy

$$\begin{cases} \frac{d}{dx}\mathrm{BesselJ}(0,x) = -\mathrm{BesselJ}(1,x), \\ \frac{d}{dx}\mathrm{BesselJ}(1,x) = \frac{\mathrm{BesselJ}(1,x)}{x} - \mathrm{BesselJ}(0,x). \end{cases}$$

By using these formulas, one can easily prove by induction that

$$\frac{d^n}{dx^n}\mathrm{BesselJ}(0,x) = P_n(1/x)\mathrm{BesselJ}(0,x) - Q_n(1/x)\mathrm{BesselJ}(1,x),$$

for some polynomials P_n and Q_n. For $n = 0$ we have $P_0(t) = 0$ and $Q_n(t) = 0$. A short calculation gives the following recursion relation for the polynomials

$$\begin{cases} P_{n+1}(t) = -t^2 P_n'(t) + Q_n(t), \\ Q_{n+1}(t) = -P_n(t) + tQ_n(t) - t^2 Q_n'(t), \end{cases}$$

It is known that optimization with respect to this weight approximates well minimization with respect to the L^∞-norm [2].

3.3 Singular weights

For certain applications it may be interesting to use singular weights of the form

$$\hat{\mu}_m = \frac{\hat{\nu}(\xi)}{|\xi|^{2m}}$$

where $\hat{\nu}$ is as before and m a positive integer. For example, if we want to reduce the relative error of the symbol $(i\xi)^n$ for the n-th derivative then we should minimize with respect to a weight of the above form with $m = n$. In this section we limit ourselves to these differential operators so the object function for the minimization becomes

$$\left\| (i\xi)^n - \sum_{j\in\mathbf{J}} \tilde{a}_j(x) e^{i\cdot(x_j-x)} \right\|^2_{\mu_m,\xi_{\text{opt}}} = \frac{1}{2\pi}\int_{-\infty}^{\infty} \left| (i\xi_{\text{opt}}\xi)^n - \sum_{j\in\mathbf{J}} \tilde{a}_j(x) e^{i\xi_{\text{opt}}\xi(x_j-x)} \right|^2 \frac{\hat{\nu}(\xi)}{|\xi|^{2m}}\, d\xi.$$

We shall assume that $m \le n$ and that $\xi^n \in L^2_{\mu_m,\xi_{\text{opt}}}(\mathbb{R})$. This includes the case $n = m = 1$ which we are most interested in and was also considered by Efraimsson in [1]. This corresponds to minimizing the relative error of the phase velocity for a hyperbolic equation, see Section 6.2. Efraimsson constructed second order accurate central difference approximations, with five and seven points, optimized with respect to the weight $1/|\xi|^2$ (in our terminology in Section 7 this corresponds to $\text{sinc}_{\text{sing}}$ (1,1) and $\text{sinc}_{\text{sing}}$ (1,2)).

We shall now explain how the coefficients for the optimal solution can be found. Clearly, in order for the integral above to be convergent we must have

$$\sum_{j\in\mathbf{J}} \tilde{a}_j(x) e^{i\xi_{\text{opt}}\xi(x_j-x)} = \mathcal{O}(\xi^m).$$

As we mentioned earlier, see equation (7), this is equivalent to that the polynomial fitting constraint is satisfied for all polynomials of degree at most $m-1$. However, we cannot write the components of the matrix $\mathbf{M}$ and the vector $\mathbf{b}$ as before since the corresponding integrals are divergent. To overcome this difficulty we use the following trick. Let $P_m(x) = \sum_{k=0}^{m-1} x^k/k!$ be the Taylor polynomial for the exponential function of degree $m-1$. It then follows from our assumption $m \le n$ and the polynomial fitting constraint that

$$\sum_{j\in\mathbf{J}} \tilde{a}_j(x) P_m(i\xi(x_j - x)) = 0$$

just because the n-th derivative of P_m is zero and the approximation gives correct result for polynomials of degree $m-1$. Since

$$e^{i\xi(x_j-x)} - P_m(i\xi(x_j - x)) = \mathcal{O}(\xi^m)$$

we can now expand the object function in a similar way as before but with slightly different mass matrix $\mathbf{M}'$ and vector $\mathbf{b}'$

$$\begin{aligned}\left\| (i\xi)^n - \sum_{j\in\mathbf{J}} \tilde{a}_j(x) e^{i\cdot(x_j-x)} \right\|^2_{\mu_m,\xi_{\text{opt}}} &= \left\| (i\xi)^n - \sum_{j\in\mathbf{J}} \tilde{a}_j(x) \left(e^{i\cdot(x_j-x)} - P_m(i\xi(x_j - x)) \right) \right\|^2_{\mu_m,\xi_{\text{opt}}} \\ &= \|(i\xi)^n\|^2_{\mu_m,\xi_{\text{opt}}} - 2\,\text{Re}\left(\mathbf{b}'\mathbf{a}^*\right)(x) + \mathbf{a}\mathbf{M}'\mathbf{a}^*(x).\end{aligned}$$

Here $\mathbf{M}'$ is the matrix with components

$$M'_{kl} = \langle e^{i\cdot(x_k-x)} - P_m(i\cdot(x_k-x)), e^{i\cdot(x_l-x)} - P_m(i\cdot(x_l-x))\rangle_{\mu_m,\xi_{\text{opt}}}$$

and $\mathbf{b}'$ is the vector with components

$$b'_l = \langle (i\cdot)^n, e^{i\cdot(x_l-x)} - P_m(i\cdot(x_l-x))\rangle_{\mu_m,\xi_{\text{opt}}}.$$

The coefficient vector $\mathbf{a}$ can now be found as the solution to the linear equation system

$$(\mathbf{a}\ \lambda)\widehat{\mathbf{M}} = (\mathbf{b}'\ \mathbf{q})$$

where $\widehat{\mathbf{M}}$ is the matrix

$$\widehat{\mathbf{M}} = \begin{pmatrix} \mathbf{M}' & \mathbf{P} \\ \mathbf{P}^* & 0 \end{pmatrix}.$$

We close this section by giving the formulas for the special case $m = n = 1$. For M'_{kl} and b'_k we have $M'_{kl} = \mathcal{M}(\xi_{\text{opt}}(x_k-x), \xi_{\text{opt}}(x_l-x))$ and $b'_k = \mathcal{B}(\xi_{\text{opt}}(x_k-x))$ where

$$\mathcal{M}(\alpha,\beta) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\left(e^{i\xi\alpha}-1\right)\left(e^{-i\xi\beta}-1\right)\frac{\hat{\nu}(\xi)}{|\xi|^2}\,d\xi = \int_0^\alpha\int_0^\beta \nu(s-t)\,dtds \tag{15}$$

and

$$\mathcal{B}(\beta) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\left(e^{-i\xi\beta}-1\right)\frac{\hat{\nu}(\xi)}{-i\xi}\,d\xi = \int_0^\beta \nu(-t)\,dt.$$

To derive the formula for $\mathcal{B}$ we only have to differentiate to obtain

$$\mathcal{B}'(\beta) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-i\xi\beta}\hat{\nu}(\xi)\,d\xi = \nu(-\beta)$$

and use that $\mathcal{B}(0) = 0$. The expression for $\mathcal{M}$ can be derived in a similar way.

3.4 Extension to higher dimensions

The above results can easily be extended to higher dimensions. It turns out that almost all formulas extend with only obvious modifications. We only need a weight function $\hat{\nu}$. In most cases it may be natural to choose a spherical symmetric function. For a spherical symmetric weight the components of the matrix $\mathbf{M}$ become

$$M_{kl} = \nu(\xi_{\text{opt}}|\mathbf{x}_k - \mathbf{x}_l|)$$

and the components of the vector $\mathbf{b}(\mathbf{x})$ for the differential operator $\frac{\partial^\alpha}{\partial\mathbf{x}^\alpha}$ become

$$b_l(\mathbf{x}) = \xi_{\text{opt}}^{N|\alpha|}\frac{\partial^\alpha\nu}{\partial\mathbf{x}^\alpha}(|\mathbf{x}-\mathbf{x}_l|),$$

where $\alpha = (\alpha_1,\ldots,\alpha_N)$ is any multi-index and N the dimension.

4 Time stepping methods

The standard method to solve ordinary differential equations numerically is to use higher order Taylor or Runge–Kutta methods. Implicit methods such as Euler backward and θ-methods are also used. The numerical scheme for these methods are constructed to give correct result for polynomials up to a certain degree. Two references for numerical methods for ordinary differential equations are [4, Chapter 7] and [3]. In this section, we discuss general properties of time stepping methods applied to linear evolution equations with time-independent operators. This is a preparation for the following section where we develop a method to derive time integration schemes for such equations.

4.1 Standard numerical methods

A discrete approximation of a linear partial differential equation leads often to a system of coupled ordinary differential equations of the type

$$\begin{cases} u_t = Au, & t > 0, \\ u|_{t=0} = u_0, \end{cases} \tag{16}$$

where $u(t) = (u_1(t), \dots, u_K(t))^T$ is a vector and A a $L \times L$ matrix. If the equation is obtained from a finite difference approximation then the components of u are approximate values of the solution at the grid points. Since the number of variables L might be very large we normally have to apply a numerical method to solve this equation.

Let $\Delta t > 0$ be the time step and put $t_k = k\Delta t$, $k = 0, 1, 2, \dots$. We seek approximations w^k of the vector $u(t)$ at the mesh points t_k: $w^k \approx u(t_k)$, $k = 0, 1, 2, \dots$. Let us give some examples of numerical methods.

Example 4.1 (Forward Euler and higher order Taylor methods). *In the forward Euler method one calculates w^k by using the truncated first order Taylor approximation for u: $u(t_k) \approx u(t_{k-1}) + hu_t(t_{k-1})$. Since $u_t(t_{k-1}) = Au(t_{k-1})$, we have the vector w^k*

$$w^k = w^{k-1} + hAw^{k-1} = (\mathrm{I} + hA)w^{k-1},$$

where I is the identity operator and $w^0 = u_0$. By iterating the this formula k times, it follows that

$$w^k = (\mathrm{I} + Ah)^k u_0.$$

This can be compared with the exact solution to (16) which is given by the exponential matrix $u(t) = \exp(At)u_0$. The exponential matrix can be defined through the series expansion

$$\exp(At) = \sum_{n=0}^{\infty} \frac{A^n t^n}{n!}.$$

If we adjust the time step such that $h = t/k$, we see that the forward Euler method approximates $\exp(At)$ by

$$\exp(At) \approx \left(\mathrm{I} + \frac{At}{k}\right)^k.$$

Even though the right hand side converges to $\exp(At)$ *as* $k \to \infty$ *the convergence is slow. To resolve this problem one can use higher order Taylor methods where the exponential function is approximated by more terms in its Taylor series*

$$\exp(Ah) \approx \sum_{n=0}^{N} \frac{h^n A^n}{n!}.$$

It turns out that the standard second and fourth order Runge–Kutta methods coincide with the second and fourth order Taylor methods for linear and time independent operators.

4.2 Approximation of the exponential matrix

The forward Euler method and higher order Taylor methods are examples of explicit methods, the approximation w^k can be calculated from w^{k-1} without solving a linear equation system. Explicit methods correspond to a polynomial approximation of the exponential matrix

$$\exp(A) \approx P(A) = \sum_{k=0}^{N_P} a_k A^k,$$

and

$$w^k = P(Ah)w^{k-1} = P(Ah)^k u_0.$$

In contrast to explicit methods we have implicit methods where we have to solve a linear equation system to find w^k. This corresponds to a rational approximation of the exponential matrix

$$\exp(A) \approx P_R(A)Q_R(A)^{-1}$$

where both $P_R(A)$ and $Q_R(A)$ are polynomials in A. An approximation of this type is called a Padé approximation for certain choices of P_R and Q_R.

Now, assume the matrix A has a spectral decomposition

$$A = UDU^{-1},$$

where U is the matrix of eigenvectors and $D = \mathrm{diag}(\lambda_1, \ldots, \lambda_N)$ is the diagonal matrix of eigenvalues λ_j. In terms of this decomposition, we have

$$\exp(At) = U\mathrm{diag}(e^{\lambda_1 t}, \ldots, e^{\lambda_N t})U^{-1}, \tag{17}$$

and

$$P(At) = U\mathrm{diag}(P(\lambda_1 t), \ldots, P(\lambda_N t))U^{-1}. \tag{18}$$

Moreover, if

$$R(x) = \frac{P_R(x)}{Q_R(x)},$$

then

$$P_R(At)Q_R(At)^{-1} = U\mathrm{diag}(R(\lambda_1 t), \ldots, R(\lambda_N t))U^{-1},$$

thus, $R(A)$ is well defined. This can of course be seen as a special case of the functional calculus for operators. However, what is important for us is that it shows

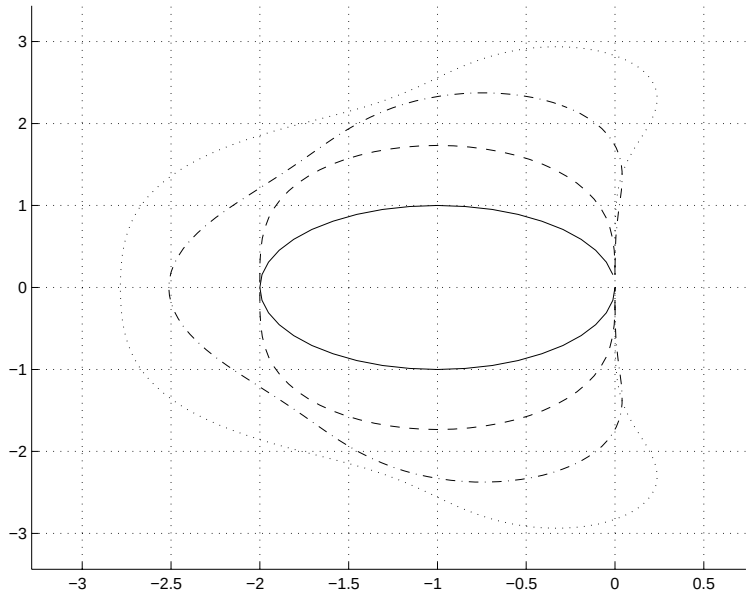


Figure 2. Stability regions for different explicit methods: Forward Euler, ———, Taylor order 3,– – ––,Taylor order 4, – · – · –, Taylor order 5, · · · · · · ·.

that in order to find a good time integration method one needs a good polynomial or rational approximation of the exponential function. For example, higher order Taylor methods correspond to approximation of the exponential function with its Taylor polynomials. In the next section we propose an optimization procedure to find the polynomial approximation. In this approach we can also take the location of the eigenvalues into account.

4.3 Stability of time integration methods

We say that a numerical method is stable if and only if w^k stays bounded as $k \to \infty$ for all initial conditions u_0. For an explicit method with polynomial P and a diagonalizable matrix A this holds if and only if $|P(\lambda_k h)| \leq 1$ for all eigenvalues. Here h is the time step. Similarly, we have $|R(\lambda_k h)| \leq 1$ for implicit methods. We define the stability region for an explicit method as the set

$$\Omega_P = \{z \in \mathbb{C} : |P(z)| < 1\}. \tag{19}$$

Thus, the method is stable provided that $\lambda_k h \in \overline{\Omega}_P$, $k = 1, 2, \ldots, L$. Figure 2 shows the stability regions for the forward Euler method and the Taylor methods of order three, four and five. If the matrix A is not diagonalizable then the situation is slightly more complicated. In order to study stability for such matrices, it is convenient to write the matrix A in Jordan normal form

$$A = UDU^{-1},$$

where U is an invertible matrix and D a Jordan block matrix. Each block in D has the form

$$D_k = \begin{pmatrix} \lambda_k & 1 & 0 & \dots & 0 \\ 0 & \lambda_k & 1 & \dots & 0 \\ 0 & 0 & \lambda_k & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 1 \\ 0 & 0 & 0 & \dots & \lambda_k \end{pmatrix}.$$

One can show that the time stepping method applied to A is stable if and only if all eigenvalues lie inside Ω_P and the ones corresponding to Jordan blocks of size larger than one lie strictly inside. For the exact solution we instead have that the solution remains bounded for all initial conditions if and only if $\mathrm{Re}\,\lambda_k \leq 0$ for all eigenvalues corresponding to Jordan blocks of size one and $\mathrm{Re}\,\lambda_k < 0$ for all other eigenvalues.

5 Optimized time integration methods

The purpose of this section is to develop a general method to derive time integration schemes for linear and time independent operators. These schemes include the standard higher order Taylor methods. The approach is very similar to the frequency optimization in Section 3, that is, we will use a constrained optimization problem to find the scheme. We recall from the last section that to find a good explicit time integration method for a linear and time independent operator is equivalent to find a good polynomial approximation of the exponential function.

5.1 The optimization problem

Let $\mathcal{P}_K$ denote the space of polynomials of degree at most K, where K is a positive integer, and let $\mathcal{H}$ be a Hilbert space of functions on $\Omega \subset \mathbb{C}$, with norm $\|\cdot\|_{\mathcal{H}}$, which includes the exponential function and the space $\mathcal{P}_K$. We will choose our polynomial $p \in \mathcal{P}_K$ as the best approximation of the exponential function in $\mathcal{H}$ subject to the constraint that p also approximates the exponential function to some order L at the origin. Let us formalize this approach. Every $p \in \mathcal{P}_K$ has a representation

$$p(t) = \sum_{k=0}^{K} a_k t^k$$

and we let $\mathbf{a} = (a_0, \ldots, a_K)$ denote the vector of coefficients for p. Since the exponential function has the power series expansion

$$e^t = \sum_{n=0}^{\infty} \frac{t^n}{n!},$$

the polynomial approximation p of e^t is of order L if and only if

$$a_n = \frac{1}{n!}, \qquad n = 0, \ldots, L-1.$$

In matrix notation we have

$$\mathbf{aP} = \mathbf{q} \tag{20}$$

where $\mathbf{q} = (q_0, \ldots, q_{L-1})$ with $q_l = 1/l!$ and $\mathbf{P} = (P_{kl})$ is the $K \times L$ matrix with $P_{ll} = 1$ for $l = 0, 1, \ldots, L-1$ and all other components are zero. The optimization problem for the coefficients is

$$\inf_{\tilde{\mathbf{a}}\mathbf{P}=\mathbf{q}} \left\| \exp(\cdot) - \sum_{k=0}^{K} \tilde{a}_k p_k \right\|_{\mathcal{H}}. \tag{21}$$

where p_k is the k-th monomial as before: $p_k(t) = t^k$. The following result shows how the coefficient vector for the optimal solution can be found.

THEOREM 5.1. *Suppose that the Hilbert space $\mathcal{H}$, the matrix $\mathbf{P}$ and the vector $\mathbf{q}$ are as above, then there exists a unique extremal vector $\mathbf{a}$ to (21). The vector can be found as the solution to the linear equation system*

$$(\mathbf{a}\ \lambda) \begin{pmatrix} \mathbf{M} & \mathbf{P} \\ \mathbf{P}^* & 0 \end{pmatrix} = (\mathbf{b}\ \mathbf{q})$$

where the matrix $\mathbf{M} = (M_{kl})_{k,l=0}^{K}$ *and the row vector* $\mathbf{b} = (b_l)_{l=0}^{K}$ *have the components*

$$M_{kl} = \langle p_k, p_l \rangle_{\mathcal{H}}, \qquad k, l = 0, 1, \ldots, K$$

and

$$b_l(x) = \langle \exp(\cdot), p_l \rangle_{\mathcal{H}}, \qquad l = 0, 1, \ldots, K,$$

respectively. Moreover, $\lambda = (\lambda_0, \ldots, \lambda_L)$ *is the vector of Lagrangian multipliers for the problem and*

$$\left\| \exp(\cdot) - \sum_{k=0}^{K} a_k p_k \right\|_{\mathcal{H}}^2 = \|\exp)(\cdot)\|_{\mathcal{H}}^2 - 2\,\mathrm{Re}\ (\mathbf{b}\mathbf{a}^*) + \mathbf{a}\mathbf{M}\mathbf{a}^*,$$

where $\mathbf{a}^*$ *is the Hermitian conjugate of* $\mathbf{a}$*.*

Proof. The theorem follows immediately from the results in Appendix A on constrained minimization in complex Hilbert spaces, Lemma A.1. □

For minimization without the linear constraint the coefficient vector is obtained as the solution to

$$\mathbf{a}\mathbf{M} = \mathbf{b}.$$

We shall now specialize to matrices A which only have negative eigenvalues (corresponding to parabolic equations) and matrices which only have purely imaginary eigenvalues (corresponding to hyperbolic equations). More precisely, we optimize our approximation of the exponential function on the negative half axis and the imaginary axis, respectively. It follows from the theorem above that we only have to find the coefficients for the matrix $\mathbf{M}$ and vector $\mathbf{b}$ to calculate the polynomials.

5.2 Purely imaginary eigenvalues

For hyperbolic problems, such as wave propagation, we often have operators A with purely imaginary eigenvalues or eigenvalues clustering at the imaginary axis. For such operators it is natural to choose a norm of the same type as for frequency optimization of the finite difference approximation of derivatives from the previous section but on the imaginary axis. Let the weight ν and the parameter k_{opt} be as in Section 3. We define the Hilbert space $L^2_{\nu,k_{\mathrm{opt}}}(i\mathbb{R})$ as the set of functions g on the imaginary axis such that

$$\|g\|^2_{\nu,k_{\mathrm{opt}}} = \frac{1}{2\pi k_{\mathrm{opt}}} \int_{-\infty}^{\infty} |g(i\xi)|^2 \hat{\nu}(\xi/k_{\mathrm{opt}})\, d\xi < \infty.$$

The corresponding inner product is

$$\langle g_1, g_2 \rangle_{\nu,k_{\mathrm{opt}}} = \frac{1}{2\pi k_{\mathrm{opt}}} \int_{-\infty}^{\infty} g_1(i\xi)\overline{g_2(i\xi)}\hat{\nu}(\xi/k_{\mathrm{opt}})\, d\xi.$$

A linear change of variables yields

$$\langle l_1, l_2 \rangle_{\nu,k_{\mathrm{opt}}} = \frac{1}{2\pi} \int_{-\infty}^{\infty} l_1(ik_{\mathrm{opt}}\xi)\overline{l_2(ik_{\mathrm{opt}}\xi)}\hat{\nu}(\xi)\, d\xi.$$

The formula for the components of the vector $\mathbf{b}$ and the matrix $\mathbf{M}$ becomes

$$M_{kl} = \langle p_k, p_l \rangle_{\nu,k_{\text{opt}}} = (-1)^l k_{\text{opt}}^{k+l} \frac{1}{2\pi} \int_{-\infty}^{\infty} (i\xi)^{k+l} \hat{\nu}(\xi)\, d\xi \\ = (-1)^l k_{\text{opt}}^{k+l} \nu^{(k+l)}(0)$$

and

$$b_l = \langle \exp(\cdot), p_l \rangle_{\nu,k_{\text{opt}}} = (-k_{\text{opt}})^l \frac{1}{2\pi} \int_{-\infty}^{\infty} (i\xi)^l \exp(ik_{\text{opt}}\xi) \hat{\nu}(\xi)\, d\xi \\ = (-k_{\text{opt}})^l \nu^{(l)}(k_{\text{opt}}).$$

At a first glance it may look as $\mathbf{M}$ is not symmetric. However, since ν is an even function it follows that all odd derivatives of ν vanishes at the origin. In the same fashion as for finite difference approximations we can use singular weights of the form

$$\hat{\mu}_m = \frac{\hat{\nu}(\xi)}{|\xi|^{2m}}.$$

5.3 Negative eigenvalues

Large negative eigenvalues are often associated with parabolic problems such as the heat equation. For such equations it may be convenient to optimize the approximation on the negative real axis. Let μ be a positive function on the positive half axis $\mathbb{R}_+ = \{x \in \mathbb{R} : x > 0\}$ and let $h > 0$. We assume that the weight is such that $p\mu \in L^1(\mathbb{R}_+)$ for all polynomials p. Define $L^2_{\mu,h_{\text{opt}}}(\mathbb{R}_-)$ as the Hilbert space of functions f on the negative half axis such that

$$\|f\|^2_{\mu,h_{\text{opt}}} = \frac{1}{h_{\text{opt}}} \int_0^{\infty} |f(-t)|^2 \mu(t/h_{\text{opt}})\, dt < \infty.$$

The corresponding inner product is

$$\langle f, g \rangle_{\mu,h_{\text{opt}}} = \frac{1}{h_{\text{opt}}} \int_0^{\infty} f(-t)\overline{g(-t)}\mu(t/h_{\text{opt}})\, dt$$

A linear change of variables yields

$$\langle f, g \rangle_{\mu,h_{\text{opt}}} = \int_0^{\infty} f(-h_{\text{opt}}t)\overline{g(-h_{\text{opt}}t)}\mu(t)\, dt.$$

Let $\tilde{\mu}$ denote the Laplace transform of μ. We can now calculate the coefficients for the row vector $\mathbf{b}$ and the mass matrix $\mathbf{M}$. We have

$$b_l = \langle \exp(\cdot), p_l \rangle_{\mu,h_{\text{opt}}} = h_{\text{opt}}^l \int_0^{\infty} e^{-ht}(-t)^l \mu(t)\, dt = h_{\text{opt}}^l \tilde{\mu}^{(l)}(h_{\text{opt}}), \quad (22)$$

and

$$M_{kl} = \langle p_k, p_l \rangle_{\mu,h_{\text{opt}}} = h_{\text{opt}}^{k+l} \int_0^{\infty} (-t)^{k+l} \mu(t)\, dt = h_{\text{opt}}^{k+l} \tilde{\mu}^{(k+l)}(0). \quad (23)$$

The parameter h_{opt} and the weight μ should be chosen in accordance with the size of the eigenvalues of the matrix A and the time step. If A is a finite difference

approximation of some differential operator, then typically the small eigenvalues for the differential operator agree well with some of the small eigenvalues for A whereas the large eigenvalues for A do not correspond to any of the eigenvalues for the differential operator. Therefore, we are more interested to model the small eigenvalues accurately in our time integration method and the constant $h_{\rm opt}$ should be chosen relatively this set of eigenvalues. However, the large eigenvalues are often responsible for instabilities in the time integration. We give one example of a weight of this type.

Example 5.2 (The box weight). *Define*

$$\mu(t) = \begin{cases} 1, & 0 \leq t \leq 1, \\ 0, & t > 1. \end{cases}$$

Laplace transformation of the weight gives

$$\widetilde{\mu}(s) = \frac{1 - e^{-s}}{s}.$$

A calculation of the derivatives gives

$$\widetilde{\mu}^{(n)}(s) = (-1)^n \left(\frac{1}{s^{n+1}} + e^{-s} \sum_{k=0}^{n} \frac{n!}{k!} \frac{1}{s^{n-k+1}} \right).$$

6 Wave propagation

Recall that all electromagnetic waves in free space propagate with the same velocity, the speed of light, independent of the shape of the wave. This is not true for electromagnetic waves in certain materials, for example glass and water, where waves of different frequencies propagate with different velocities. We say that the material is *dispersive*. The rainbow is a well known effect due to this phenomena. In numerical solutions of hyperbolic equations it may happen that the velocity of a wave depends on the frequency although this is not true for the governing equation. For problems in several dimensions the speed may also be different in different directions. This phenomena is called *numerical dispersion* and is in most cases unavoidable but it is important to reduce its effects.

The purpose of this section is to explain dispersion and numerical dispersion and other related quantities for a simple wave equation in one space dimension. We also define a precise concept for how many points per wave length we need for different schemes in order for the relative error in the phase velocity to be less than a given tolerance. A similar measure is defined for time integration schemes. These concepts will then be used in Section 7 in order to evaluate the different schemes which we have derived and to compare them with standard methods. All the concept which we consider can also be extended to systems in higher dimensions. We show how this is performed on the two dimensional $\mathrm{TM_z}$ mode of the Maxwell equations on an infinite Cartesian grid.

6.1 Dispersion relations

Let us consider the simple advection equation in one space dimension

$$\begin{cases} u_t(x,t) + cu_x(x,t) = 0, & x \in \mathbb{R}, \quad t > 0, \\ u(x,0) = u_0(x), & x \in \mathbb{R}. \end{cases} \tag{24}$$

The solution to this equation is given by

$$u(x,t) = u_0(x - ct),$$

which is a wave traveling to the right with velocity c. A *dispersion relation* is an equation which relates the wave number ξ for a monochromatic wave

$$u(x,t) = e^{i(\xi x - \omega t)}.$$

with the angular velocity ω. If we substitute the monochromatic wave into (24), we obtain

$$i(c\xi - \omega)e^{i(\xi x - \omega t)} = 0.$$

After a simplification we get

$$\omega = c\xi$$

which is the dispersion relation for the advection equation. The *phase velocity* is the velocity a monochromatic wave with wave number ξ propagates with and is given by

$$v_p = \frac{\omega(\xi)}{\xi}.$$

For the advection equation it equals c, independent of the wave number.

6.2 Numerical dispersion

We are now in position to study numerical dispersion and phase velocity. We will consider both exact and numerical time integration. For numerical time integration, where we combine spatial and time discretization, it is possible, more or less, to study the dispersion error due to the spatial and time approximation separately.

6.2.1 Exact time integration

Now let us consider a semi-discrete approximation of the advection equation on a grid with mesh size Δx

$$u_l'(t) + cDu_l(t) = 0, \qquad l \in \mathbb{Z}, \tag{25}$$

where $u_l(t) \approx u(x_l, t)$, $x_l = l\Delta x$, and Du_l a discrete approximation of the derivative with respect to x. Here we consider central differences approximations of the derivative of the form

$$Du_l(t) = \frac{1}{\Delta x} \sum_{j=-N}^{N} a_j u_{l+j}(t).$$

Exactly as for the continuous advection equation we seek solutions to (25) of the form $u(x,t) = e^{i(\xi x - \omega t)}$. We substitute this into the equation to obtain

$$-i\omega e^{i(\xi x - \omega t)} + \frac{c}{\Delta x} \left(\sum_{j=-N}^{N} a_j e^{i\xi j \Delta x} \right) e^{i(\xi x - \omega t)} = 0.$$

Clearly, this ansatz diagonalizes the equation just as in the continuous case. If we use the definition of the effective wave number,

$$i\tilde{\xi}(\xi) = \frac{1}{\Delta x} \left(\sum_{j=-N}^{N} a_j e^{i\xi j \Delta x} \right),$$

we obtain after simplification the numerical dispersion relation

$$\omega(\xi) = c\tilde{\xi}(\xi).$$

Compared to the continuous dispersion we see that the wave number is replaced by the effective wave number. The numerical phase velocity is now given by

$$v_p = \frac{\omega(\xi)}{\xi} = c\frac{\tilde{\xi}(\xi)}{\xi}.$$

A reasonable requirement on a numerical scheme is that the relative error of the phase velocity is less than a certain quantity for all waves with wave number less than a given maximal wave number $\xi_{\max}$. For exact time integration the relative error of the phase velocity becomes

$$F(\xi) = \left| \frac{c - v_p(\xi)}{c} \right| = \left| 1 - \frac{\tilde{\xi}(\xi)}{\xi} \right|. \tag{26}$$

In our comparisons between different finite different approximations we will use the following definition of $\xi_{\max}$ and points per wave length (PPW).

DEFINITION 6.1. *Given a $\kappa > 0$, we define $\xi_{\max}$ as the largest number such that*

$$F(\xi) \leq \kappa, \qquad \textit{for all} \qquad \xi \in [-\xi_{\max}, \xi_{\max}].$$

The number of points per wave length (PPW) for the specified error is given by

$$PPW = \frac{2\pi}{\xi_{\max}\Delta x}.$$

To see the relevance of this concept we consider a monochromatic wave which we propagate a number of wave lengths, say 10. During the propagation the total phase error of the numerical wave accumulates. If the relative error in the phase is 2% then the total phase error after a propagation of ten wave lengths is

$$\phi = \frac{2\pi \cdot 10 \cdot 2}{100} = \frac{2\pi}{5} \approx 1.257 \text{ radians.}$$

After a propagation of 20 wave lengths the phase error is 2.514 radians. Thus, the larger the computational region is relative to the wave length the more accurate the scheme has to be.

6.2.2 Numerical time integration

We will here use the same notation as in Section 4 for the time stepping. This means that the matrix operator A is given by $-cD$ so the time integration has the following form

$$w^k = P(-chD)w^{k-1}, \qquad k = 1, 2, 3, \dots,$$

where D is the discrete approximation of the derivative. Here P is either a polynomial for an explicit method or a rational function for an implicit method, h is the time step and the vectors w^k, $k = 0, 1, 2, \dots$, are approximations of the solution to (24) at the points (x_l, t_k). Again we seek solutions of exponential form $w_l^k = e^{i(\xi x_l - \omega t_k)}$. If we substitute this into the equation we obtain

$$e^{i(\xi x_l - \omega t_k)} = P(-ich\tilde{\xi}(\xi))e^{i(\xi x_l - \omega t_{k-1})}.$$

A simplification yields

$$e^{-ih\omega} = P(-ich\tilde{\xi}(\xi)),$$

which is the numerical dispersion relation in this case. To express ω as a function of ξ we need to take the logarithm of the right hand side

$$\omega(\xi) = \frac{i\log(P(-ihc\tilde{\xi}(\xi))}{h}$$

Unfortunately, there is no guarantee that the the angular velocity is real valued. In fact, for polynomials one can show that $\omega(\xi)$ has a non-zero imaginary part except at a finite set of points. The imaginary part corresponds to numerical damping of the wave if $\operatorname{Im}\omega(\xi) < 0$ (dissipation) and numerical amplification if $\operatorname{Im}\omega(\xi) > 0$.

The phase velocity is then given by

$$\tilde{v}_p(\xi) = \operatorname{Re}\left(\frac{i\log(P(-ich\tilde{\xi}(\xi))}{h\xi}\right).$$

We use the tilde sign to distinguish between the phase velocities for exact and numerical time integration. To evaluate and compare different time stepping methods for hyperbolic problems we want to separate the error in phase velocity due to the spatial approximation and the time stepping. Although this is not completely possible it can be done to the first order of the errors. To do this we introduce the function

$$Q(x) = \frac{i \log(P(-ix))}{x}. \tag{27}$$

The phase velocity $\tilde{v}_p(\xi)$ can now be written as a product between Q and v_p

$$\tilde{v}_p(\xi) = c \,\mathrm{Re}\, \left(Q(hc\tilde{\xi}(\xi)) \right) \frac{\tilde{\xi}(\xi)}{\xi} = \mathrm{Re}\, \left(Q(hc\tilde{\xi}(\xi)) \right) v_p(\xi). \tag{28}$$

In analogy with the definition of $\xi_{\max}$, we define the parameter $k_{\max}$ to control the relative error due to time integration. We also define k_{stab} to specify the stability interval for imaginary eigenvalues for the method.

DEFINITION 6.2. *Given a $\kappa > 0$, we define $k_{\max}$ as the largest number such that*

$$|1 - Q(x)| \leq \kappa, \qquad \textit{for all} \quad x \in [-k_{\max}, k_{\max}],$$

The parameter k_{stab} is defined as largest number such that

$$|P(ix)| \leq 1, \qquad \textit{for all} \quad x \in [-k_{\mathrm{stab}}, k_{\mathrm{stab}}].$$

Let us motivate the definition of $k_{\max}$. It follows from (28) that the relative error of $\tilde{v}_p(\xi)$ can be estimated to first order by the sum of the relative error of $Q(hc\tilde{\xi}(\xi))$ and the relative error of $\tilde{v}_p(\xi)$. If $hc\tilde{\xi}(\xi) \in [-k_{\max}, k_{\max}]$ then we have for the relative error for $Q(hc\tilde{\xi}(\xi))$

$$\left| 1 - Q(hc\tilde{\xi}(\xi)) \right| \leq \max_{x \in [-k_{\max}, k_{\max}]} |1 - Q(x)| \leq \kappa.$$

The parameter k_{stab} is such that $i[-k_{\mathrm{stab}}, k_{\mathrm{stab}}]$ is the intersection of the stability region Ω_p for P, equation (19), with the imaginary axis.

6.3 Wave propagation in higher dimension and systems

The above argument can of course be generalized to higher dimension and systems as well. The requirement is that we have a constant coefficient differential equation and a Cartesian and equidistant grid. We will here show how this can be accomplished for the two dimensional $\mathrm{TM_z}$ mode of the Maxwell equations on an infinite Cartesian grid. The main result is that the relative error in the phase velocity in any direction is bounded by the maximum of relative error in the phase velocities along the coordinate axes (see equation (32).

The treatment here is similar to [5, Section 4.2]. The Maxwell equations in $\mathrm{TM_z}$ mode are

$$\left\{ \begin{array}{rcl} \frac{\partial H_x}{\partial t} & = & -\frac{1}{\mu} \frac{\partial E_z}{\partial y}, \\ \frac{\partial H_y}{\partial t} & = & \frac{1}{\mu} \frac{\partial E_z}{\partial x}, \\ \frac{\partial E_z}{\partial t} & = & \frac{1}{\varepsilon} \left(\frac{\partial H_y}{\partial x} - \frac{\partial H_x}{\partial y} \right), \end{array} \right. \tag{29}$$

For exact time integration we seek numerical solutions which approximate the fields at the points (x_k, y_l) where $x_k = k\Delta x$ and $y_l = l\Delta y$ with $k, l \in \mathbb{Z}$. For example, for the electric field component E_z we have

$$E_z|_{k,l}(t) \approx E_z(x_k, y_l, t).$$

The approximations of the derivatives are along the coordinate axes. For example, for the derivative of E_z with respect to x we have

$$\left.\frac{\partial E_z}{\partial x}\right|_{k,l}(t) \approx \frac{1}{\Delta x}\sum_{j=-N_x}^{N_x} a_j^x\, E_z|_{k+j,l}(t)$$

We define the effective wave number along the x and y direction by

$$\widetilde{\xi}_x(\xi_x) = -i\frac{1}{\Delta x}\sum_{j=-N_x}^{N_x} a_j^x e^{ix_j\xi_x}$$

and

$$\widetilde{\xi}_y(\xi_y) = -i\frac{1}{\Delta y}\sum_{j=-N_y}^{N_y} a_j^y e^{iy_j\xi_y}$$

As before we seek solutions to the discrete $\mathrm{TM_z}$ mode equation of exponential form

$$\left\{\begin{array}{lcl} H_x|_{k,l}(t) &=& h_x e^{i(\xi_x x_k+\xi_y y_l-\omega t)},\\ H_y|_{k,l}(t) &=& h_y e^{i(\xi_x x_k+\xi_y y_l-\omega t)},\\ E_z|_{k,l}(t) &=& e_z e^{i(\xi_x x_k+\xi_y y_l-\omega t)}, \end{array}\right. \tag{30}$$

which is a plane wave traveling along the direction (ξ_x, ξ_y). Substituting these expressions into the discrete TM_z mode equations yields after simplification

$$\left\{\begin{array}{lcl} -i\omega h_x &=& -i\frac{1}{\mu}\widetilde{\xi}_y(\xi_y)e_z,\\ -i\omega h_y &=& i\frac{1}{\mu}\widetilde{\xi}_x(\xi_x)e_z,\\ -i\omega e_z &=& i\frac{1}{\varepsilon}\left(\widetilde{\xi}_x(\xi_x)h_y - \widetilde{\xi}_y(\xi_y)h_x\right). \end{array}\right. \tag{31}$$

If we combine these equations, we obtain the following dispersion relation

$$\omega^2(\xi_x, \xi_y) = c^2\left(\widetilde{\xi}_x(\xi_x)^2 + \widetilde{\xi}_y(\xi_y)^2\right),$$

where $c = \frac{1}{\sqrt{\varepsilon\mu}}$ is the speed of light. The phase velocity for the wave is

$$\nu(\xi_x, \xi_y) = \frac{\omega(\xi_x, \xi_y)}{\sqrt{\xi_x^2 + \xi_y^2}} = c\frac{\sqrt{\widetilde{\xi}_x(\xi_x)^2 + \widetilde{\xi}_y(\xi_y)^2}}{\sqrt{\xi_x^2 + \xi_y^2}}.$$

The phase velocities along the x and y directions are defined as in the one dimensional case

$$\nu_x(\xi_x) = c\frac{\widetilde{\xi}_x(\xi_x)}{\xi_x}, \qquad \nu_y(\xi_y) = c\frac{\widetilde{\xi}_y(\xi_y)}{\xi_y}.$$

It follows that

$$\nu(\xi_x, \xi_y) = \frac{\sqrt{\xi_x^2 \nu_x(\xi_x)^2 + \xi_y^2 \nu_y(\xi_y)^2}}{\sqrt{\xi_x^2 + \xi_y^2}}.$$

The phase velocity $\nu(\xi_x, \xi_y)$ is thus some kind of weighted mean value of the phase velocities along the coordinate directions. This give us the following bound

$$\min(\nu_x(\xi_x), \nu_y(\xi_y)) \leq \nu(\xi_x, \xi_y) \leq \max(\nu_x(\xi_x), \nu_y(\xi_y)),$$

and we can estimate the relative error in the phase velocity by

$$F(\xi_x, \xi_y) \leq \max(F_x(\xi_x), F_y(\xi_y)). \tag{32}$$

7 Numerical tests

The purpose of this section is to test the methods we have developed and to compare them with some standard methods. We begin by testing the finite difference approximations.

7.1 Tests of finite difference methods

We limit ourselves to central difference approximations of the first derivative on equidistant grids. We will focus on dispersion error for the advection equation which we measure by the relative error in the phase velocity as defined by (26). In terms of the effective wave number we have

$$F(\xi) = \left| 1 - \frac{\tilde{\xi}(\xi)}{\xi} \right|.$$

The number of points per wave length (PPW) and $\xi_{\max}$ are estimated for the different tolerances $\kappa = 5, 2, 1, 0.5, 0.2, 0.1, 0.05, 0.02$ and 0.01 percent.

Since we use central difference approximations the number of points in the stencil is odd and can therefore be written

$$\text{number of points} = 1 + 2N_{\text{Tay}} + 2N_{\text{opt}}.$$

although $a_0 = 0$ for all central difference approximations of the first derivative. Thus, $2N_{\text{Tay}} + 1$ is the number of the coefficients which are used for the polynomial fitting and the remaining $2N_{\text{opt}}$ are used for the optimization. For the first derivative, the order of the approximation is therefore $2N_{\text{Tay}}$. In all tables for optimized schemes we use the names of the function ν from examples 3.4-3.6. For example, $\text{sinc}\,(1, 2)$ indicates that the weight from Example 3.4 is used, $N_{\text{Tay}} = 1$ and $N_{\text{opt}} = 2$. A subscript sing indicate that we use the corresponding singular weight $\hat{\nu}(\xi)/|\xi|^2$, for example $\text{sinc}_{\,\text{sing}}$, (see Subsection 3.3).

Table 1 shows $\xi_{\max}$ and the number of points per wave length which are needed for different tolerances for the Taylor approximations of order 2, 4, 6 and 8 (this correspond to $N_{\text{Tay}} = 1, 2, 3, 4$ and $N_{\text{opt}} = 0$). The definition of $\xi_{\max}$ was given in Definition 6.1. We see that the number of points per wave length grows very fast for the second order approximation as the tolerance decreases whereas we still only need nine points per wave length for the 8-th order approximation for $\kappa = 0.01\%$. In the tables for the optimized schemes we have chosen ξ_{opt} so that we need as few points per wave length as possible for the given tolerance κ, function ν, N_{Tay} and N_{opt}. Table 2 and 3 shows the result for different seven point schemes. The DRP scheme due to Tam and Webb corresponds to $\text{sinc}\,(2, 1)$ for $\xi_{\text{opt}} = 1.1$. We see that the best result for dispersion errors are obtained for $\text{BesselJ}_{\text{sing}}(0, 3)$. This may not be so surprising since it is known that optimization with respect to the corresponding norm ought to be close to optimization in the supremum norm [2]. This is also confirmed by Figure 3 which displays the relative phase velocities for three different seven point schemes including the sixth order Taylor approximation and $\text{BesselJ}_{\text{sing}}(0, 3)$. The parameter ξ_{opt} is chosen to give the best result for the tolerance $\kappa = 0.5\%$. If we compare the results the sixth order Taylor approximation and $\text{BesselJ}_{\text{sing}}(0, 3)$ we see that the optimized scheme needs about 1.48-1.74 less points per wave length for a given accuracy. In

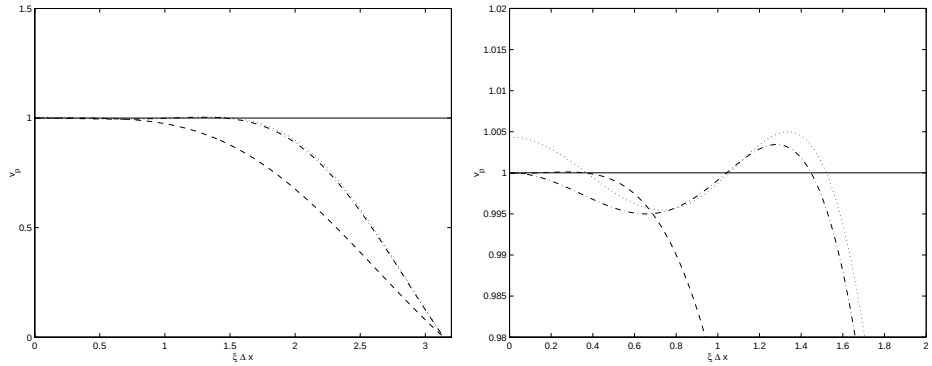


Figure 3. Relative phase velocity for different schemes: Exact, ———, Taylor order 6,– – ––, $\mathrm{sinc}\,(1,2)$, – · – · –,$\mathrm{BesselJ}_{\mathrm{sing}}(0,3)$, · · · · · · · . The right figure is a magnification of the left.

Scheme	Taylor (1,0)		Taylor (2,0)		Taylor (3,0)		Taylor (4,0)	
Tolerence	$\xi_{\max}$	PPW	$\xi_{\max}$	PPW	$\xi_{\max}$	PPW	$\xi_{\max}$	PPW
$\kappa = 5\%$	0.552	11.4	1.15	5.46	1.49	4.23	1.70	3.70
$\kappa = 2\%$	0.347	18.1	0.902	6.97	1.25	5.03	1.48	4.25
$\kappa = 1\%$	0.245	25.6	0.753	8.39	1.10	5.71	1.34	4.70
$\kappa = 0.5\%$	0.173	36.2	0.630	9.98	0.972	6.47	1.21	5.18
$\kappa = 0.2\%$	0.110	57.3	0.499	12.6	0.827	7.60	1.07	5.87
$\kappa = 0.1\%$	0.077	81.1	0.418	15.0	0.733	8.57	0.98	6.44
$\kappa = 0.05\%$	0.055	115	0.351	17.9	0.652	9.65	0.889	7.07
$\kappa = 0.02\%$	0.035	181	0.279	22.5	0.557	11.29	0.788	7.97
$\kappa = 0.01\%$	0.024	257	0.234	26.8	0.495	12.7	0.720	8.72

Table 1. Number of points per wave length for a given tolerance for different schemes.

three dimensions the gain is about $3.2 - 5.3$. Table 4-6 shows the results for some three, five and nine point schemes.

7.2 Tests of time integration methods

In our tests of time integration methods we will compare the standard higher order Taylor methods with optimized explicit methods for different weights and parameters. As for the finite difference approximations, we will focus on the dispersion errors due to the time integration. Thus we will use the function Q defined in equation (27) and Definition 6.2 of $k_{\max}$ and k_{stab}. We assume that all polynomial approximations $P(x) = \sum_{n=0}^{N} a_n x^n$ of the exponential function are of at least order 1, that is, $a_0 = 1$ (otherwise we cannot define Q at the origin). We split the degree N of the polynomial into two parts

$$N = N_{\mathrm{Tay}} + N_{\mathrm{opt}}.$$

The integer $N_{\mathrm{Tay}} + 1$ is the number of coefficients used for polynomial fitting and the remaining N_{opt} are used for the optimization. This means that the order of the approximation is at least $N_{\mathrm{Tay}} + 1$. We shall use the same tolerances κ as before and in analogy with $k_{\max}$ and k_{opt} we choose k_{opt} to give a maximal $k_{\max}$ for fixed weight function ν, and integers N_{Tay} and N_{opt}. Since stability is important

Scheme	sinc (0,3)			sinc (1,2)			sinc (2,1)		
Tolerance	ξ_{opt}	$\xi_{\max}$	PPW	ξ_{opt}	$\xi_{\max}$	PPW	ξ_{opt}	$\xi_{\max}$	PPW
$\kappa = 5\%$	1.942	2.025	3.10	2.208	2.152	2.92	2.097	2.007	3.13
$\kappa = 2\%$	1.697	1.767	3.56	1.923	1.885	3.33	1.794	1.727	3.64
$\kappa = 1\%$	1.529	1.590	3.95	1.730	1.699	3.70	1.595	1.540	4.08
$\kappa = 0.5\%$	1.375	1.429	4.40	1.554	1.528	4.11	1.419	1.372	4.58
$\kappa = 0.2\%$	1.192	1.237	5.08	1.345	1.325	4.74	1.215	1.178	5.33
$\kappa = 0.1\%$	1.068	1.108	5.67	1.204	1.187	5.29	1.082	1.049	5.99
$\kappa = 0.05\%$	0.956	0.992	6.34	1.077	1.062	5.92	0.963	0.935	6.72
$\kappa = 0.02\%$	0.825	0.855	7.35	0.928	0.916	6.86	0.826	0.802	7.83
$\kappa = 0.01\%$	0.737	0.764	8.22	0.829	0.818	7.68	0.735	0.714	8.80

Table 2. Number of points per wave length for a given tolerance for different schemes.

Scheme	$\text{sinc}_{\text{sing}}$ (0,3)			BesselJ (1,2)			$\text{BesselJ}_{\text{sing}}$ (0,3)		
Tolerance	ξ_{opt}	$\xi_{\max}$	PPW	ξ_{opt}	$\xi_{\max}$	PPW	ξ_{opt}	$\xi_{\max}$	PPW
$\kappa = 5\%$	2.339	2.189	2.871	2.080	2.134	2.944	2.220	2.202	2.853
$\kappa = 2\%$	2.030	1.925	3.265	1.821	1.868	3.364	1.950	1.941	3.237
$\kappa = 1\%$	1.825	1.739	3.612	1.641	1.683	3.733	1.763	1.757	3.575
$\kappa = 0.5\%$	1.639	1.568	4.007	1.476	1.514	4.151	1.590	1.587	3.960
$\kappa = 0.2\%$	1.419	1.363	4.611	1.280	1.312	4.790	1.383	1.381	4.549
$\kappa = 0.1\%$	1.272	1.223	5.138	1.147	1.175	5.347	1.242	1.241	5.064
$\kappa = 0.05\%$	1.138	1.096	5.734	1.027	1.052	5.975	1.114	1.113	5.644
$\kappa = 0.02\%$	0.982	0.946	6.639	0.885	0.907	6.929	0.963	0.962	6.529
$\kappa = 0.01\%$	0.877	0.846	7.427	0.791	0.810	7.756	0.861	0.861	7.301

Table 3. Number of points per wave length for a given tolerance for different schemes.

Scheme	sinc (0,1)			sinc (1,1)			sinc (0,2)		
Tolerance	ξ_{opt}	$\xi_{\max}$	PPW	ξ_{opt}	$\xi_{\max}$	PPW	ξ_{opt}	$\xi_{\max}$	PPW
$\kappa = 5\%$	0.699	0.767	8.19	1.740	1.667	3.77	1.510	1.600	3.93
$\kappa = 2\%$	0.445	0.488	12.88	1.393	1.341	4.68	1.221	1.291	4.87
$\kappa = 1\%$	0.316	0.346	18.17	1.176	1.135	5.54	1.036	1.094	5.74
$\kappa = 0.5\%$	0.223	0.245	25.68	0.992	0.958	6.56	0.876	0.925	6.79
$\kappa = 0.2\%$	0.141	0.155	40.57	0.790	0.765	8.22	0.701	0.740	8.49
$\kappa = 0.1\%$	0.100	0.110	57.37	0.665	0.644	9.76	0.591	0.624	10.07
$\kappa = 0.05\%$	0.071	0.077	81.12	0.560	0.542	11.59	0.498	0.526	11.95
$\kappa = 0.02\%$	0.045	0.049	128.26	0.446	0.432	14.55	0.397	0.419	15.01
$\kappa = 0.01\%$	0.032	0.035	181.38	0.375	0.363	17.30	0.334	0.352	17.83

Table 4. Number of points per wave length for a given tolerance for different schemes.

Scheme	$\mathrm{sinc}_{\mathrm{sing}}$ (0,2)			$\mathrm{Gauss}_{\mathrm{sing}}$ (0,2)			$\mathrm{BesselJ}_{\mathrm{sing}}$ (0,2)		
Tolerance	ξ_{opt}	ξ_{max}	PPW	ξ_{opt}	ξ_{max}	PPW	ξ_{opt}	ξ_{max}	PPW
$\kappa = 5\%$	1.933	1.741	3.608	2.711	1.724	3.644	1.768	1.748	3.595
$\kappa = 2\%$	1.545	1.414	4.444	1.959	1.393	4.511	1.429	1.421	4.423
$\kappa = 1\%$	1.305	1.202	5.228	1.590	1.180	5.323	1.213	1.208	5.199
$\kappa = 0.5\%$	1.101	1.018	6.169	1.308	0.998	6.295	1.028	1.025	6.131
$\kappa = 0.2\%$	0.879	0.816	7.704	1.022	0.798	7.878	0.823	0.821	7.650
$\kappa = 0.1\%$	0.741	0.689	9.125	0.852	0.672	9.347	0.694	0.693	9.064
$\kappa = 0.05\%$	0.623	0.580	10.827	0.712	0.566	11.092	0.585	0.585	10.748
$\kappa = 0.02\%$	0.497	0.463	13.578	0.564	0.451	13.923	0.466	0.466	13.482
$\kappa = 0.01\%$	0.418	0.390	16.128	0.473	0.380	16.543	0.393	0.393	16.002

Table 5. Number of points per wave length for a given tolerance for different schemes.

Scheme	sinc (1,3)			$\mathrm{sinc}_{\mathrm{sing}}$ (0,4)			$\mathrm{BesselJ}_{\mathrm{sing}}$ (0,4)		
Tolerance	ξ_{opt}	ξ_{max}	PPW	ξ_{opt}	ξ_{max}	PPW	ξ_{opt}	ξ_{max}	PPW
$\kappa = 5\%$	2.441	2.397	2.621	2.546	2.427	2.589	2.457	2.442	2.573
$\kappa = 2\%$	2.211	2.182	2.880	2.300	2.215	2.837	2.242	2.234	2.812
$\kappa = 1\%$	2.050	2.027	3.099	2.131	2.061	3.048	2.088	2.083	3.016
$\kappa = 0.5\%$	1.899	1.880	3.342	1.974	1.915	3.282	1.942	1.938	3.241
$\kappa = 0.2\%$	1.713	1.698	3.701	1.780	1.732	3.628	1.759	1.757	3.576
$\kappa = 0.1\%$	1.582	1.569	4.004	1.645	1.603	3.920	1.630	1.628	3.859
$\kappa = 0.05\%$	1.460	1.449	4.336	1.518	1.481	4.242	1.508	1.507	4.170
$\kappa = 0.02\%$	1.311	1.302	4.826	1.364	1.332	4.716	1.358	1.357	4.629
$\kappa = 0.01\%$	1.207	1.199	5.239	1.257	1.228	5.115	1.253	1.253	5.016

Table 6. Number of points per wave length for a given tolerance for different schemes.

Scheme	Taylor (1,0)		Taylor (2,0)		Taylor (3,0)		Taylor (4,0)	
Tolerance	$k_{\max}$	k_{stab}	$k_{\max}$	k_{stab}	$k_{\max}$	k_{stab}	$k_{\max}$	k_{stab}
$\kappa = 5\%$	0.108	0.000	0.483	0.000	1.345	1.732	1.468	2.828
$\kappa = 2\%$	0.041	0.000	0.316	0.000	1.189	1.732	1.147	2.828
$\kappa = 1\%$	0.020	0.000	0.228	0.000	1.106	1.732	0.961	2.828
$\kappa = 0.5\%$	0.010	0.000	0.164	0.000	1.049	1.732	0.809	2.828
$\kappa = 0.2\%$	0.004	0.000	0.106	0.000	0.430	1.732	0.647	2.828
$\kappa = 0.1\%$	0.002	0.000	0.075	0.000	0.323	1.732	0.548	2.828
$\kappa = 0.05\%$	0.001	0.000	0.054	0.000	0.248	1.732	0.463	2.828
$\kappa = 0.02\%$	0.000	0.000	0.034	0.000	0.178	1.732	0.372	2.828
$\kappa = 0.01\%$	0.000	0.000	0.024	0.000	0.140	1.732	0.315	2.828

Table 7. Maximal imaginary eigenvalues for different tolerances and stability parameters.

Scheme	sinc (1,2)			sinc (0,3)			Gauss (1,2)		
Tolerance	k_{opt}	$k_{\max}$	k_{stab}	k_{opt}	$k_{\max}$	k_{stab}	k_{opt}	$k_{\max}$	k_{stab}
$\kappa = 5\%$	2.498	2.222	0.000	2.262	2.046	0.000	3.274	2.365	0.000
$\kappa = 2\%$	1.730	1.590	0.000	1.828	1.638	0.000	2.005	1.474	0.000
$\kappa = 1\%$	1.257	1.271	0.000	1.525	1.377	0.000	1.460	1.185	0.000
$\kappa = 0.5\%$	0.938	1.107	0.000	1.276	1.174	0.000	1.052	1.054	0.000
$\kappa = 0.2\%$	0.655	0.797	0.000	0.582	0.991	0.000	0.684	0.994	0.000
$\kappa = 0.1\%$	0.506	0.528	0.000	0.530	0.557	0.000	0.596	0.541	0.000
$\kappa = 0.05\%$	0.395	0.400	0.000	0.406	0.412	0.000	0.465	0.405	0.000
$\kappa = 0.02\%$	0.286	0.285	0.000	0.291	0.290	0.000	0.338	0.287	0.000
$\kappa = 0.01\%$	0.225	0.223	0.000	0.228	0.225	0.000	0.266	0.224	0.000

Table 8. Maximal imaginary eigenvalues for different tolerances and stability parameters.

for time integration, the variable k_{stab} is calculated for all schemes. The notation of different schemes is the same as before.

In Table 7 the constants $k_{\max}$ and k_{stab} are given for Taylor schemes of order 2 to 5 for tolerances between 0.01 and 5 percent. Table 8 and 9 shows the result for some approximations of degree 3 and Table 10 and 11 approximations of degree 4. Apart from the stability constraint it seems as the best result again is obtained for $\text{BesselJ}_{\text{sing}}$. However, the interval of stability $[-k_{\text{stab}}, k_{\text{stab}}]$ does not include $[-k_{\max}, k_{\max}]$ in most cases so the gain in accuracy cannot be used. It may be necessary to include a stability constraint into the optimization problem to cope with this problem. The reason why this is such a big problem is that the imaginary axis is the boundary for the stability region for exact time integration. Therefore the optimization technique have a chance to work better for problems with dissipation since the eigenvalues are then strictly included in the left half plane but we will not investigate this issue here.

Scheme	$\mathrm{sinc}_{\mathrm{sing}}$ (0,3)			BesselJ (0,3)			$\mathrm{BesselJ}_{\mathrm{sing}}$ (0,3)		
Tolerance	k_{opt}	k_{max}	k_{stab}	k_{opt}	k_{max}	k_{stab}	k_{opt}	k_{max}	k_{stab}
$\kappa = 5\%$	2.700	2.183	0.000	2.076	2.040	0.000	2.349	2.149	0.000
$\kappa = 2\%$	2.282	1.795	0.000	1.706	1.658	0.000	1.977	1.773	0.000
$\kappa = 1\%$	1.718	1.380	0.000	1.424	1.397	0.000	1.511	1.392	0.000
$\kappa = 0.5\%$	1.099	1.081	0.000	1.191	1.193	0.000	0.984	1.098	0.000
$\kappa = 0.2\%$	0.643	0.995	0.000	0.550	0.994	0.000	0.597	1.001	0.000
$\kappa = 0.1\%$	0.563	0.546	0.000	0.491	0.553	0.000	0.504	0.541	0.000
$\kappa = 0.05\%$	0.436	0.408	0.000	0.376	0.411	0.000	0.390	0.406	0.000
$\kappa = 0.02\%$	0.314	0.288	0.000	0.269	0.290	0.000	0.281	0.288	0.000
$\kappa = 0.01\%$	0.246	0.224	0.000	0.211	0.225	0.000	0.221	0.224	0.000

Table 9. Maximal imaginary eigenvalues for different tolerances and stability parameters.

Scheme	$\mathrm{sinc}_{\mathrm{sing}}$ (0,4)			BesselJ (0,4)			$\mathrm{BesselJ}_{\mathrm{sing}}$ (0,4)		
Tolerance	k_{opt}	k_{max}	k_{stab}	k_{opt}	k_{max}	k_{stab}	k_{opt}	k_{max}	k_{stab}
$\kappa = 5\%$	3.039	2.925	0.869	2.121	2.784	1.217	2.746	2.915	0.915
$\kappa = 2\%$	2.192	2.216	0.615	1.674	1.904	0.955	2.075	2.224	0.680
$\kappa = 1\%$	1.774	1.692	0.494	1.404	1.553	0.798	1.673	1.705	0.544
$\kappa = 0.5\%$	1.468	1.383	0.407	1.178	1.289	0.669	1.382	1.393	0.448
$\kappa = 0.2\%$	1.161	1.085	0.321	0.936	1.018	0.528	1.091	1.093	0.353
$\kappa = 0.1\%$	0.977	0.911	0.270	0.787	0.854	0.446	0.919	0.918	0.296
$\kappa = 0.05\%$	0.825	0.768	0.224	0.662	0.719	0.369	0.775	0.773	0.253
$\kappa = 0.02\%$	0.660	0.614	0.179	0.526	0.573	0.295	0.621	0.619	0.201
$\kappa = 0.01\%$	0.559	0.519	0.161	0.443	0.482	0.261	0.525	0.523	0.178

Table 10. Maximal imaginary eigenvalues for different tolerances and stability parameters.

Scheme	$\mathrm{sinc}_{\mathrm{sing}}$ (1,3)			BesselJ (1,3)			$\mathrm{BesselJ}_{\mathrm{sing}}$ (1,3)		
Tolerance	k_{opt}	k_{max}	k_{stab}	k_{opt}	k_{max}	k_{stab}	k_{opt}	k_{max}	k_{stab}
$\kappa = 5\%$	2.734	2.744	0.000	2.532	2.782	0.000	2.528	2.765	0.000
$\kappa = 2\%$	1.998	1.960	0.000	1.815	1.958	0.000	1.849	1.943	0.000
$\kappa = 1\%$	1.643	1.556	0.000	1.486	1.554	0.000	1.520	1.551	0.000
$\kappa = 0.5\%$	1.369	1.281	0.000	1.236	1.279	0.000	1.267	1.279	0.000
$\kappa = 0.2\%$	1.087	1.010	0.000	0.980	1.008	0.000	1.006	1.009	0.000
$\kappa = 0.1\%$	0.916	0.849	0.000	0.826	0.848	0.000	0.848	0.849	0.000
$\kappa = 0.05\%$	0.774	0.716	0.000	0.698	0.715	0.000	0.716	0.716	0.000
$\kappa = 0.02\%$	0.620	0.573	0.000	0.559	0.572	0.000	0.574	0.573	0.000
$\kappa = 0.01\%$	0.525	0.485	0.000	0.473	0.484	0.000	0.486	0.485	0.000

Table 11. Maximal imaginary eigenvalues for different tolerances and stability parameters.

8 Conclusions and future work

We have developed methods to derive both finite difference approximations of derivatives and numerical schemes for time integration for differential equations with time independent operators. We have compared them with the standard approximations obtained by polynomial fitting. These methods can be extended straightforwardly to higher dimensions on both structured and unstructured grids. Unlike polynomial fitting in several variables where the Vandermonde matrix may become singular for certain distributions of grid points, the optimization problem will always have a solution unless the constraint is overdetermined.

The numerical tests for finite difference approximation shows that we can reduce the number of points per wave length by a factor 1.5-1.7 by using optimized schemes compared to standard schemes and still have the same accuracy for the phase velocity for numerical solutions of hyperbolic problems. Here we have compared central difference approximation on equidistant grid with equally large stencils.

The results for the time integration schemes which are optimized for imaginary eigenvalues are not that promising. Although the accuracy seems to be quite good, stability problems appear. In many cases there is no interval of stability on the imaginary axis. This could perhaps be expected since the imaginary axis is the boundary of the left half plane and all eigenvalues have to belong to the left half plane in order for exact solution to remain bounded for all initial condition as $t \rightarrow \infty$. To resolve this problem one should perhaps incorporate a stability constraint into the optimization problem. This may be a problem for further investigations.

Other problems that might be interesting to study further is the generalizations to higher dimensions, handling of boundary conditions, and derive optimized summation by parts operators. Summation by parts operators are used to force the numerical scheme to be stable. The optimization procedure is quite general and is therefore applicable to many different problems.

References

[1] G. Efraimsson, *A numerical method for the first-order wave equation with discontinuous initial data*, Numer. Methods Partial Differential Equations **14** (1998), no. 3, 353–365.

[2] K. Forsberg, Private communication, 2001, Swedish Defence Research Agency, FOI.

[3] E. Hairer, S. P. Nørsett, and G. Wanner, *Solving ordinary differential equations. I*, second ed., Springer-Verlag, Berlin, 1993, Nonstiff problems.

[4] J. Stoer and R. Bulirsch, *Introduction to numerical analysis*, second ed., Springer-Verlag, New York, 1993, Translated from the German by R. Bartels, W. Gautschi and C. Witzgall.

[5] A. Taflove and S. C. Hagness, *Computational electrodynamics: the finite-difference time-domain method*, second ed., Artech House Inc., Boston, MA, 2000, With 1 CD-ROM (Windows).

[6] C. K. W. Tam, *Computational aeroacoustics: Issues and methods*, AIAA J. **33** (1995), no. 10, 1788–1796.

[7] C. K. W. Tam and K. A. Kurbatskii, *A wavenumber based extrapolation and interpolation method for use in conjunction with high-order finite difference schemes*, J. Comput. Phys. **157** (2000), no. 2, 588–617.

[8] C. K. W. Tam and H. Shen, *Direct computation of nonlinear acoustic pulses using high-order finite difference schemes*, AIAA paper (1993), no. 93-4325.

[9] C. K. W. Tam and J. C. Webb, *Dispersion-relation-preserving finite difference schemes for computational acoustics*, J. Comput. Phys. **107** (1993), no. 2, 262–281.

Appendix A

Minimization in complex Hilbert spaces

The purpose of this section is to derive formulas for the extremal function for a minimization problem in a complex Hilbert space subject to a linear constraint. Let E be a finite dimensional subspace of a Hilbert space $\mathcal{H}$ with basis $\{e_k\}_{k=1}^K$, that is, every vector $a \in E$ has an unique representation

$$a = \sum_{k=1}^{K} a_k e_k. \tag{33}$$

with $a_k \in \mathbb{C}$. We denote the coordinate vector for a by $\mathbf{a} = (a_1, \dots, a_K)$. The linear constraint is given by

$$p_l(a) = q_l, \qquad l = 1, \dots, L, \tag{34}$$

where $\{p_l\}_{l=1}^L$ are linear functionals on $\mathcal{H}$ and $\mathbf{q} = (q_1, \dots, q_L) \in \mathbb{C}^L$ is a complex vector. We assume that the restriction of the linear functionals $\{p_l\}_{l=1}^L$ to E are linear independent which implies that the dimension of the affine subspace

$$\{\tilde{a} \in E : p_l(a) = q_l \text{ for all } l = 1, \dots, L\}.$$

is $K - L$, in particular, $L \geq K$. If we expand $p_l(a)$ in its coordinates we see that (34) is equivalent to the matrix equation

$$\mathbf{q} = \mathbf{a}\mathbf{P}$$

for the coordinate vector $\mathbf{a}$. Here $\mathbf{P} = (P_{kl})_{k=1,l=1}^{k=K,l=L}$ is the $K \times L$ matrix with $P_{kl} = p_l(e_k)$ and $\mathbf{q} = (q_1, \dots, q_L)$. For an arbitrary $x \in \mathcal{H}$ we consider the following minimization problem

$$\inf_{\mathbf{a}\tilde{\mathbf{P}}} \left\| x - \sum_{k=1}^{K} \tilde{a}_k e_k \right\|_{\mathcal{H}}. \tag{35}$$

The following theorem shows how the coordinate vector for the optimal solution can be found. First we need some notation. The mass matrix $\mathbf{M} = (M_{kl})_{k,l=1}^N$ for the basis $\{e_j\}_{j=1}^N$ is

$$M_{k,l} = \langle e_k, e_l \rangle_{\mathcal{H}}, \qquad k, l = 1, \dots, K.$$

With this notation the norm of a vector $a \in E$ is

$$\|a\|_{\mathcal{H}}^2 = \mathbf{a}\mathbf{M}\mathbf{a}^*$$

where $\mathbf{a}^* = (\bar{a}_1, \dots, \bar{a}_K)^T$ is the Hermitian conjugate of the coordinate vector $\mathbf{a}$.

LEMMA A.1. *Assume that* $\mathbf{M}$, $\mathbf{P}$ *and* $\mathbf{q}$ *are as above. Then for each* $x \in \mathcal{H}$ *there exists an unique optimal solution* $a \in E$ *to the minimization problem (35). The coordinate vector* $\mathbf{a} = (a_1, \dots, a_K)$ *for* a *is given as the solution to the linear problem*

$$(\mathbf{a}\ \lambda)\widehat{\mathbf{M}} = (\mathbf{b}\ \mathbf{q})$$

where $\widehat{\mathbf{M}}$ *is the matrix*

$$\widehat{\mathbf{M}} = \begin{pmatrix} \mathbf{M} & \mathbf{P} \\ \mathbf{P}^* & 0 \end{pmatrix},$$

and $\mathbf{b} = (b_1, \ldots, b_K)$ *is row vector with components*

$$b_k = \langle x, e_k \rangle_{\mathcal{H}}, \qquad k = 1, \ldots, K.$$

Here $\lambda = (\lambda_1, \ldots, \lambda_L)$ *is the vector with Lagrangian multipliers for the problem.*

Proof. The existence and uniqueness of an optimal solution follows since we optmize a strictly convex norm over an affine finite dimensional subspace. Recall the following result for constrained optimization due to Lagrange. A necessary condition for a point $\mathbf{x}$ to be a minimum for the problem

$$\inf\{f(\mathbf{x}) : g_l(\mathbf{x}) = 0 \text{ for all } l = 1, \ldots, L\}$$

where $\mathbf{x} \in \mathbb{R}^K$ and $L \leq K$, is that the Lagrange function

$$L(\mathbf{x}, \Lambda) = f(\mathbf{x}) + \sum_{l=1}^{L} \lambda_l g_l(\mathbf{x})$$

where $\Lambda = (\lambda_1, \ldots, \lambda_L)$, satisfies

$$\frac{\partial L}{\partial x_k}(\mathbf{x}) = 0, \qquad k = 1, \ldots, K$$

and

$$g_l(\mathbf{x}) = 0, \qquad l = 1, \ldots, L.$$

In the rest of the proof we restrict ourselves to real Hilbert spaces. For complex hilbert spaces we only have to split up all complex variable into real and imaginary parts and treat them as two separate real variables. With the definition of the vectors $\mathbf{a}$, $\mathbf{b}$ and the matrix $\mathbf{M}$, we have for the object function $f(\mathbf{a}) = \|x - a\|_{\mathcal{H}}^2$

$$\begin{aligned} f(\mathbf{a}) &= \|x\|_{\mathcal{H}}^2 - 2\langle a, x \rangle_{\mathcal{H}} + \|x\|_{\mathcal{H}}^2 \\ &= \|x\|_{\mathcal{H}}^2 - 2\sum_{k=1}^{K} a_k b_k + \sum_{k,l=1}^{K} a_k M_{kl} a_l, \end{aligned}$$

where $\mathbf{b}$ is defined as above. The constraints are

$$g_l(\mathbf{a}) = 2\left(\sum_{k=1}^{K} a_k p_l(e_k) - q_l\right), \qquad l = 1, \ldots, L.$$

The factor 2 is a matter of convenience. The Lagrange function becomes

$$L(\mathbf{a}, \Lambda) = \|x\|_{\mathcal{H}}^2 - 2\sum_{k=1}^{K} a_k b_k + \sum_{k,l=1}^{K} a_k M_{kl} a_l + 2\sum_{l=1}^{L} \lambda_l \left(\sum_{k=1}^{K} a_k p_l(e_k) - q_l\right).$$

Differentiation with respect to a_i yields

$$0 = \frac{\partial L}{\partial x_i}(\mathbf{a}, \Lambda) = -2b_i + 2\sum_{k=1}^{K} a_k M_{ki} - 2\sum_{l=1}^{L} \lambda_l p_l(e_i).$$

We have here used that $\mathbf{M}$ is symmetric: $M_{ki} = \langle e_k, e_i \rangle_{\mathcal{H}} = \langle e_i, e_k \rangle_{\mathcal{H}} = M_{ik}$. Lagrange's criteria in matrix form becomes

$$\mathbf{aM} + \Lambda \mathbf{P}^T = \mathbf{b}$$

and the constraints are

$$\mathbf{aP} = \mathbf{q}.$$

This is the equation system given in the lemma in the real case. It is easy to see that it has a unique solution under the assumptions that $\{e_k\}_{k=1}^K$ is a basis for E and the constraints are linear independent. □

Issuing organisation
FOI – Swedish Defence Research Agency
Division of Aeronautics, FFA
SE-172 90 STOCKHOLM

Report number, ISRN
FOI-R–0407–SE

Report type
Scientific report

Month year
March 19, 2002

Project number
E 840303 CEM

Customers code
3. Aeronautical Research

Research area code
6. Electric Warfare

Sub area code
62. Stealth Technology

Author(s)
Stefan Jakobsson

Project manager
Jan Nordström

Approved by
Bengt Winzell
Head, Computational Aerodynamics Department

Scientifically and technically responsible
Jan Nordström

Report title
Frequency optimized computation methods

Abstract
In this paper we develop an alternative method to derive finite difference approximations of derivatives. The purpose is to find schemes which work for a broader range of frequencies than the usual approximations based on polynomial fitting and Taylor's Theorem to the expense of less accuracy for low frequencies. The numerical schemes are obtained as solutions to constrained optimizations problems in a weighted L^2-norm in the frequency domain. We examine the accuracy of these schemes and compare them with the standard approximations. We also use the same approach to derive numerical schemes for time integration for differential equations with time independent operators. To test the accuracy of the different schemes, we study dispersion errors for a simple wave equation in one space dimension. We examine the number of points per wave length which is needed in order for the relative error in the phase velocity to be below a certain bound. A similar examination is carried out for the different time integration schemes.

Keywords
Finite differences, numerical dispersion

Further bibliographic information

ISSN
ISSN 1650-1942

Pages
49

Language
English

Price
Acc. to price list

Security classification
Unclassified

Utgivare	Rapportnummer, ISRN	Klassificering
Totalförsvarets Forskningsinstitut – FOI Avdelningen för Flygteknik, FFA SE-172 90 STOCKHOLM	FOI-R–0407–SE	Vetenskaplig rapport
	Månad år March 19, 2002	Projektnummer E 840303 CEM
	Verksamhetsgren 3. Flygteknisk forskning	
	Forskningsområde 6. Telekrig	
	Delområde 62. Signaturanpassning	
Författare Stefan Jakobsson	Projektledare Jan Nordström	
	Godkänd av Bengt Winzell Chef, Institutionen för beräkningsaerodynamik	
	Tekniskt och/eller vetenskapligt ansvarig Jan Nordström	

Rapporttitel

Frekvensoptimerade beräkningsmetoder

Sammanfattning

I denna artikel utvecklar vi en alternativ metod för att härleda finita differens approximationer av derivator. Syftet är att hitta approximationer som fungerar för ett större intervall av frekvenser än de vanliga standard scheman till priset att vi får sämre noggrannhet för låga frekvenser. Koefficienterna för approximationerna ges som lösningar till minimering problem under bivillkor i viktade L^2-rum i frekvens domän. Vi använder samma idé för att härleda tidstegningsscheman för linjära differential ekvationer för tidsoberoende operatorer.

För att studera och jämföra noggrannheten för de olika scheman studerar vi dispersionsfel för en enkel vågekvation i en rumsdimension. Vi undersöker hur många punkter per våglängd som behövs för att det relativa felet i fashastigheten skall vara mindre än vissa givna toleransnivåer. En motsvarande undersökning gör för tidsscheman.

Nyckelord

Finita differenser, numerisk dispersion

Övriga bibliografiska uppgifter

ISSN	Antal sidor	Språk
ISSN 1650-1942	49	Engelska
Distribution enligt missiv Distribution	Pris Enligt prislista	
	Sekretess Öppen	