# Fourth order schemes for hyperbolic equations : heterogeneous case

Jukka Tuomela

INRIA

UNITÉ DE RECHERCHE
INRIA-ROCQUENCOURT

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau
Rocquencourt
B.P.105
78153 Le Chesnay Cedex
France
Tél.(1)39635511

Rapports de Recherche

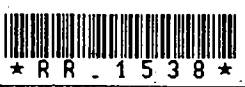
N° 1538

*Programme 6*
*Calcul Scientifique, Modélisation et*
*Logiciel numérique par Ordinateur*

# FOURTH ORDER SCHEMES FOR HYPERBOLIC EQUATIONS : HETEROGENEOUS CASE

**Jukka TUOMELA**

**Octobre 1991**



# Fourth Order Schemes for Hyperbolic Equations: Heterogeneous Case

# Schémas d'Ordre Quatre pour Équations Hyperboliques: Milieu Hétérogène

Jukka Tuomela

INRIA, Rocquencourt
BP 105
78153 Le Chesnay Cedex
France

Helsinki University of Technology
02150 Espoo
Finland

## Abstract

With a simple finite difference operator we construct fourth order schemes in space and in time for wave equation, Maxwell equations and linearized elastodynamic equations using the modified equation approach. The schemes remain stable for arbitrary heterogeneous mediums because the relevant difference operators are always positive definite. We also present some dispersion curves to show the accuracy of the schemes.



## Résumé

A l'aide d'un opérateur aux différences finies simple nous construirons quelques schémas d'ordre quatre en espace et en temps pour l'équation des ondes, les équations de Maxwell et les équations d'élastodynamique linéaire en utilisant la technique d'équation modifiée. Ces schémas sont stables pour les milieux hétérogènes quelconques car certains opérateurs aux différences finies utilisés sont toujours définis positifs. Nous présenterons aussi quelques courbes de dispersion pour montrer la précision de ces schémas.

# 1 Introduction

In this report we introduce and analyse methods for solving wave equation, Maxwell equation and linearized elastodynamic equation. All methods are based on a relatively simple difference operator which simplifies the analysis; in fact it is essentially sufficient to treat thoroughly the one dimensional wave equation, the other cases then follow the same idea although of course the various computations become more complicated. To construct our fourth order schemes we use the modified equation approach, see [CJ], [CO], [SB] and [TU]. Another way to construct accurate schemes is given in [HO]. There are two disadvantages in Holberg's method: the difference operators are very long and the time discretization remains of the second order. The problem with the modified equation approach has been the difficulty to construct stable schemes in the heterogeneous case. In our method this problem disappears because we obtain naturally a discrete variational formulation with positive definite operators which then gives the stability immediately. The difference operators in our method are also rather long, but only implicitly: the same simple operator is used several times; this simplifies implementation.

Another interesting feature appears when we treat the linearized elastodynamic equation. In the continuous case there are shear and pressure waves which travel with different speeds. The shear waves are transversal and the pressure waves are longitudinal: their oscillations are then orthogonal to each other; our method preserves this orthogonality. In addition, and this is probably more important, the approximation is 'uniform' with respect to Lamé coefficients. This means that accuracy is essentially only determined by the number of points per wavelength instead of depending also on the ratio of Lamé coefficients.

We shall present various dispersion curves to show the accuracy of the scheme. As regards imposing the boundary conditions we evidently encounter the same problems as with the difference methods in general: if the boundary is not rectangular some special treatment is necessary, and of course the stability of the resulting initial boundary value problem is not an easy matter. However, the standard homogeneous Dirichlet and Neumann boundary conditions are obviously stable because they are 'built in' the discrete variational formulation.

# 2 One Dimensional Wave Equation

## 2.1 Constant Coefficients

Let us start by considering the one dimensional wave equation

$$u_{tt} - u_{xx} = 0$$

For the moment we forget the time derivatives and treat only the space derivatives. All the developments will be based on the following difference operator.

$$\mathcal{S}u = \frac{1}{12}(u(x+3h/2) + 9u(x+h/2) - 9u(x-h/2) - u(x-3h/2)) \qquad (2.1)$$

To approximate the second derivative we naturally then use

$$\mathcal{T}u = -\mathcal{S}^2 u =$$
$$\frac{1}{144}(164u(x) - 63u(x \pm h) - 18u(x \pm 2h) - u(x \pm 3h))$$

Note that the operator $\mathcal{T}$ does not use the 'intermediate' values $u(x \pm ih/2)$. Using the Taylor's expansion we get

$$\begin{aligned} \frac{1}{h}\mathcal{S}u &= u_x + \frac{h^2}{8}u_{xxx} + O(h^4) \\ \frac{1}{h^2}\mathcal{T}u &= -u_{xx} - \frac{h^2}{4}u_{xxxx} + O(h^4) \end{aligned}$$

So we obtain a fourth order scheme

$$u_{tt} + \frac{1}{h^2}(\mathcal{T} + \frac{1}{4}\mathcal{T}^2)u = 0$$

Then using the ordinary three point time discretization and compensating the second order error term we have the fully discrete scheme.

$$u^{n+1} - 2u^n + u^{n-1} + \alpha^2(\mathcal{T} + \frac{3-\alpha^2}{12}\mathcal{T}^2)u^n = 0 \qquad (2.2)$$

where $\alpha = \delta t/h$ and $\delta t$ is the time step. For more details about these kind of 'error compensating' or 'modified equation' schemes we refer to [CJ], [CO], [SB] and [TU]. Let us define

$$\mathcal{A} = \mathcal{T} + \frac{3-\alpha^2}{12}\mathcal{T}^2$$

Then we have

**Proposition 1** *The operator $\mathcal{A}$ is positive and the scheme (2.2) is stable if*

$$\alpha \leq \sqrt{\frac{39 - 3\sqrt{61}}{8}} \simeq 1.39 \qquad (2.3)$$

**Proof** First we recall that the stability condition of the scheme

$$u^{n+1} - 2u^n + u^{n-1} + \alpha^2 \mathcal{A}u^n = 0$$

can be written as

$$\alpha^2 \|\mathcal{A}\| \leq 4 \tag{2.4}$$

To calculate the norm of $\mathcal{A}$ we can use the Fourier analysis and we start by considering first the operator $\mathcal{T}$ for which we get

$$\mathcal{T} e^{ikx} = \frac{1}{36}(\sin(3kh/2) + 9\sin(kh/2))^2 e^{ikx}$$

This shows that $\mathcal{T}$ is a positive operator and consequently also $\mathcal{A}$ because the coefficient in front of $\mathcal{T}^2$ is positive for the (small) values of $\alpha$ which interest us. Simple calculations then show that the maximum is attained at $kh = \pi$ so

$$\|\mathcal{T}\| = (-1+9)^2/36 = 16/9 \tag{2.5}$$

Then evidently

$$\|\mathcal{A}\| = \|\mathcal{T}\| \left(1 + \frac{3-\alpha^2}{12} \|\mathcal{T}\|\right)$$

Putting all this into (2.4) gives then the stability limit.

■

## 2.2 Variable Coefficients

Then we turn to the variable coefficient case and take the following model equation

$$\rho(x) u_{tt} - \partial_x(\mu(x) u_x) = 0 \tag{2.6}$$

Here $c(x) = \sqrt{\mu(x)/\rho(x)}$ is the (local) wave speed and $z(x) = \sqrt{\mu(x)\rho(x)}$ the wave impedance. We suppose that $\rho$, $\mu \in L^\infty$ and that

$$0 < \rho_* \leq \rho(x) \leq \rho^*$$
$$0 < \mu_* \leq \mu(x) \leq \mu^*$$

From this we get also the inequalities

$$0 < c_* = \sqrt{\mu_*/\rho^*} \leq c(x) \leq \sqrt{\mu^*/\rho_*} = c^*$$
$$0 < z_* = \sqrt{\mu_*\rho_*} \leq z(x) \leq \sqrt{\mu^*\rho^*} = z^*$$

Note that these limits of $c$ and $z$ are not necessarily attained at any point. Now in the variable coefficient case we cannot use Fourier analysis, so we have to use energy methods. To proceed let us introduce some convenient notations. It is helpful to interpret the difference operators as operating on the sequences instead of functions so let us consider the space of square summable sequences $l^2(\mathbb{Z}) = \{\{x_i\} \mid \sum x_i^2 <$

$\infty\}$ with the inner product $(x,y) = \sum x_i y_i$. To simplify the notation the spaces are supposed to be real, but of course we could as well consider complex sequences. Then we will need the shift operator $\mathcal{E}$ defined by $\mathcal{E}u = \{u_{i+1}\}$ which has the following evident (but fundamental) property

$$(\mathcal{E}u, v) = (u, \mathcal{E}^{-1}v) \tag{2.7}$$

In other words $\mathcal{E}^{-1}$ is the transpose of $\mathcal{E}$. With this operator we can write $\mathcal{T}$ as

$$\mathcal{T} = -\frac{1}{144}(\mathcal{E}^{3/2} + 9\mathcal{E}^{1/2} - 9\mathcal{E}^{-1/2} - \mathcal{E}^{-3/2})^2$$

One might think that there is a problem treating the fractional powers, but there is a way out: we can factor $\mathcal{T}$ as

$$\mathcal{T} = -\frac{1}{144}(\mathcal{E}^2 + 9\mathcal{E} - 9 - \mathcal{E}^{-1})(\mathcal{E} + 9 - 9\mathcal{E}^{-1} - \mathcal{E}^{-2})$$

The right hand side can also be considered as a rational function with the argument $\mathcal{E}$ in which case it is called the Z-transform of $\mathcal{T}$. The natural step to take is then to define

$$\mathcal{S}^+ = \frac{1}{12}(\mathcal{E}^2 + 9\mathcal{E} - 9 - \mathcal{E}^{-1})$$
$$\mathcal{S}^- = \frac{1}{12}(\mathcal{E} + 9 - 9\mathcal{E}^{-1} - \mathcal{E}^{-2})$$

Using (2.7) we have then immediately

$$-(\mathcal{S}^- u, v) = (u, \mathcal{S}^+ v)$$

Finally to $\mu$ and $\rho$ we associate the following (diagonal and positive) operators

$$\begin{aligned} \mathcal{M}u &= \{(\mu_{i+1} + \mu_i)u_i/2\} \\ \mathcal{R}u &= \{\rho_i u_i\} \end{aligned}$$

Then defining $\mathcal{T}$ by

$$\mathcal{T} = -\mathcal{S}^- \mathcal{M} \mathcal{S}^+$$

we can discretize the equation (2.6) as follows

$$\mathcal{R}(u^{n+1} - 2u^n + u^{n-1}) + \alpha^2(\mathcal{T} + \frac{3 - c^2\alpha^2}{12c^2}\mathcal{T}\mathcal{R}^{-1}\mathcal{T})u^n = 0 \tag{2.8}$$

We first give some properties of $\mathcal{T}$.

**Proposition 2** *The operator $\mathcal{T}$ is coercive and positive:*

$$(\mathcal{T}u,u) \geq \frac{7\mu_*}{16}\sum(u_{i+1}-u_i)^2$$

*and we have the energy inequality*

$$(\mathcal{T}u,u) \leq \frac{41\mu^*}{18}\sum u_i^2$$

*In addition* $\|\mathcal{T}\| \geq 16\mu_*/9$.

**Proof** Using the properties of the shift operator we get

$$\begin{aligned}(\mathcal{T}u,u) &= -(\mathcal{S}^-\mathcal{M}\mathcal{S}^+u,u)\\ &= (\mathcal{M}\mathcal{S}^+u,\mathcal{S}^+u)\\ &= \frac{1}{288}\sum(\mu_{i+1}+\mu_i)(u_{i+2}+9u_{i+1}-9u_i-u_{i-1})^2\end{aligned}$$

Consequently

$$\begin{aligned}(\mathcal{T}u,u) &\geq \frac{\mu_*}{144}\sum(u_{i+2}+9u_{i+1}-9u_i-u_{i-1})^2\\ &= \frac{\mu_*}{144}\sum 63(u_{i+1}-u_i)^2+18(u_{i+2}-u_i)^2+(u_{i+3}-u_i)^2 \qquad (2.9)\\ &\geq \frac{7\mu_*}{16}\sum(u_{i+1}-u_i)^2\end{aligned}$$

which proves the first statement. On the other hand

$$\begin{aligned}(\mathcal{T}u,u) &\leq \frac{\mu^*}{144}\sum(u_{i+2}+9u_{i+1}-9u_i-u_{i-1})^2\\ &= \frac{\mu^*}{144}\sum 63(u_{i+1}-u_i)^2+18(u_{i+2}-u_i)^2+(u_{i+3}-u_i)^2\\ &\leq \frac{\mu^*}{144}\sum 4(63+18+1)u_i^2\\ &= \frac{41\mu^*}{8}\sum u_i^2\end{aligned}$$

Here we have simply used the fact that $2ab \leq a^2+b^2$. Then supposing that $\mu(x)=\mu_*$ and considering the sequence $u_i=(-1)^i$ when $|i|\leq N$ and zero otherwise where $N$ is arbitrary we see that when $N$ grows $(\mathcal{T}u,u)\to 16\mu_*/9\sum u_i^2$ so that $\|\mathcal{T}\|\geq 16\mu_*/9$.

■

Evidently the value of the constant $41/18$ could be improved because the sums of the differences $(u_{i+l}-u_i)^2$ are not independent. This is most easily done with Fourier methods which gives

**Proposition 3**

$$\|\mathcal{T}\| \leq 16\mu^*/9$$

**Proof** Having first derived the inequality

$$(\mathcal{T}u, u) \leq \frac{\mu^*}{144}\sum(u_{i+2} + 9u_{i+1} - 9u_i - u_{i-1})^2$$

we note that this is like the constant coefficient case with the constant $\mu^*$. So (2.5) multiplied by $\mu^*$ gives the result.

■

To analyse the full scheme (2.8) we first define the operator $\mathcal{A}$ by

$$\mathcal{A} = \mathcal{T} + \frac{3 - c^2\alpha^2}{12c^2}\mathcal{T}\mathcal{R}^{-1}\mathcal{T} \tag{2.10}$$

Then we have

**Theorem 1** *The operator $\mathcal{A}$ is positive if $c^*\alpha \leq \sqrt{3}$ and the scheme (2.8) is stable if*

$$c^*\alpha \leq \sqrt{\frac{27 + 12\lambda^* - 3\sqrt{16\lambda^{*2} + 72\lambda^* - 27}}{8}} \tag{2.11}$$

*where $\lambda^* = \mu^*\rho^*/\mu_*\rho_* \geq 1$.*

In particular, putting $c^* = \lambda^* = 1$ we find the value given in (2.3). Note that this condition is sufficient but not necessary because the quantity whose upper limit $\lambda^*$ is may never attain it. To get sharper bounds we would have to make more precise assumptions about the behavior of $\mu$ and $\rho$.

**Proof** First note that $\mathcal{T}\mathcal{R}^{-1}\mathcal{T}$ is positive because

$$(\mathcal{T}\mathcal{R}^{-1}\mathcal{T}u, v) = (\mathcal{R}^{-1}\mathcal{T}u, \mathcal{T}v)$$

so evidently $\mathcal{A}$ is positive when $c^*\alpha \leq \sqrt{3}$. Then the inequality (2.4) can be written as

$$\alpha^2 \left\|\mathcal{R}^{-1}\mathcal{A}\right\| \leq 4$$

which leads to

$$\alpha^2 \|\mathcal{A}\| \leq 4\rho_* \tag{2.12}$$

Because of the positiveness we have (taking for the moment $\alpha$ and $c$ as fixed quantities)

$$\begin{aligned} \|\mathcal{A}\| &\leq \|\mathcal{T}\|\left(1 + \frac{3 - c^2\alpha^2}{12c^2}\left\|\mathcal{R}^{-1}\mathcal{T}\right\|\right) \\ &\leq \|\mathcal{T}\|\left(1 + \frac{3 - c^2\alpha^2}{12c^2\rho_*}\|\mathcal{T}\|\right) \end{aligned}$$

Then substituting the value of $\|\mathcal{T}\|$ and putting this into (2.12) we get

$$16c^2c^{*4}\alpha^4 - 12c^{*2}(4c^{*2} + 9c^2)\alpha^2 + 243c^2 \geq 0$$

Defining $1 \leq \lambda = \mu^*\rho/\mu\rho_* \leq \lambda^*$ and solving for $\alpha$ we find

$$c^*\alpha \leq \sqrt{\frac{27 + 12\lambda - 3\sqrt{16\lambda^2 + 72\lambda - 27}}{8}}$$

Then elementary verifications show that the right hand side attains its minimum when $\lambda = \lambda^*$ which gives the result.

■

We can easily see that

$$\frac{27 + 12\lambda - 3\sqrt{16\lambda^2 + 72\lambda - 27}}{8} < 3$$

for all $\lambda \geq 1$ so that we have

**Corollary 1** *If the scheme (2.8) is stable the operator $\mathcal{A}$ is positive.*

Finally let us consider the accuracy of (2.8). Using the Taylor's expansion we find that

$$\mathcal{R}(u^{n+1} - 2u^n + u^{n-1}) + \alpha^2(\mathcal{T} + \frac{3 - c^2\alpha^2}{12c^2}\mathcal{T}\mathcal{R}^{-1}\mathcal{T})u^n$$
$$= \rho u_{tt} - \partial_x(\mu u_x) +$$
$$\frac{1}{4\rho^2\mu}\Big(2\rho\mu(\rho\mu' - \mu\rho')u_{xxx} + \big(\mu\rho(\rho\mu'' - \mu\rho'') + 2(\rho^2\mu'^2 + \mu^2\rho'^2) - 5\mu\rho\mu'\rho'\big)u_{xx}$$
$$+\big(\mu'\rho'(2\mu\rho' - \rho\mu') + \rho\mu''(\rho\mu' - \mu\rho') - \rho\mu(\rho'\mu'' + \mu'\rho'')\big)u_x\Big)h^2 + O(h^4 + \delta t^4)$$

It is seen that the method is not completely of the fourth order. However, evidently the second order terms vanish if $\mu$ and $\rho$ are constant. Let us then remark that the choice of the operators $\mathcal{R}$ and $\mathcal{M}$ is not critical: they can be chosen in any consistent (and reasonable!) way, that is if $\rho = \mu = 1$ they must reduce to identity operators and they must be positive. Consequently by choosing $\mathcal{R}$ and $\mathcal{M}$ appropriately it may be possible to make some of the terms in the above expansion vanish, or at least reduce their coefficients.

# 3 More Dimensions

## 3.1 Isotropic Case

Next we will generalize the above schemes to the three dimensional case. The basic equation is then

$$\rho(x)u_{tt} - \nabla \cdot (\mu(x)\nabla u) = 0 \tag{3.1}$$

The natural generalization of (2.1) is evidently

$$\begin{aligned}\mathcal{S}_x u &= \frac{1}{48}(u(x+3h/2, y\pm h/2, z\pm h/2) + 9u(x+h/2, y\pm h/2, z\pm h/2) \\ &\quad -9u(x-h/2, y\pm h/2, z\pm h/2) - u(x-3h/2, y\pm h/2, z\pm h/2)) \quad (3.2)\end{aligned}$$

Of course this gives also the two dimensional scheme when we suppose that $u$ does not depend on $z$. Taylor's expansion gives

$$\frac{1}{h}\mathcal{S}_x u = u_x + \frac{h^2}{8}\Delta u_x + O(h^4)$$

To analyse the situation we use the shift operator as in the one dimensional case; however now we need shifts in all three directions. Anyway expanding $\mathcal{S}_x^2$ yields

$$\begin{aligned}\mathcal{S}_x^2 &= \frac{1}{2304}\Big((\mathcal{E}_x^{3/2} + 9\mathcal{E}_x^{1/2} - 9\mathcal{E}_x^{-1/2} - \mathcal{E}_x^{-3/2}) \\ &\quad (\mathcal{E}_y^{1/2} - \mathcal{E}_y^{-1/2})(\mathcal{E}_z^{1/2} - \mathcal{E}_z^{-1/2})\Big)^2 \\ &= \frac{1}{2304}\Big((\mathcal{E}_x^2 + 9\mathcal{E}_x - 9 - \mathcal{E}_x^{-1})(\mathcal{E}_y + 1)(\mathcal{E}_z + 1)\Big) \\ &\quad \Big((\mathcal{E}_x + 9 - 9\mathcal{E}_x^{-1} - \mathcal{E}_x^{-2})(\mathcal{E}_y^{-1} + 1)(\mathcal{E}_z^{-1} + 1)\Big)\end{aligned}$$

This suggests that we define

$$\begin{aligned}\mathcal{S}_x^+ &= \frac{1}{48}\Big((\mathcal{E}_x^2 + 9\mathcal{E}_x - 9 - \mathcal{E}_x^{-1})(\mathcal{E}_y + 1)(\mathcal{E}_z + 1)\Big) \\ \mathcal{S}_x^- &= \frac{1}{48}\Big((\mathcal{E}_x + 9 - 9\mathcal{E}_x^{-1} - \mathcal{E}_x^{-2})(\mathcal{E}_y^{-1} + 1)(\mathcal{E}_z^{-1} + 1)\Big)\end{aligned}$$

In the same way we define $\mathcal{S}_y^+$, $\mathcal{S}_y^-$, $\mathcal{S}_z^+$ and $\mathcal{S}_z^-$. Evidently we have

$$-(\mathcal{S}_x^- u, v) = (u, \mathcal{S}_x^+ v)$$

where the inner product is now taken in $l^2(\mathbb{Z}^3)$. The operators $\mathcal{M}$ and $\mathcal{R}$ become

$$\begin{aligned}\mathcal{M}u &= \{(\mu_{i+1,j,k} + \mu_{i,j,k} + \mu_{i+1,j+1,k} + \mu_{i,j+1,k} + \\ &\quad \mu_{i+1,j,k+1} + \mu_{i,j,k+1} + \mu_{i+1,j+1,k+1} + \mu_{i,j+1,k+1})u_{ijk}/8\} \\ \mathcal{R}u &= \{\rho_{ijk}u_{ijk}\}\end{aligned}$$

Here also the operators $\mathcal{M}$ and $\mathcal{R}$ could be defined in any reasonable way: the result would in any case be a stable and consistent scheme. The above choice just seemed to be simplest and the most natural. Finally we can define

$$\mathcal{T} = -\mathcal{S}_x^- \mathcal{M}\mathcal{S}_x^+ - \mathcal{S}_y^- \mathcal{M}\mathcal{S}_y^+ - \mathcal{S}_z^- \mathcal{M}\mathcal{S}_z^+ \quad (3.3)$$

so that with these notations the scheme can be written as before

$$\mathcal{R}(u^{n+1} - 2u^n + u^{n-1}) + \alpha^2(\mathcal{T} + \frac{3 - c^2\alpha^2}{12c^2}\mathcal{T}\mathcal{R}^{-1}\mathcal{T})u^n = 0$$

Obviously $\mathcal{T}$ is a positive operator and the situation is like in one dimensional case except that we still have to calculate the norm of $\mathcal{T}$. To facilitate the treatment of evaluating the inner products we introduce some useful formalism. Let us suppose that we want to calculate

$$(\mathcal{P}u, \mathcal{Q}v)$$

where $\mathcal{P}$ and $\mathcal{Q}$ are two difference operators and $u,\ v \in l^2(\mathbb{Z}^3)$. Without loss of generality we can take $\mathcal{P}$ and $\mathcal{Q}$ to be polynomials with variables $\mathcal{E}_x$, $\mathcal{E}_y$ and $\mathcal{E}_z$. Let us call the symbol of $\mathcal{P}$ the polynomial we get when we replace the operator $\mathcal{E}_x$ by $\xi_1$, $\mathcal{E}_y$ by $\xi_2$ and $\mathcal{E}_z$ by $\xi_3$. In the same way we define the symbol of $\mathcal{Q}$ but using the variables $\eta_1$, $\eta_2$ and $\eta_3$. Now it is not hard to see that if we calculate the product $p(\xi_1, \xi_2, \xi_3)q(\eta_1, \eta_2, \eta_3)$ then to every monomial $\xi_1^{a_1}\xi_2^{a_2}\xi_3^{a_3}\eta_1^{c_1}\eta_2^{c_2}\eta_3^{c_3}$ there corresponds the following sum in the inner product

$$\sum_{i,j,k} u_{i+a_1,j+a_2,k+a_3}v_{i+c_1,j+c_2,k+c_3}$$

Then we notice that when we calculate the product of $p$ and $q$ we can simplify it by the rules $\xi_i\eta_i = 1$ without changing the value of the inner product. If in addition $u = v$ we can further reduce the result because in every monomial we can interchange $\xi$'s and $\eta$'s without affecting the result: for instance $\xi_1^2\eta_2\eta_3^3 \sim \eta_1^2\xi_2\xi_3^3$.
So we can easily check if two inner products are the same: we write the corresponding symbols, use the simplification rules and see if the reduced polynomials are the same. With this tool we first give an energy result.

**Proposition 4** *The operator $\mathcal{T}$ is coercive and positive:*

$$(\mathcal{T}u, u) \geq \frac{7\mu_*}{256}\sum_{i,j,k}$$
$$(u_{i+1,j+1,k+1} + u_{i+1,j+1,k} - u_{i,j,k+1} - u_{i,j,k})^2 + (u_{i+1,j,k+1} + u_{i+1,j,k} - u_{i,j+1,k+1} - u_{i,j+1,k})^2 +$$
$$(u_{i+1,j+1,k+1} + u_{i+1,j,k+1} - u_{i,j+1,k} - u_{i,j,k})^2 + (u_{i+1,j+1,k} + u_{i+1,j,k} - u_{i,j+1,k+1} - u_{i,j,k+1})^2 +$$
$$(u_{i+1,j+1,k+1} + u_{i,j+1,k+1} - u_{i+1,j,k} - u_{i,j,k})^2 + (u_{i+1,j+1,k} + u_{i,j+1,k} - u_{i+1,j,k+1} - u_{i,j,k+1})^2$$

*and we have the energy inequality*

$$(\mathcal{T}u, u) \leq \frac{10\mu^*}{3}\sum u_{i,j,k}^2$$

*In addition* $\|\mathcal{T}\| \geq 16\mu_*/9$.

**Proof** Proceeding as in one dimensional case we find that

$$\begin{aligned}(\mathcal{T}u,u) &= -(\mathcal{S}_x^-\mathcal{M}\mathcal{S}_x^+u,u)-(\mathcal{S}_y^-\mathcal{M}\mathcal{S}_y^+u,u)-(\mathcal{S}_z^-\mathcal{M}\mathcal{S}_z^+u,u)\\ &= (\mathcal{M}\mathcal{S}_x^+u,\mathcal{S}_x^+u)+(\mathcal{M}\mathcal{S}_y^+u,\mathcal{S}_y^+u)+(\mathcal{M}\mathcal{S}_z^+u,\mathcal{S}_z^+u)\end{aligned}$$

So as before we naturally have

$$\begin{aligned}(\mathcal{T}u,u) &\geq \mu_*\Big((\mathcal{S}_x^+u,\mathcal{S}_x^+u)+(\mathcal{S}_y^+u,\mathcal{S}_y^+u)+(\mathcal{S}_z^+u,\mathcal{S}_z^+u)\Big)\\ (\mathcal{T}u,u) &\leq \mu^*\Big((\mathcal{S}_x^+u,\mathcal{S}_x^+u)+(\mathcal{S}_y^+u,\mathcal{S}_y^+u)+(\mathcal{S}_z^+u,\mathcal{S}_z^+u)\Big)\end{aligned}$$

Let us analyse for instance the term $(\mathcal{S}_x^+u,\mathcal{S}_x^+u)$, others being exactly analogous. Now the symbol of $\mathcal{S}_x^+$ (multiplied by 48) is evidently $p(\xi_1,\xi_2,\xi_3)=(\xi_1^3+9\xi_1^2-9\xi_1-1)(\xi_2+1)(\xi_3+1)$. Then using the simplification rules and factoring we find

$$p(\xi_1,\xi_2,\xi_3)p(\eta_1,\eta_2,\eta_3)=(164-63(\xi_1+\eta_1)-18(\xi_1^2+\eta_1^2)-\xi_1^3-\eta_1^3)(2+\xi_2+\eta_2)(2+\xi_3+\eta_3) \tag{3.4}$$

Now we look for the form analogous to (2.9) so we try the polynomial $\tilde{p}(\xi_1,\xi_2,\xi_3)=63(\xi_1-1)(\xi_2+1)(\xi_3+1)+18(\xi_1^2-1)(\xi_2+1)(\xi_3+1)+(\xi_1^3-1)(\xi_2+1)(\xi_3+1)$. Then multiplying $\tilde{p}(\xi_1,\xi_2,\xi_3)\tilde{p}(\eta_1,\eta_2,\eta_3)$ and simplifying we find the same expression as in (3.4). To get the contribution of the other operators we find that the symbol of $\mathcal{S}_y^+$ is $\tilde{p}(\xi_2,\xi_1,\xi_3)$ and that of $\mathcal{S}_z^+$ is $\tilde{p}(\xi_3,\xi_2,\xi_1)$. Taking into account only the terms with 63 and writing this out in terms of indices gives

$$(\mathcal{T}u,u)\geq\frac{7\mu_*}{256}\sum_{i,j,k}$$

$$\begin{aligned}&(u_{i+1,j+1,k+1}+u_{i+1,j+1,k}+u_{i+1,j,k+1}+u_{i+1,j,k}-u_{i,j+1,k+1}-u_{i,j+1,k}-u_{i,j,k+1}-u_{i,j,k})^2+\\ &(u_{i+1,j+1,k+1}+u_{i+1,j+1,k}+u_{i,j+1,k+1}+u_{i,j+1,k}-u_{i+1,j,k+1}-u_{i+1,j,k}-u_{i,j,k+1}-u_{i,j,k})^2+\\ &(u_{i+1,j+1,k+1}+u_{i,j+1,k+1}+u_{i+1,j,k+1}+u_{i,j,k+1}-u_{i+1,j+1,k}-u_{i,j+1,k}-u_{i+1,j,k}-u_{i,j,k})^2\end{aligned}$$

Then elementary verifications show that this can be 'factored' as in the first statement of the proposition.

To get the upper bound we could simply expand everything out and then use $2ab\leq a^2+b^2$. However, with a little effort we can do better. First define $y\in\mathbb{R}^8$ as follows

$$\begin{aligned}y_1 &= u_{i+1,j+1,k+1}\\ y_2 &= u_{i+1,j+1,k}\\ y_3 &= u_{i+1,j,k+1}\\ y_4 &= u_{i+1,j,k}\\ y_5 &= u_{i,j+1,k+1}\\ y_6 &= u_{i,j+1,k}\\ y_7 &= u_{i,j,k+1}\\ y_8 &= u_{i,j,k}\end{aligned}$$

Then the terms with 63 can be written as

$$\sum_{i,j,k}(y_1+y_2+y_3+y_4-y_5-y_6-y_7-y_8)^2+$$
$$(y_1+y_2-y_3-y_4+y_5+y_6-y_7-y_8)^2+$$
$$(y_1-y_2+y_3-y_4+y_5-y_6+y_7-y_8)^2$$

For each fixed triple $(i,j,k)$ we estimate the maximum value of the above expression. To this end we introduce the matrix

$$M=\begin{pmatrix} 3 & 1 & 1 & -1 & 1 & -1 & -1 & -3 \\ 1 & 3 & -1 & 1 & -1 & 1 & -3 & -1 \\ 1 & -1 & 3 & 1 & -1 & -3 & 1 & -1 \\ -1 & 1 & 1 & 3 & -3 & -1 & -1 & 1 \\ 1 & -1 & -1 & -3 & 3 & 1 & 1 & -1 \\ -1 & 1 & -3 & -1 & 1 & 3 & -1 & 1 \\ -1 & -3 & 1 & -1 & 1 & -1 & 3 & 1 \\ -3 & -1 & -1 & 1 & -1 & 1 & 1 & 3 \end{pmatrix}$$

Then the terms in the above sum are nothing but $y^tMy$ for different triples of indices. Next calculating the eigenvalues of $M$ we find that three of them are 8 and five are zero. From this it follows that $y^tMy \leq 8y^ty$ which means that

$$\sum_{i,j,k} y^tMy \leq 64\sum_{i,j,k} u_{ijk}^2$$

For other terms, that is the terms of the form $u_{i+2,j,k}$ and $u_{i+3,j,k}$ we simply use $2ab \leq a^2+b^2$ so that the energy inequality becomes

$$\begin{aligned}(\mathcal{T}u,u) &\leq \frac{\mu^*}{2304}\sum_{i,j,k}(63\cdot 64+3\cdot 18\cdot 64+3\cdot 64)u_{ijk}^2 \\ &= \frac{10\mu^*}{3}\sum_{i,j,k}u_{ijk}^2\end{aligned}$$

The final result follows if we consider the sequence $u_{ijk}=(-1)^i$ when $|i|+|j|+|k|\leq N$ and zero otherwise.

■

It is interesting to note that for the 'chess board' sequence $u_{ijk}=(-1)^{ijk}$ we have formally $\mathcal{T}u=0$ (however, $\mathcal{T}$ is anyway injective because this sequence does not belong to $l^2(\mathbb{Z}^3)$). This can be interpreted by saying that the method filters the very high frequency components.

Note the particular form of (3.4). There is an easy explanation; let us call a difference operator *rectangular* if its symbol can be written as $p_1(\xi_1)p_2(\xi_2)p_3(\xi_3)$. Then we can formulate the following result which is a direct consequence of the rules $\xi_i\eta_i=1$.

**Proposition 5** *If $\mathcal{P}$ and $\mathcal{Q}$ are rectangular operators whose symbols are $p$ and $q$ then there are polynomials $r_i$ and $\tilde{r}_i$, $1 \leq i \leq 3$ such that*

$$p(\xi_1,\xi_2,\xi_3)q(\eta_1,\eta_2,\eta_3) = \big(r_1(\xi_1)+\tilde{r}_1(\eta_1)\big)\big(r_2(\xi_2)+\tilde{r}_2(\eta_2)\big)\big(r_3(\xi_3)+\tilde{r}_3(\eta_3)\big)$$

*In particular if $\mathcal{P} = \mathcal{Q}$ then we can take $r_i = \tilde{r}_i$.*

■

As before the bounds obtained by the energy method are not optimal, so to get better bounds we use Fourier analysis.

**Proposition 6**

$$\|\mathcal{T}\| \leq \frac{16\mu^*}{9}$$

**Proof** Transforming the operator in (3.2) we get

$$\hat{\mathcal{S}}_x = i\Big(\sin(3k_1h/2)\cos(k_2h/2)\cos(k_3h/2) + 9\sin(k_1h/2)\cos(k_2h/2)\cos(k_3h/2)\Big)/6 \tag{3.5}$$

Treating similarly $\mathcal{S}_y$ and $\mathcal{S}_z$ one obtains

$$\|\mathcal{T}\| \leq \max_{k\in\mathbb{R}^3} \frac{\mu^*}{36}\Big($$
$$\Big(\sin(3k_1h/2)\cos(k_2h/2)\cos(k_3h/2) + 9\sin(k_1h/2)\cos(k_2h/2)\cos(k_3h/2)\Big)^2 +$$
$$\Big(\sin(3k_2h/2)\cos(k_1h/2)\cos(k_3h/2) + 9\sin(k_2h/2)\cos(k_1h/2)\cos(k_3h/2)\Big)^2 +$$
$$\Big(\sin(3k_3h/2)\cos(k_2h/2)\cos(k_1h/2) + 9\sin(k_3h/2)\cos(k_2h/2)\cos(k_1h/2)\Big)^2\Big)$$

To simplify the problem we introduce the following convenient notation.

$$\begin{aligned} a &= \cos^2(k_1h/2) \\ b &= \cos^2(k_2h/2) \\ c &= \cos^2(k_3h/2) \end{aligned}$$

Using these new variables we get

$$\|\mathcal{T}\| \leq \max_{0\leq a,b,c\leq 1} \mu^* F(a,b,c) =$$
$$\max_{0\leq a,b,c\leq 1} \frac{4\mu^*}{9}\big(4(ab+bc+ca) - abc(3a+3b+3c+a^2+b^2+c^2)\big) \tag{3.6}$$

Then forgetting the factor $4\mu^*/9$ our problem then becomes

$$\max f = 4(ab+bc+ca) - abc(3a+3b+3c+a^2+b^2+c^2)$$
$$0 \leq a, b, c \leq 1$$

Because of the symmetry we have only three cases.

**Case 1** Let us take $c = 0$, so that $f = 4ab$ whose maximum is evidently 4 when $0 \leq a, b \leq 1$.

**Case 2** Taking $c = 1$ gives $f = 4(ab+b+a) - ab(3a+3b+a^2+b^2+4)$. When $a$ and/or $b$ are either zero or one the maximum is again four. So there remains the interior points $0 < a, b < 1$. Taking the gradient we find that the only zero in the domain is $a = b = (\sqrt{33}-1)/8$. Substituting this into $f$ yields $(165\sqrt{33}-117)/256 \simeq 3.25$.

**Case 3** Finally the most difficult case $0 < a, b, c < 1$. First it is straight forward to calculate that the gradient vanishes at $a = b = c = (\sqrt{21}-1)/5$ and the corresponding value of $f$ is $(14688 - 1008\sqrt{21})/3125 \simeq 3.22$. Then we have to check that there are no other zeros of the gradient in the domain. First consider the possibility $c = b \neq a$. There are then two equations

$$\begin{aligned} f_a &= b(8 - 6ab - 3a^2b - 6b^2 - 2b^3) = 0 \\ f_b &= 4a + 4b - 3a^2b - a^3b - 9ab^2 - 4ab^3 = 0 \end{aligned}$$

Then tedious calculations show that the only possible solution is $a = b$ given above. This leaves us with the case $a \neq b \neq c$. The equations are then

$$\begin{aligned} f_a &= 4c + 4b - 3c^2b - c^3b - 3cb^2 - cb^3 - 6abc - 3a^2bc = 0 \\ f_b &= 4a + 4c - 3a^2c - a^3c - 3ac^2 - ac^3 - 6abc - 3ab^2c = 0 \\ f_c &= 4a + 4b - 3a^2b - a^3b - 3ab^2 - ab^3 - 6abc - 3abc^2 = 0 \end{aligned}$$

Then calculating $f_c - f_b$ and $f_c - f_a$ and dividing the first by $b - c$ and the second by $a - c$ yields

$$4 - 3a^2 - a^3 - 3ab - ab^2 - 3ac + 2abc - ac^2 = 0$$
$$4 - 3b^2 - b^3 - 3ab - a^2b - 3bc + 2abc - bc^2 = 0$$

Then subtracting the second equation from the first and dividing by $b - a$ one obtains

$$3a + a^2 + 3b + b^2 + 3c + c^2 = 0$$

Remembering that $0 < a, b, c < 1$ this is clearly impossible. So all in all the maximum value is 4 and putting back the factor $4\mu^*/9$ gives the final result.

■

This is the same bound as in one dimensional case which means that

**Proposition 7** *The stability condition of the three dimensional scheme is exactly the same as in one dimensional case given in (2.11).*

This is a little surprising since normally we expect that stability condition is stricter in higher dimensions (there is often a factor $1/\sqrt{n}$ where $n$ is the dimension of the space). Intuitively we may think that it is the high frequency components which 'tend' to be unstable. As noted above $\mathcal{T}$ annihilates some of them, so it is 'natural' that the stability condition remains 'good'.

## 3.2 (Almost) Anisotropic Case

To further generalize our analysis we now take $\mu$ to be a symmetric positive definite matrix. Trying to avoid the confusion with different indices we will denote its components by $\mu_{ij}$ and use the notation $M_{ijk}$ to mean the value of the matrix at a certain point.

As we have seen above the variable coefficients do not cause any problems of stability or consistency. In fact, for the method described above once we have the scheme in the constant coefficient case we can readily write down the corresponding variable coefficient scheme.

In the anisotropic case we take the same approach: first we treat the case $\mu$ constant and then extend it to variable $\mu$. Since $\rho$ remains a scalar we can suppose to begin with that $\rho = 1$. So our model problem can be written as

$$u_{tt} - \nabla \cdot (\mu \nabla u) = 0$$

To construct the fourth order scheme we start by considering the error in time discretization. Using the usual three point scheme leads to

$$\begin{aligned} \frac{u^{n+1} - 2u^n + u^{n-1}}{\delta t^2} &= u_t t + \frac{\delta t^2}{12} u_{tttt} + O(\delta t^4) \\ &= u_{tt} + \frac{\delta t^2}{12} \nabla \cdot \Big(\mu \nabla (\nabla \cdot (\mu \nabla u))\Big) + O(\delta t^4) \end{aligned}$$

This yields the term which must be added to the original equation in order to compensate the second order error in time. Now to have an efficient scheme the second order error in space should be of the same form. Let us first expand in terms of indices this supplementary term

$$\nabla \cdot \Big(\mu \nabla (\nabla \cdot (\mu \nabla u))\Big) = \sum_{i,j,k,l} \mu_{ij} \mu_{kl} \frac{\partial^4 u}{\partial x_i \partial x_j \partial x_k \partial x_l} \tag{3.7}$$

Here it is more convenient to use notation $x = \begin{pmatrix} x_1 & x_2 & x_3 \end{pmatrix}$ instead of $\begin{pmatrix} x & y & z \end{pmatrix}$. Denoting the generic term of the sum by $W_{ij}^{kl}$ we have the following symmetry relations.

$$W_{ij}^{kl} = W_{ji}^{kl}$$

$$
\begin{aligned}
W_{ij}^{kl} &= W_{ij}^{lk} \\
W_{ij}^{kl} &= W_{kl}^{ij}
\end{aligned}
\tag{3.8}
$$

Now it is clear that the operator (3.2) as such is not sufficient, evidently the coefficients have to depend on $\mu$. To analyse the situation let us note first of all that for any consistent symmetric approximation $\mathcal{D}_{x_l}u$ to $u_{x_l}$ we have

$$
\mathcal{D}_{x_l}u = u_{x_l} + \frac{h^2}{2}\sum_{ij} w_{ij}^{l}\frac{\partial^3 u}{\partial x_l \partial x_i \partial x_j} + O(h^4)
$$

where $w_{ij}^{l}$ are some numbers. In this way one can then define the discrete gradient by $\mathcal{G} = \begin{pmatrix} \mathcal{D}_{x_1} & \mathcal{D}_{x_2} & \mathcal{D}_{x_3} \end{pmatrix}^t$ and the discrete divergence by $\mathcal{D} = \begin{pmatrix} \mathcal{D}_{x_1} & \mathcal{D}_{x_2} & \mathcal{D}_{x_3} \end{pmatrix}$. Multiplying this discrete gradient by $\mu$ and finally taking the discrete divergence we obtain

$$
\begin{aligned}
\mathcal{D}\mu\mathcal{G}u &= \sum \mu_{ij}\mathcal{D}_{x_i}\mathcal{D}_{x_j}u \\
&= \nabla\cdot(\mu\nabla u) + h^2\sum_{ijkl}\mu_{kl}w_{ij}^{l}\frac{\partial^4 u}{\partial x_i\partial x_j\partial x_k\partial x_l}
\end{aligned}
$$

Now the symmetry relations (3.8) are verified only if $w_{ij}^{l}$ do not depend on $l$ and then comparing to (3.7) we conclude that the compensation is possible only if $w_{ij} = c\mu_{ij}$, where $c$ is some constant. We summarize this in

**Proposition 8** *To construct a fourth order scheme we need a difference operator $\mathcal{D}_{x_k}$ such that*

$$
\mathcal{D}_{x_k}u = u_{x_k} + ch^2\sum_{ij}\mu_{ij}\frac{\partial^3 u}{\partial x_i\partial x_j\partial x_k} + O(h^4)
$$

■

So how to construct such an operator? To explore the possibilities we introduce the operator

$$
\begin{aligned}
\mathcal{S}_{x_1}u \;=\; & b_1\big(u(x_1+3h/2,x_2+h/2,x_3+h/2) - u(x_1-3h/2,x_2-h/2,x_3-h/2)\big) + \\
& b_2\big(u(x_1+3h/2,x_2+h/2,x_3-h/2) - u(x_1-3h/2,x_2-h/2,x_3+h/2)\big) + \\
& b_3\big(u(x_1+3h/2,x_2-h/2,x_3+h/2) - u(x_1-3h/2,x_2+h/2,x_3-h/2)\big) + \\
& b_4\big(u(x_1+3h/2,x_2-h/2,x_3-h/2) - u(x_1-3h/2,x_2+h/2,x_3+h/2)\big) + \\
& a_1\big(u(x_1+h/2,x_2+h/2,x_3+h/2) - u(x_1-h/2,x_2-h/2,x_3-h/2)\big) + \\
& a_2\big(u(x_1+h/2,x_2+h/2,x_3-h/2) - u(x_1-h/2,x_2-h/2,x_3+h/2)\big) + \\
& a_3\big(u(x_1+h/2,x_2-h/2,x_3+h/2) - u(x_1-h/2,x_2+h/2,x_3-h/2)\big) + \\
& a_4\big(u(x_1+h/2,x_2-h/2,x_3-h/2) - u(x_1-h/2,x_2+h/2,x_3+h/2)\big)
\end{aligned}
$$

In the same way we define $\mathcal{S}_{x_2}$ and $\mathcal{S}_{x_3}$. Now first of all $\mathcal{S}_{x_1}u$ has to give a consistent approximation to $u_{x_1}$ so that using Taylor's expansion we get the equations

$$\begin{aligned} a_1 + a_2 + a_3 + a_4 + 3b_1 + 3b_2 + 3b_3 + 3b_4 &= 1 \\ a_1 + a_2 - a_3 - a_4 + b_1 + b_2 - b_3 - b_4 &= 0 \\ a_1 - a_2 + a_3 - a_4 + b_1 - b_2 + b_3 - b_4 &= 0 \end{aligned}$$

Solving this we get

$$\begin{aligned} b_2 &= (1 - 4a_1 - 4a_2 + 2a_3 + 2a_4 - 6b_1)/6 \\ b_3 &= (1 - 4a_1 + 2a_2 - 4a_3 + 2a_4 - 6b_1)/6 \\ b_4 &= a_1 - a_4 + b_1 \end{aligned}$$

Substituting these values into $\mathcal{S}_{x_1}$ and expanding we get the following

$$\begin{aligned} \frac{1}{h}\mathcal{S}_{x_1}u &= u_{x_1} + \frac{h^2}{24}\Big(\big(9 - 8(a_1 + a_2 + a_3 + a_4)\big)\frac{\partial^3 u}{\partial x_1^3} \\ &+ 24(-a_1 - a_2 + a_3 + a_4)\frac{\partial^3 u}{\partial x_1^2 \partial x_2} + 24(-a_1 + a_2 - a_3 + a_4)\frac{\partial^3 u}{\partial x_1^2 \partial x_3} \\ &+ (-6 + 48a_1 - 24a_4 + 72b_1)\frac{\partial^3 u}{\partial x_2 \partial x_3 \partial x_1} + 3\frac{\partial^3 u}{\partial x_2^2 \partial x_1} + 3\frac{\partial^3 u}{\partial x_3^2 \partial x_1}\Big) + O(h^4) \end{aligned}$$

Because $\mu$ is symmetric we obtain the following six equations.

$$\begin{aligned} c\mu_{11} &= 9 - 8(a_1 + a_2 + a_3 + a_4) \\ c\mu_{22} &= 3 \\ c\mu_{33} &= 3 \\ c\mu_{12} &= 12(-a_1 - a_2 + a_3 + a_4) \\ c\mu_{31} &= 12(-a_1 + a_2 - a_3 + a_4) \\ c\mu_{23} &= -3 + 24a_1 - 12a_4 + 36b_1 \end{aligned}$$

Of course this is impossible if $\mu_{22} \neq \mu_{33}$. Difference operators $\mathcal{S}_{x_2}$ and $\mathcal{S}_{x_3}$ give similar conditions so to have a solution we must demand that $\mu_{11} = \mu_{22} = \mu_{33}$. To simplify the notation (but without the loss of generality) let us take this value to be one so that $c = 3$ and this leaves us with the following four equations.

$$\begin{aligned} a_1 + a_2 + a_3 + a_4 &= 3/4 \\ -a_1 - a_2 + a_3 + a_4 &= \mu_{12}/4 \\ -a_1 + a_2 - a_3 + a_4 &= \mu_{31}/4 \\ 8a_1 - 4a_4 + 12b_1 &= \mu_{23} + 1 \end{aligned}$$

Solving this we get

$$\begin{aligned} a_2 &= (3 - 8a_1 - \mu_{12})/8 \\ a_3 &= (3 - 8a_1 - \mu_{31})/8 \\ a_4 &= (8a_1 + \mu_{12} + \mu_{31})/8 \\ b_1 &= (2 - 8a_1 + \mu_{12} + \mu_{31} + 2\mu_{23})/24 \end{aligned}$$

Evidently we have to choose $a_1$ such that when $\mu_{12} = \mu_{23} = \mu_{31} = 0$ we get the original (3.2). So any reasonable choice is of the form $a_1 = c_1\mu_{12} + c_2\mu_{31} + c_3\mu_{23} + 3/16$. The most symmetric choice seems to be $c_1 = -1/16$, $c_2 = -1/16$ and $c_3 = 1/16$. So finally all the coefficients are given by

$$\begin{aligned} a_1 &= (3 - \mu_{12} - \mu_{31} + \mu_{23})/16 \\ a_2 &= (3 - \mu_{12} + \mu_{31} - \mu_{23})/16 \\ a_3 &= (3 + \mu_{12} - \mu_{31} - \mu_{23})/16 \\ a_4 &= (3 + \mu_{12} + \mu_{31} + \mu_{23})/16 \\ b_1 &= (1 + 3\mu_{12} + 3\mu_{31} + 3\mu_{23})/48 \\ b_2 &= (1 + 3\mu_{12} - 3\mu_{31} - 3\mu_{23})/48 \\ b_3 &= (1 - 3\mu_{12} + 3\mu_{31} - 3\mu_{23})/48 \\ b_4 &= (1 - 3\mu_{12} - 3\mu_{31} + 3\mu_{23})/48 \end{aligned}$$

To obtain the coefficients for $\mathcal{S}_{x_2}$ and $\mathcal{S}_{x_3}$ we simply permute the indices as follows.

$$\begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix} \qquad \begin{pmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{pmatrix}$$

The permutation on the left is for $\mathcal{S}_{x_2}$ and the permutation on the right is for $\mathcal{S}_{x_3}$. Next, as before we want to get rid of 'intermediate' points. To this end we again calculate with symbols of operators. Denoting the symbol of $\mathcal{S}_{x_1}$ by $p_1$ we obtain

$$\begin{aligned} p_1^2 = \Big(& a_1(\xi_1\xi_2\xi_3 - 1) + a_2(\xi_1\xi_2 - \xi_3) + a_3(\xi_1\xi_3 - \xi_2) + a_4(\xi_1 - \xi_2\xi_3) + \\ & (b_1(\xi_1^2\xi_2\xi_3 - \xi_1^{-1}) + b_2(\xi_1^2\xi_2 - \xi_1^{-1}\xi_3) + b_3(\xi_1^2\xi_3 - \xi_1^{-1}\xi_2) + b_4(\xi_1^2 - \xi_1^{-1}\xi_2\xi_3)\Big) \\ \Big(& a_1(1 - \xi_1^{-1}\xi_2^{-1}\xi_3^{-1}) + a_2(\xi_3^{-1} - \xi_1^{-1}\xi_2^{-1}) + a_3(\xi_2^{-1} - \xi_1^{-1}\xi_3^{-1}) + \\ & a_4(\xi_2^{-1}\xi_3^{-1} - \xi_1^{-1}) + (b_1(\xi_1 - \xi_1^{-1}\xi_2^{-1}\xi_3^{-1}) + b_2(\xi_1\xi_3^{-1} - \xi_1^{-2}\xi_2^{-1}) + \\ & b_3(\xi_1\xi_2^{-1} - \xi_1^{-2}\xi_3^{-1}) + b_4(\xi_1\xi_2^{-1}\xi_3^{-1} - \xi_1^{-2})\Big) \end{aligned}$$

Then we use the first factor on the right to define the operator $\mathcal{S}_{x_1}^+$ and the second factor to define $\mathcal{S}_{x_1}^-$. Exactly as before we have also the fundamental property

$$-(\mathcal{S}_{x_1}^- u, v) = (u, \mathcal{S}_{x_1}^+ v)$$

In the same way we define also $\mathcal{S}^+_{x_2}$ etc. With these operators we then define the discrete gradient as $\mathcal{G} = (\mathcal{S}^+_{x_1}, \mathcal{S}^+_{x_2}, \mathcal{S}^+_{x_3})^t$ and the discrete divergence as $\mathcal{D} = (\mathcal{S}^-_{x_1}, \mathcal{S}^-_{x_2}, \mathcal{S}^-_{x_3})$; $\mathcal{G} : l^2(\mathbb{Z}^3) \to l^2(\mathbb{Z}^3)^3$ and $\mathcal{D} : l^2(\mathbb{Z}^3)^3 \to l^2(\mathbb{Z}^3)$. Then we obtain

**Proposition 9**

$$\mathcal{D}^t = -\mathcal{G}$$

**Proof** By definition, $\mathcal{B}$ is the transpose of D if

$$(\mathcal{D}X, y)_1 = (X, \mathcal{B}y)_3$$

for all $X \in l^2(\mathbb{Z}^3)^3$ and for all $y \in l^2(\mathbb{Z}^3)$ and on the left we have the inner product in $l^2(\mathbb{Z}^3)$ and on the right the inner product in $l^2(\mathbb{Z}^3)^3$. The latter is defined by

$$(X, Y)_3 = \sum_{ijk} X_{ijk} \cdot Y_{ijk}$$

So using these definitions we obtain

$$\begin{aligned}(\mathcal{D}X, y)_1 &= (\mathcal{S}^-_{x_1}X^1 + \mathcal{S}^-_{x_2}X^2 + \mathcal{S}^-_{x_3}X^3, y)_1 \\ &= -(X^1, \mathcal{S}^+_{x_1}y)_1 - (X^2, \mathcal{S}^+_{x_2}y)_1 - (X^3, \mathcal{S}^+_{x_3}y)_1 \\ &= -(X, \mathcal{G}y)_3\end{aligned}$$

Exactly in the same way we can show that $\mathcal{G}^t = -\mathcal{D}$.

■

Now putting $\tilde{\mu} = \mu_{ii} = \mu_{jj} = \mu_{kk} \neq 1$ we get the corresponding coefficients simply by replacing $\mu_{ij}$ by $\mu_{ij}/\tilde{\mu}$. To get the complete scheme we still have to define

$$\begin{aligned}\mathcal{M}X &= \{(M_{i+1,j,k} + M_{i,j,k} + M_{i+1,j+1,k} + M_{i,j+1,k} + \\ &\quad M_{i+1,j,k+1} + M_{i,j,k+1} + M_{i+1,j+1,k+1} + M_{i,j+1,k+1})X_{ijk}/8\} \\ \mathcal{R}u &= \{\rho_{ijk}u_{ijk}\}\end{aligned}$$

where $X \in l^2(\mathbb{Z}^3)^3$. Then we define

$$\mathcal{T} = -\mathcal{D}\mathcal{M}\mathcal{G}$$

so that the whole scheme is still of the form

$$\mathcal{R}(u^{n+1} - 2u^n + u^{n-1}) + \alpha^2(\mathcal{T} + \frac{3 - c^2\alpha^2}{12c^2}\mathcal{T}\mathcal{R}^{-1}\mathcal{T})u^n = 0 \qquad (3.9)$$

where $c^2 = \tilde{\mu}/\rho$ and we recall that $\mu$ is the diagonal element of $M$. Note that now $c$ cannot really be interpreted as the signal speed. So defining $\mathcal{A}$ as in (2.10) we obtain

**Proposition 10** *The operator $\mathcal{T}$ is positive and symmetric. This implies that $\mathcal{A}$ is also positive if $\alpha c \leq \sqrt{3}$.*

**Proof** The positivity of $\mathcal{T}$ is immediate:

$$\begin{aligned}(\mathcal{T}u, u)_1 &= -(\mathcal{D}\mathcal{M}\mathcal{G}u, u)_1 \\ &= (\mathcal{M}\mathcal{G}u, \mathcal{G}u)_3 \geq 0\end{aligned}$$

Because $M$ is symmetric it follows that $\mathcal{T}$ is also symmetric, that is

$$(\mathcal{T}u, v) = (u, \mathcal{T}v)$$

so that $\mathcal{T}\mathcal{R}^{-1}\mathcal{T}$ is also positive. Consequently the operator $\mathcal{A}$ is positive for $c\alpha \leq \sqrt{3}$. ■

This positivity property then implies

**Corollary 2** *The scheme (3.9) is stable for sufficiently small $\alpha$.*

It seems difficult to calculate the stability condition more precisely in general, that is without making some more or less arbitrary assumptions on the components of $\mu$. Instead, let us consider some numerical examples. For simplicity we take the diagonal to be one, as well as $\rho$. Then the stability condition can be written as

$$\alpha^2 \|\mathcal{T}\| \left(1 + \frac{3 - \alpha^2}{12} \|\mathcal{T}\|\right) \leq 4$$

Denoting the norm of $\mathcal{T}$ by $\hat{t}$ and solving this one obtains

$$\alpha \leq \sqrt{\frac{12 + 3\hat{t} - \sqrt{9\hat{t}^2 + 72\hat{t} - 48}}{2\hat{t}}}$$

Then we present in the following tables some test cases. In the first column there are the parameters $\mu_{12}$, $\mu_{23}$ and $\mu_{31}$ (in that order), in the second column there are the square roots of the eigenvalues, in the third and fourth column are the direction angles of the eigenvectors corresponding to the eigenvalues which are on the same row and finally in the fifth column there are first the norm of $\mathcal{T}$ and then the stability limit.

We note that there seems to be no direct relation between the maximum eigenvalue and the stability limit; for instance in the first and second test cases the maximum eigenvalues are quite close, although there is big difference in the stability limits. Then we recall that we had one free parameter in our difference operators which might be used to improve the stability condition.

Finally let us remark that taking our basic operator a little longer it would be possible to treat the arbitrary tensors using the same ideas as above. Note also that there does

| | | | | |
|---|---|---|---|---|
| −0.4 | 1.27 | 137 | 30 | 2.21 |
| 0.2 | 0.76 | 38 | 15 | 1.18 |
| −0.3 | 0.90 | 106 | −55 | |

Table 1: First test case.

| | | | | |
|---|---|---|---|---|
| 0.1 | 1.02 | 14 | −6 | 3.38 |
| 0.8 | 0.38 | 110 | −44 | 0.85 |
| −0.2 | 1.34 | 97 | 45 | |

Table 2: Second test case.

| | | | | |
|---|---|---|---|---|
| 0.6 | 1.48 | 48 | 35 | 3.04 |
| 0.7 | 0.72 | 167 | 34 | 0.92 |
| 0.5 | 0.53 | 107 | −36 | |

Table 3: Third test case.

| | | | | |
|---|---|---|---|---|
| −0.2 | 0.71 | 31 | 39 | 2.35 |
| −0.1 | 1.19 | 165 | 41 | 1.13 |
| −0.4 | 1.04 | 99 | −25 | |

Table 4: Fourth test case.

not in general exist any orthogonal transformation allowing to reduce any symmetric matrix to the form where diagonal elements are equal; this can be seen already in two dimensions. However, in some problems of elasticity, there appear naturally tensors with equal diagonal elements, so that maybe this condition is not after all very restrictive. Of course it would be nice to be able to characterize such tensors, but we are not aware that such a characterization exists.

In some situations one could achieve a greater generality by using different $h$ in different directions. However, to be able to profit these extra parameters one has to suppose that the eigenvectors of the tensor remain constant; different $h$'s would be useful essentially if the tensor is diagonal and diagonal elements are not the same.

# 4 Accuracy Considerations

## 4.1 Group Speed and Phase Speed

When we want to examine the accuracy of any scheme there are two things to be taken into account: the speed (absolute value) and the direction of propagation. In the wave phenomena there are two relevant wave speeds, namely the phase speed and the group speed and they are defined for the plane waves as

$$\begin{aligned} v_\varphi &= \frac{\omega}{|k|^2}k \\ v_g &= \nabla\omega \end{aligned}$$

where the gradient is taken with respect to $k$. Of course for the continuous problem we have $v_\varphi = v_g$ if the coefficients are constant, but for the discrete problem they are different. In fact the group speed is more important, because the 'energy' propagates (approximately) with the group speed. We recall the heuristic reasoning leading to this conclusion (see [BCL], see also [TR] for the discussion on the relevance of the group speed in wave phenomena). Consider the following function.

$$u(x,t) = \int a(k)e^{i(k\cdot x-\omega t)}dk \tag{4.1}$$

Taking a sufficiently nice $a$ and $\omega = |k|$, this evidently is a solution of (3.1) when $\rho = \mu = 1$. However, we can as well take $\omega$ to be any reasonable function of $k$. The function $u$ is called a wave packet if the support of $a$ is 'small'. For definiteness we suppose that $\mathrm{supp}(a) \subset B(\tilde{k},\varepsilon) = \{\, k \in \mathbb{R}^n \,|\, |k-\tilde{k}| < \varepsilon \,\}$. We normalize the amplitude of $u$ to be of the order of unity, that is

$$\int |a(k)|\, dk = O(1)$$

In fact, if we imagine that the support of $a$ gets smaller and smaller, then essentially $a$ approaches the Dirac measure. Next, in $B(\tilde{k},\varepsilon)$ we can write

$$\omega(k) = \omega(\tilde{k}) + \nabla\omega\cdot(k-\tilde{k}) + O(\varepsilon^2)$$

where the gradient is evaluated at $\tilde{k}$. Substituting this into (4.1) and putting $\delta k = k-\tilde{k}$ gives

$$\begin{aligned} u(x,t) &= \int a(k)\exp\Big(i((\tilde{k}+\delta k)\cdot x - (\omega(\tilde{k}) + \nabla\omega\cdot\delta k + O(\varepsilon^2))t)\Big)dk \\ &= e^{i(\tilde{k}\cdot x-\omega(\tilde{k})t)}\int a(k)\exp\Big(i(\delta k\cdot x - (\nabla\omega\cdot\delta k + O(\varepsilon^2))t)\Big)dk \\ &= e^{i(\tilde{k}\cdot x-\omega(\tilde{k})t)}\int a(k)\exp\Big(i(\delta k\cdot x - \nabla\omega\cdot\delta k\, t\Big)dk + t\,O(\varepsilon^2) \end{aligned}$$

In the last 'equality' the normalization condition was used. Then using the 'initial condition' $u_0(x) = u(x,0)$ the above expression further simplifies to

$$u(x,t) = e^{i\,t(\nabla\omega\cdot\tilde{k}-\omega(\tilde{k}))}u_0(x-\nabla\omega\,t) + t\,O(\varepsilon^2)$$

So there is a phase change and an 'error' of the order $tO(\varepsilon^2)$. Taking the absolute values we get

$$|u(x,t)| = |u_0(x-\nabla\omega\,t)| + t\,O(\varepsilon^2)$$

So if $t$ and $\epsilon$ are sufficiently small the profile of the absolute value travels with the group speed. In particular, the 'edge' or 'wave front' of the signal travels approximately with the group speed.

Finally let us make some general remarks of the form of $\omega$. Certainly it is natural to have $\omega(k) \geq 0$ and $\omega(0) = 0$, often $\omega$ is convex (at least in the neighborhood of origin) and sometimes it is homogeneous. To see how these properties affect the group and the phase speed, we state the following simple result

**Proposition 11** *Suppose that we can write the dispersion relation as $\omega^n = g(k)$ where $n$ is positive and $g(k) \geq 0$ and $g(0) = 0$. Then if $g$ is convex we have*

$$|v_\varphi| \leq n\,|v_g|$$

*If $g$ is homogeneous of degree $m$ (that is $g(\lambda k) = \lambda^m g(k)$) then*

$$|v_\varphi| \leq \frac{n}{m}|v_g|$$

Of course a convex function need not be homogeneous and a homogeneous function need not be convex.

**Proof** For $g$ convex and differentiable we have

$$g(\tilde{k}) - g(k) \geq \nabla g(k)\cdot(\tilde{k}-k)$$

Then taking $\tilde{k} = 0$ gives

$$g(k) \leq \nabla g(k)\cdot k \leq |\nabla g(k)|\,|k|$$

Then using $n\omega^{n-1}\nabla\omega = \nabla g$ and $\omega^n = g$ we get the first result.

To prove the second part we first note that

$$\frac{d}{d\lambda}g(\lambda k) = k\cdot\nabla g(\lambda k)$$

On the other hand if $g$ is homogeneous we have

$$\frac{d}{d\lambda}g(\lambda k) = \frac{d}{d\lambda}\lambda^m g(k) = m\lambda^{m-1}g(k)$$

Putting these together, setting $\lambda = 1$ and otherwise proceeding as in the convex case leads to the second result.

■

As a simple consequence we can estimate the error in the group speed relative to the phase speed. If the order of the scheme is $p$, then the dispersion relation is asymptotically

$$\omega = |k| + C(k)h^p + O(h^{p+1})$$

Now $C(k)$ is homogeneous of degree $p+1$ so that one can conclude that the error in the group speed is about $p+1$ times greater than the error in the phase speed. Of course this can also be calculated in a completely elementary way. However, we wanted to point out that even though the discrete dispersion relations are not homogeneous, they are 'asymptotically homogeneous' and this can be important in some situations. All in all, there are then two reasons to concentrate on the group speed: it is more 'physical' and numerically more annoying.

Let us then calculate the numerical group speed for the equation (3.1) with $\rho = \mu = 1$. First recall that in the continuous case we then have $v_\varphi = v_g = k/|k|$. Using the function $F$ of (3.6) the dispersion relation is

$$4\sin^2(\omega\delta t/2) = \alpha^2 F(k,h)(1 + \frac{3-\alpha^2}{12}F(k,h))$$

The corresponding two (resp. one) dimensional case is obtained simply by putting $k_3 = 0$ (resp. $k_2 = k_3 = 0$). Taking the gradient we find

$$\nabla\omega = \frac{\alpha^2(6 + (3-\alpha^2)F)}{12\delta t \sin(\omega\delta t)}\nabla F$$

To measure the error in the direction of the group speed we use the angle $\beta$ defined by

$$\cos\beta = \frac{\nabla\omega \cdot k}{|\nabla\omega|\,|k|} = \frac{\nabla F \cdot k}{|\nabla F|\,|k|}$$

Note that $\beta$ depends only on the second order operator $F$. However, expanding we find that

$$\beta = O(h^4) \tag{4.2}$$

For more details about this phenomenon see [TU].

## 4.2 Some Dispersion Curves

Let us take up directly the three dimensional case. Without the loss of generality we can take $\rho = \mu = |k| = 1$ and parametrize $k$ as

$$\begin{aligned} k_1 &= \cos\varphi\cos\theta \\ k_2 &= \sin\varphi\cos\theta \\ k_3 &= \sin\theta \end{aligned}$$

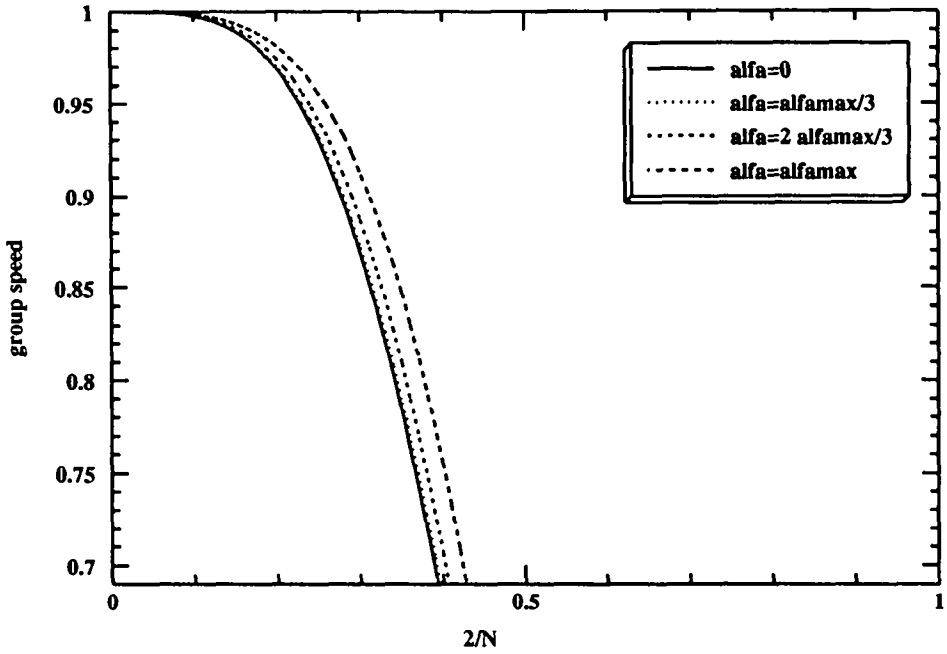



Figure 4.1: Group speed in the direction $\varphi = 20$, $\theta = 30$.

So we have four relevant parameters: $\alpha$, $\varphi$, $\theta$ and $h$. Instead of $h$, however, it is convenient to think in terms of points per wavelength which is $N = 2\pi/|k|\,h = 2\pi/h$. Now to be able to represent a wave in the grid we must have $N > 2$ or in other words $h < \pi$.

In figure 4.1 we show the group speed as function of $2/N$ for different $\alpha$; picking other values for $\theta$ and $\varphi$ does not essentially change the situation. Evidently it is best to choose $\alpha$ to be exactly at the stability limit, as is often the case with hyperbolic equations. On the other hand the error does not depend very sensitively on the choice of $\alpha$. Next in figure 4.2 the variation of the group speed is shown as a function of the direction, using the optimal $\alpha$ and $N = 5$. Figure 4.3 is similar but with $N = 9$. It is seen that this is sufficient to get an error which is less than 2 %. In figure 4.4 we show the error in the direction of the group speed as a function of $N$ and finally in figure 4.5 there is the same error when $N = 5$. Here it is more difficult to say what should be an acceptable level of error, but the maximum error less than 2 degrees with already $N = 5$ seems rather satisfactory. Anyway, when $N$ is less than about four the approximation degrades rapidly and consequently these high frequencies are only numerical noise. Note that the error in the direction approaches 180 degrees when $N$ tends to 2; this happens with any scheme and is due to the periodicity of the discrete dispersion relation.

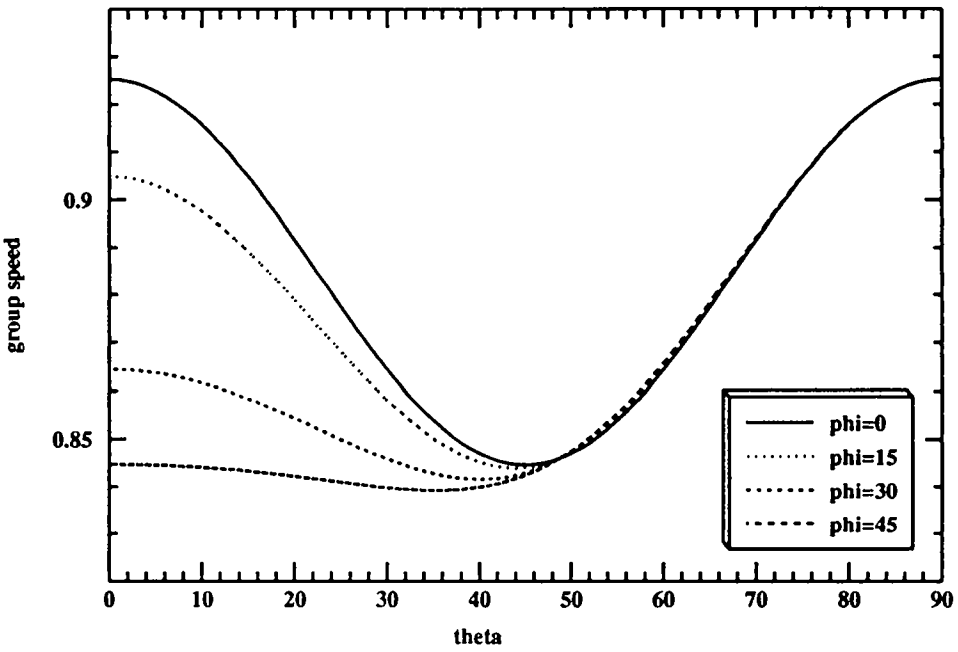



Figure 4.2: Group speed with $N = 5$.

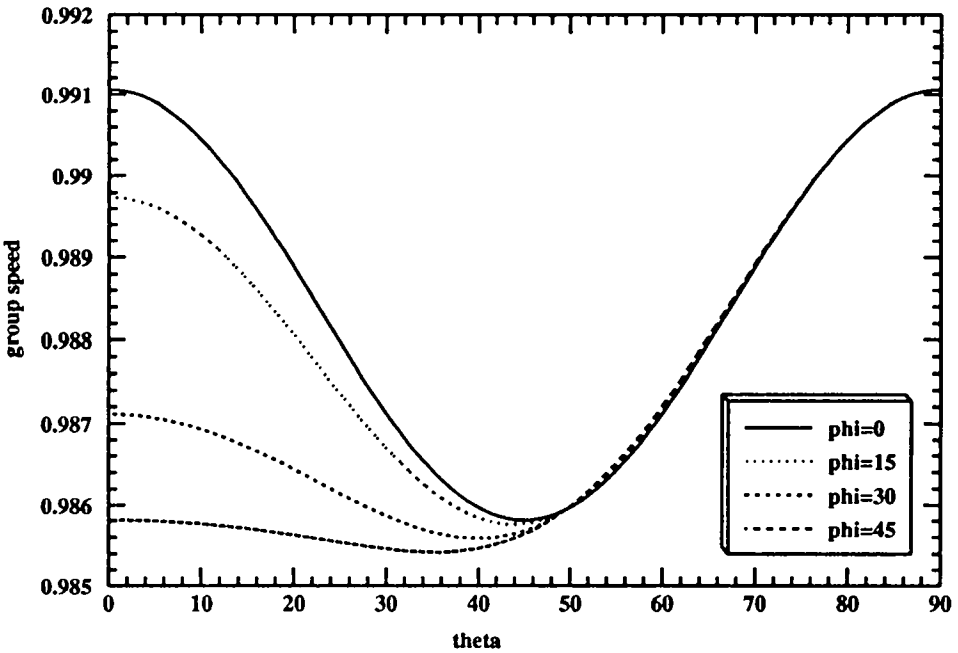



Figure 4.3: Group speed with $N = 9$.

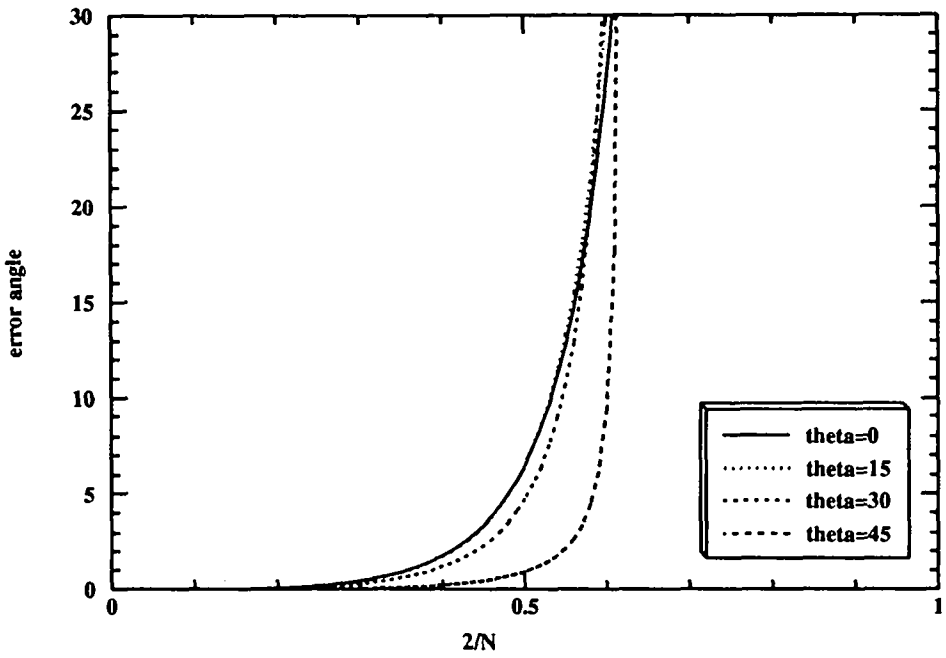



Figure 4.4: Error in the direction of the group speed with $\varphi = 20$.

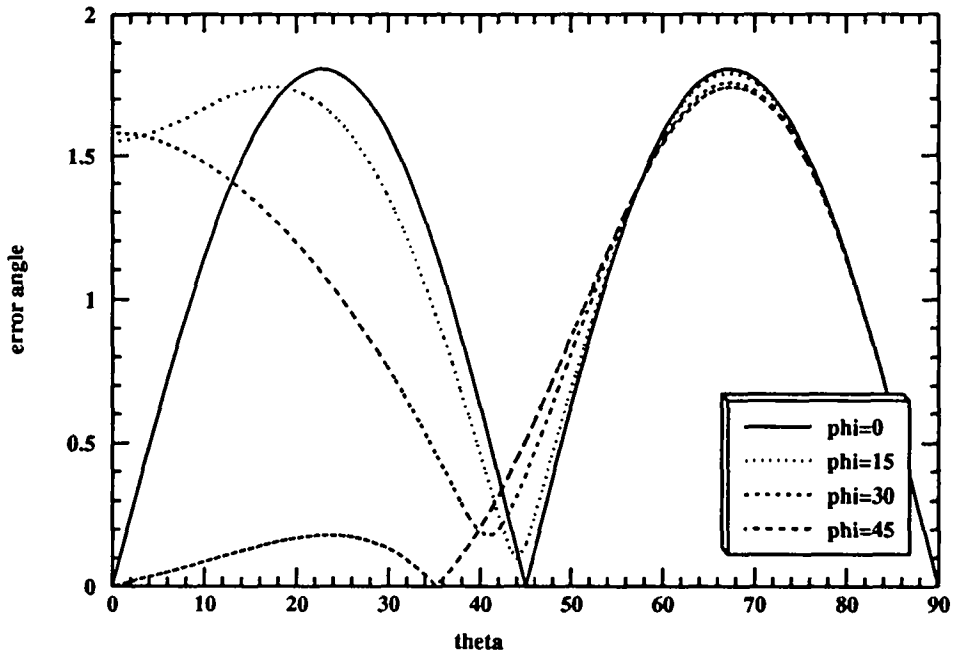



Figure 4.5: Error in the direction of the group speed with $N = 5$.

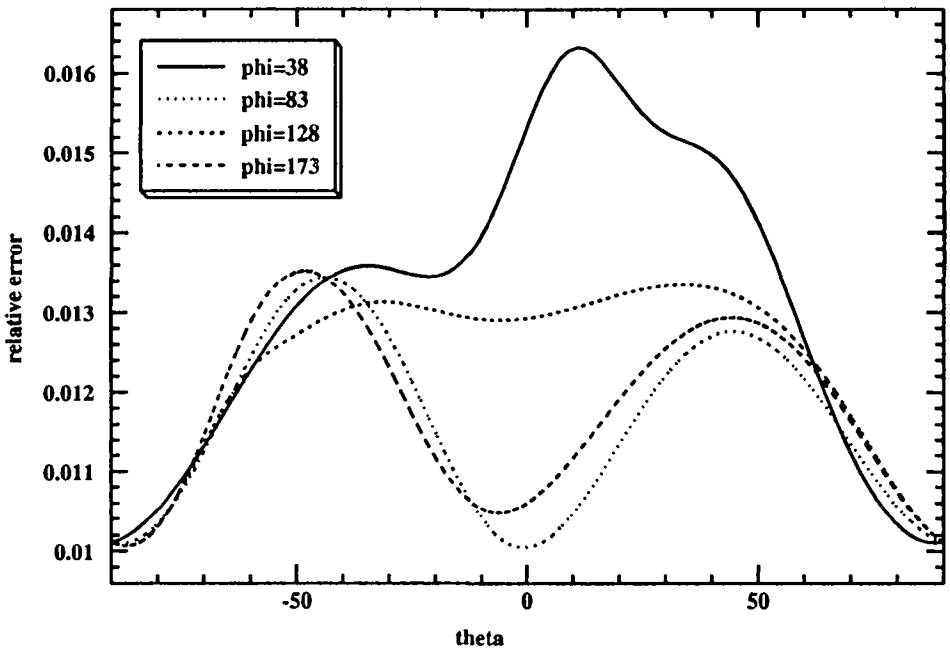



Figure 4.6: Relative error of the group speed; first test case.

Then we give the dispersion curves corresponding to the test cases in tables 1, 2, 3 and 4. Because now the speed is different in different directions we plot directly the relative error as a function of the direction. In anisotropic case the dispersion relation is $\omega^2 = k^t \mu k$ and the group speed is

$$v_g = \frac{\mu k}{\omega} = \frac{\mu k}{\sqrt{k^t \mu k}}$$

Evidently in the pictures $\omega$ was kept constant and the length of $k$ was adjusted according to the dispersion relation. Comparing to the corresponding tables, it is seen that the maximum error occurs when $k$ is the eigenvector with the minimum eigenvalue, which seems rather natural. In all pictures we have taken $N = 8$ with respect to the smallest possible wave speed. Then comparing to the scalar case we see that errors are of the same magnitude. Note, however, that the variation of the error is typically more rapid than in the isotropic case.

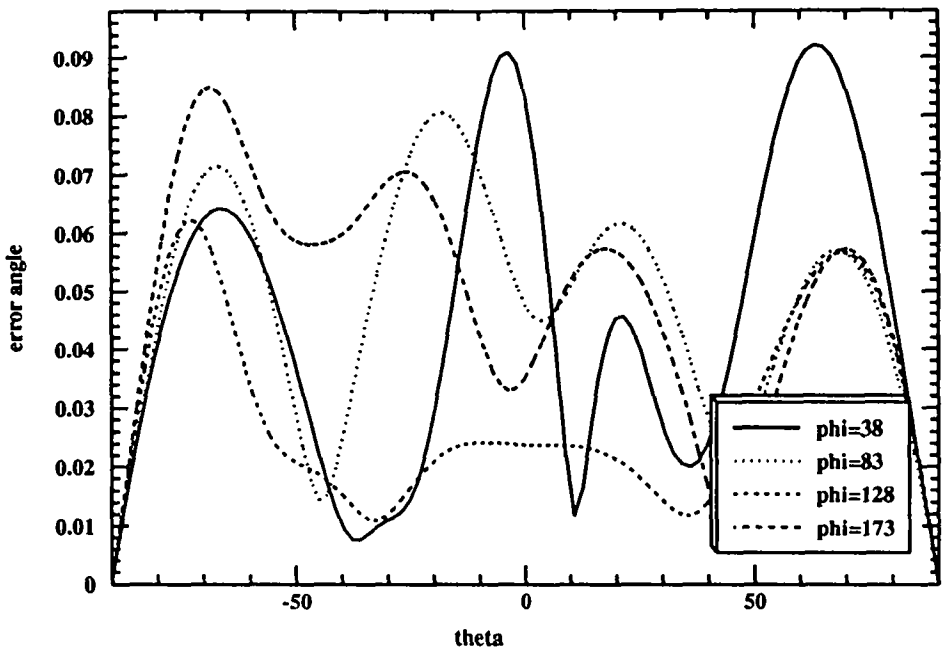



Figure 4.7: Error in the direction of the group speed; first test case.

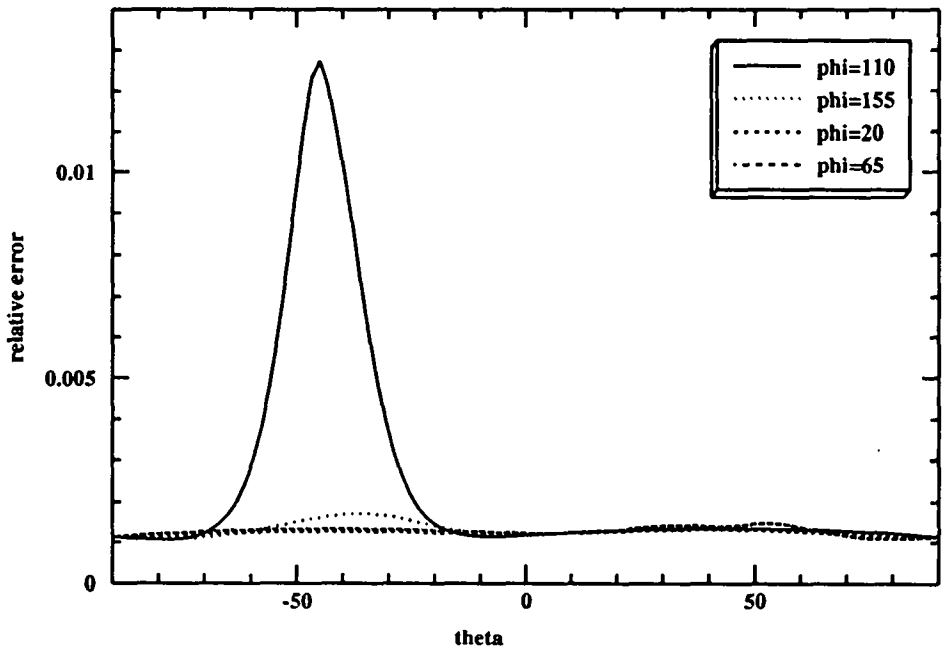



Figure 4.8: Relative error of the group speed; second test case.

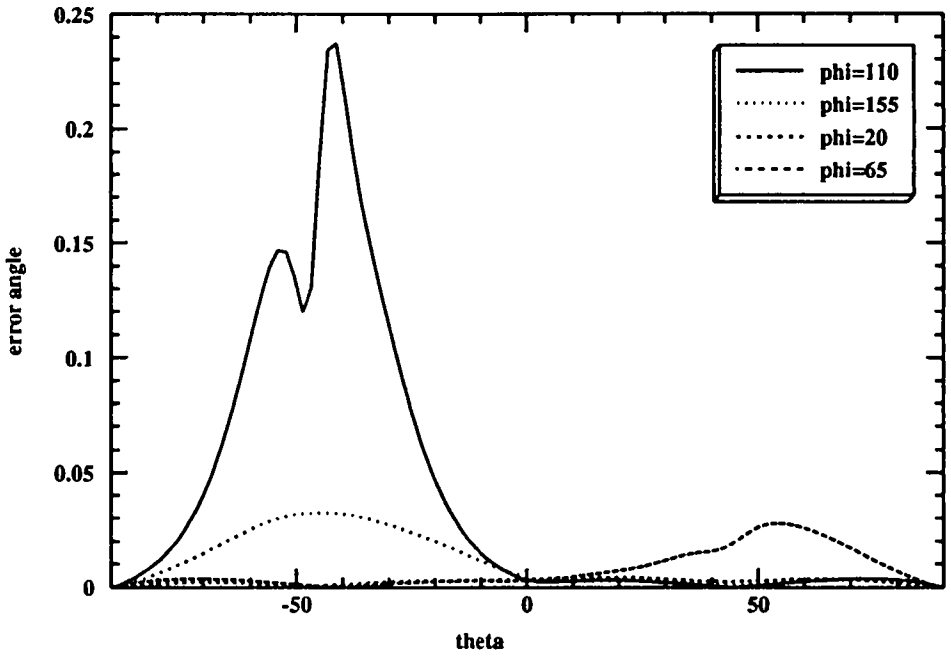



Figure 4.9: Error in the direction of the group speed; second test case.

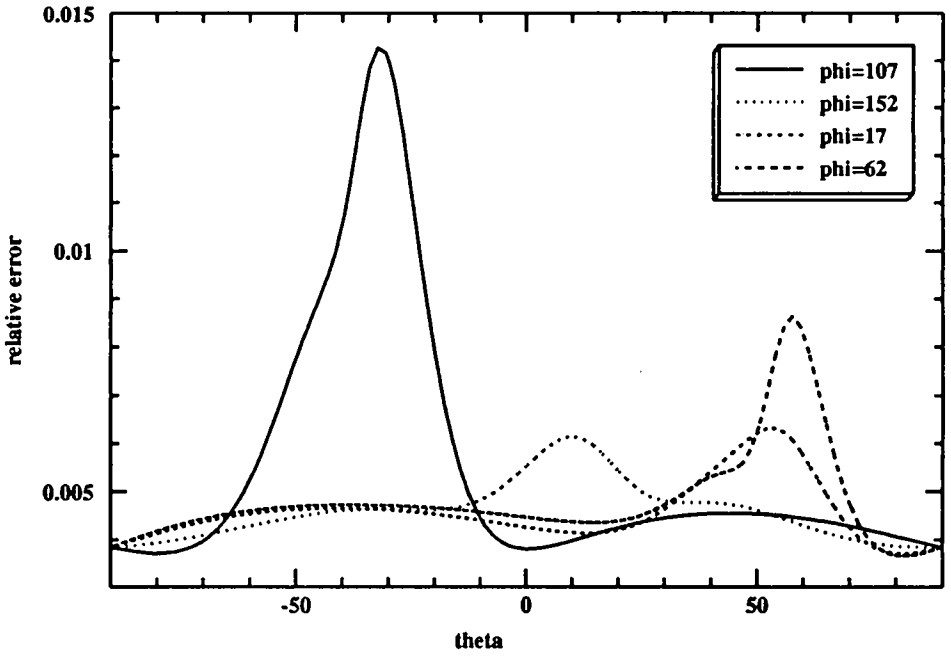



Figure 4.10: Relative error of the group speed; third test case.

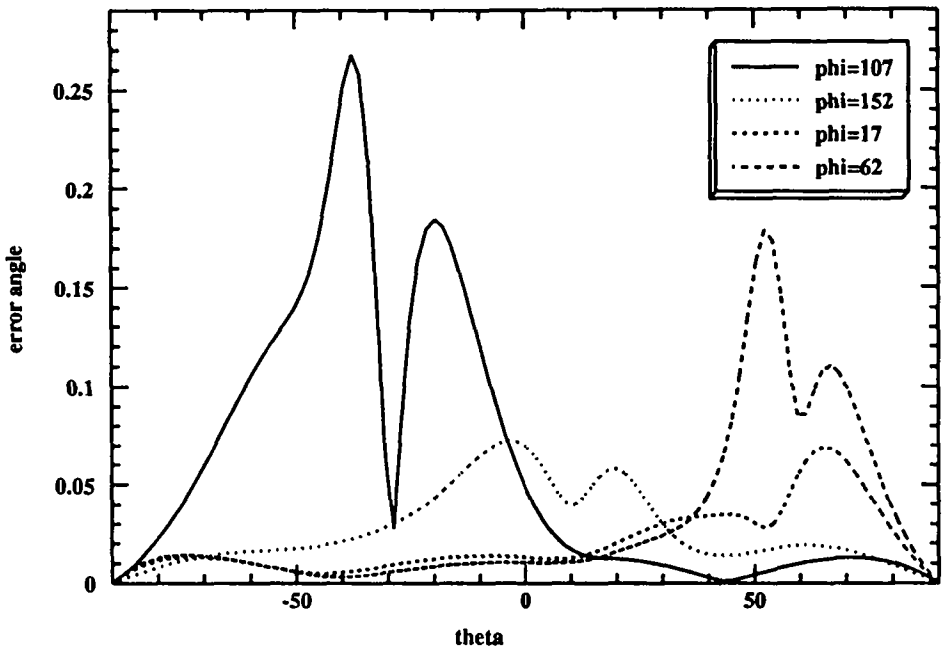



Figure 4.11: Error in the direction of the group speed; third test case.

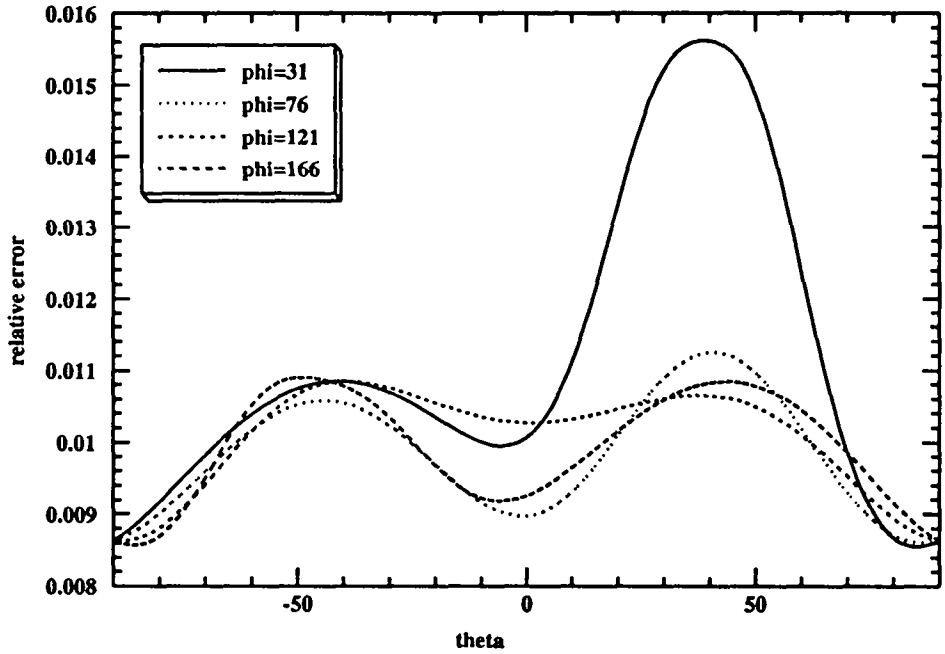



Figure 4.12: Relative error of the group speed; fourth test case.

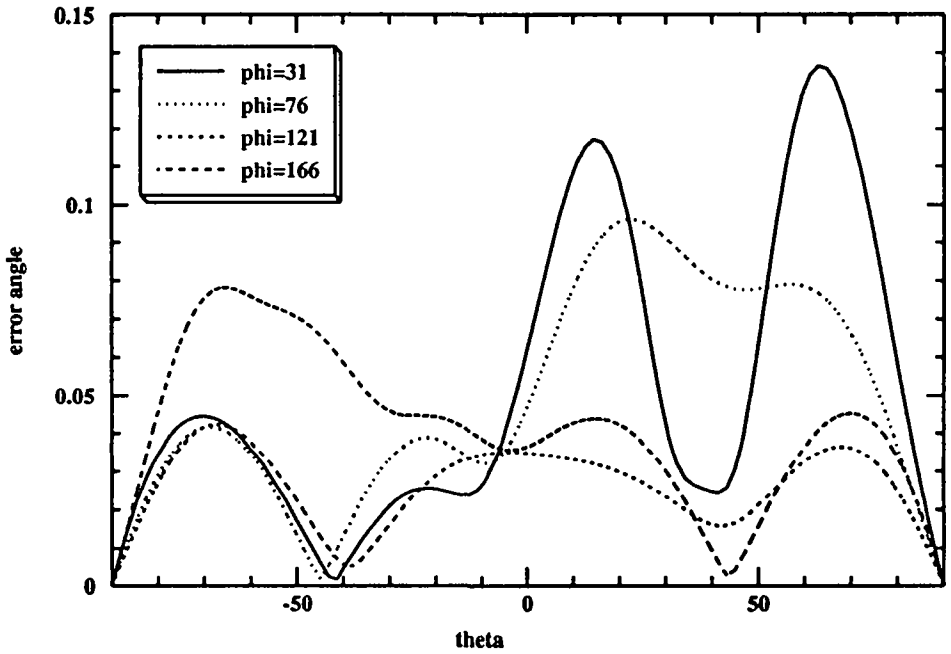



Figure 4.13: Error in the direction of the group speed; fourth test case.

# 5 Maxwell Equations

## 5.1 First Order System

We will rapidly indicate how to use the preceding ideas to construct fourth order schemes to Maxwell equations. Here we will in fact explicitly use the intermediate points: because of the structure of the Maxwell equations it is possible to consider that the (discrete) electric and magnetic fields are defined in different points. It is not so unnatural as it sounds at first; in fact (see [BO]) when interpreting electric and magnetic fields as differential forms instead of vector fields it is natural to take one to be 1-form and the other 2-form. So when discretizing the forms, there is a priori no reason to associate the degrees of freedom of 1-forms and 2-forms at the same points. Anyway, considering first the homogeneous and isotropic case we write the Maxwell equations as

$$\begin{aligned} D_t - \nabla \times B &= 0 \\ B_t + \nabla \times D &= 0 \end{aligned}$$

Using (3.2) we can discretize $\nabla\times$ by

$$\mathcal{T}_\times = \begin{pmatrix} 0 & -\mathcal{S}_z & \mathcal{S}_y \\ \mathcal{S}_z & 0 & -\mathcal{S}_x \\ -\mathcal{S}_y & \mathcal{S}_x & 0 \end{pmatrix} \tag{5.1}$$

Taylor's expansion then shows that

$$\frac{1}{h}\mathcal{T}_\times B = \nabla \times B + \frac{h^2}{8}\Delta\nabla \times B + O(h^4) = \nabla \times B - \frac{h^2}{8}(\nabla\times)^3 B + O(h^4)$$

To treat the time discretization it is convenient to take the other field at instants $n\,\delta t$ and the other at $(n+1/2)\delta t$. This gives

$$\begin{aligned} \frac{B^{n+1/2} - B^{n-1/2}}{\delta t} &= B_t + \frac{\delta t^2}{24} B_{ttt} + O(\delta t^4) = B_t + \frac{\delta t^2}{24}(\nabla\times)^3 D + O(\delta t^4) \\ \frac{D^{n+1} - D^n}{\delta t} &= D_t + \frac{\delta t^2}{24} D_{ttt} + O(\delta t^4) = D_t - \frac{\delta t^2}{24}(\nabla\times)^3 B + O(\delta t^4) \end{aligned}$$

Of course in the latter equation we approximate the derivative of $D$ at instant $n+1/2$. So we can write down the fourth order scheme as follows

$$\begin{aligned} \frac{D^{n+1} - D^n}{\delta t} - \frac{1}{h}\left(\mathcal{T}_\times B^{n+1/2} + \frac{3-\alpha^2}{24}(\mathcal{T}_\times)^3 B^{n+1/2}\right) &= 0 \\ \frac{B^{n+1/2} - B^{n-1/2}}{\delta t} + \frac{1}{h}\left(\mathcal{T}_\times D^n + \frac{3-\alpha^2}{24}(\mathcal{T}_\times)^3 D^n\right) &= 0 \end{aligned}$$

where $\alpha = \delta t/h$ as usual. Then we have the following result.

**Theorem 2** *The stability condition of the above scheme is*

$$\alpha \leq 1.39$$

**Proof** We use the Fourier analysis which now leads to eigenvalue problems because we have vector valued functions. Let us first introduce some notations. We look for the plane wave solutions $B = B_0 \exp(i(k \cdot x - \omega t))$ and $D = D_0 \exp(i(k \cdot x - \omega t))$, where $B_0$ and $D_0$ are some constant vectors. Let us denote the vector $(D_0, B_0)$ by $Z \in \mathbb{R}^6$. Then we define the (discrete Maxwell) operator $\mathcal{M}$ as

$$\mathcal{M} = \begin{pmatrix} 0 & -\mathcal{T}_\times \\ \mathcal{T}_\times & 0 \end{pmatrix}$$

Then using the Fourier transforms of operators as in (3.5) we get

$$\hat{\mathcal{M}} = i\,M = \begin{pmatrix} 0 & -\hat{\mathcal{T}}_\times \\ \hat{\mathcal{T}}_\times & 0 \end{pmatrix} = i \begin{pmatrix} 0 & -T_\times \\ T_\times & 0 \end{pmatrix} \tag{5.2}$$

where $T$ is a $3 \times 3$ real antisymmetric matrix and consequently $M$ is a $6 \times 6$ real symmetric matrix. Evidently we have $\hat{\mathcal{T}}_\times^3 = -i\,T_\times^3$. Then finally the eigenvalue problem takes the form

$$2\sin(\omega\delta t/2)Z = \alpha(M + \frac{3-\alpha^2}{24}M^3)Z$$

Of course it is sufficient to analyse the eigenvalues of $M$. Now $T$ is antisymmetric so that its eigenvalues are purely imaginary; then elementary considerations show that if $i\lambda$ is an eigenvalue of $T$ (where $\lambda \in \mathbb{R}$) then $\lambda$ is an eigenvalue of $M$. Now the eigenvalues of $T$ are $i\lambda$, $-i\lambda$ and zero, so that $M$ has three double eigenvalues: $\pm\lambda$ and zero. Then doing the calculations we find that we have already computed the eigenvalues of $T$ or more precisely we have calculated $\lambda^2$: taking the function $F$ in (3.6) we have $F = \lambda^2$ which then leads to the final dispersion relation

$$4\sin^2(\omega\delta t/2) = \alpha^2 F(1 + \frac{3-\alpha^2}{24}F)^2 \tag{5.3}$$

Consequently the stability condition is

$$\alpha^2 F(1 + \frac{3-\alpha^2}{24}F)^2 \leq 4$$

The maximum of $F$ was found to be $16/9$; to analyse the situation let us denote $\alpha^2$ by $x$ and $F$ by $y$ and consider the following function.

$$g(x, y) = xy(1 + (3 - x)y/24)^2$$

where $x \geq 0$ and $0 \leq y \leq 16/9$. To check the inequality we consider the following problem

$$\begin{aligned} &\max g(x,y) \\ &0 \leq x \leq x^* \\ &0 \leq y \leq 16/9 \end{aligned}$$

where $x^*$ is arbitrary. In addition to the trivial cases $g(0,y) = g(x,0) = 0$ we have three different possibilities.

**Case 1** In the interior of the domain we find that the gradient of $g$ can vanish only when $x = 3 + 24/y$ which corresponds to the value zero of $g$.

**Case 2** Put $y = 16/9$. Then there is a local maximum at $x = 11/6$ and the value of $g$ is $42592/6561 \simeq 6.49 > 4$. Then solving $g(x, 16/9) = 4$ gives

$$x = 11 - \frac{(1+i\sqrt{3})121 \cdot 2^{-4/3}}{(1237+81i\sqrt{4087})^{1/3}} - (1-i\sqrt{3})2^{-8/3}(1237+81i\sqrt{4087})^{1/3} \simeq 1.93$$

The above solution is really exactly a real number, in spite of appearances. The other two solutions are also real, but they are bigger than the above solution, so they can be ignored.

**Case 3** Put $x^* = 1.93$. Then we find that $g$ is monotonically increasing function of $y$ so that the maximum is at $(x^*, 16/9)$.

Now taking the square root of the above value of $x$ gives the final stability condition.

■

Note that the stability condition is almost the same as in (2.3). This is not so surprising when we notice that expanding (5.3) we get

$$4\sin^2(\omega\delta t/2) = \alpha^2 F(1 + \frac{3-\alpha^2}{12}F + \frac{(3-\alpha^2)^2}{576}F^2)$$

which is the same as for the wave equation except the last term which we expect to be 'small'. We will not treat the variable coefficients in case of the full system of Maxwell equations. Instead we proceed to the case where only electric or magnetic field is required.

## 5.2 Second Order System

Often it is not necessary to solve the whole system of Maxwell equations: sometimes it is sufficient to calculate only the electric or magnetic field. When this is the case one can proceed in the following way. Starting then from the original equations

$$\begin{aligned} D_t - \nabla \times H &= 0 \\ B_t + \nabla \times E &= 0 \end{aligned}$$

and taking the time derivative of the first equation and then using the second equation and the constitutive relations $B = \mu H$ and $D = \varepsilon E$ we get

$$\varepsilon E_{tt} + \nabla \times \frac{1}{\mu} \nabla \times E = 0 \tag{5.4}$$

We suppose that $\mu$ and $\varepsilon$ are scalar positive functions which are independent of time. The speed of propagation and the impedance are then given by

$$\begin{aligned} c &= 1/\sqrt{\varepsilon\mu} \\ z &= \sqrt{\varepsilon/\mu} \end{aligned}$$

Note that $\mu$ (permeability) here is different from $\mu$ in the previous sections and also different from the $\mu$ (one of the Lamé coefficients) of the next section!

Now the above equation resembles very much (3.1) and indeed it can be treated in the same spirit. We recall that $\nabla\times$ is a symmetric operator, so that we have the variational formulation

$$(\varepsilon E_{tt}, v) + (\frac{1}{\mu} \nabla \times E, \nabla \times v) = 0 \quad \forall v \in H(\nabla\times)$$

where $H(\nabla\times)$ is the space of $L^2$ functions $v$ such that $\nabla \times v$ is also in $L^2$. Now to discretize (5.4) we cannot directly use (5.1) (because of half steps), so we define $\mathcal{T}_x^+$ and $\mathcal{T}_x^-$ by

$$\mathcal{T}_x^{\pm} = \begin{pmatrix} 0 & -\mathcal{S}_z^{\pm} & \mathcal{S}_y^{\pm} \\ \mathcal{S}_z^{\pm} & 0 & -\mathcal{S}_x^{\pm} \\ -\mathcal{S}_y^{\pm} & \mathcal{S}_x^{\pm} & 0 \end{pmatrix}$$

Then it is straightforward to verify that

**Proposition 12**

$$\begin{aligned} \mathcal{T}_x^- \mathcal{T}_x^+ &= \mathcal{T}_x^2 \\ (\mathcal{T}_x^- u, v) &= (u, \mathcal{T}_x^+ v) \end{aligned}$$

*where in the latter equation the inner product is taken in* $l^2(\mathbb{Z}^3)^3$.

Then defining operators

$$\begin{aligned} \mathcal{M}u &= \{(1/\mu_{i+1,j,k} + 1/\mu_{i,j,k} + 1/\mu_{i+1,j+1,k} + 1/\mu_{i,j+1,k} + \\ &\quad 1/\mu_{i+1,j,k+1} + 1/\mu_{i,j,k+1} + 1/\mu_{i+1,j+1,k+1} + 1/\mu_{i,j+1,k+1})u_{ijk}/8\} \\ \mathcal{R}u &= \{\varepsilon_{ijk} u_{ijk}\} \\ \mathcal{W} &= \mathcal{T}_x^- \mathcal{M} \mathcal{T}_x^+ \end{aligned}$$

we can discretize the equation (5.4) as follows

$$\mathcal{R}(E^{n+1} - 2E^n + E^{n-1}) + \alpha^2 (\mathcal{W} + \frac{3 - c^2\alpha^2}{12c^2} \mathcal{W}\mathcal{R}^{-1}\mathcal{W}) E^n = 0 \tag{5.5}$$

Then we have the following result.

**Theorem 3** *The stability condition for the scheme (5.5) is given by (2.11).*

Obviously in applying (2.11) to the present situation one must identify $\varepsilon$ and $\rho$, and $\mu$ and $1/\mu$ (latter being unfortunate notational coincidence...).

**Proof** Because (5.5) is of the same form as (2.8), all we have to do is to calculate the norm of $\mathcal{W}$. First we evidently get

$$(\mathcal{W}E, E) \leq \frac{1}{\mu_*}(\mathcal{T}_x^+ E, \mathcal{T}_x^+ E)$$

Next one has to calculate the maximum value of the scalar product in the right hand side. We use the Fourier transform as usual, and then applying the preceding proposition and recalling the matrices in (5.2), it is seen that this is equivalent to finding the maximum eigenvalue of $-T_x^2$. But we have already noted that the expression for the (non zero) eigenvalues of $-T_x^2$ is the same $F$ that appears in (3.6), so that the bound for $\|\mathcal{W}\|$ is the same as the bound for the corresponding operator in case of wave equation, and consequently the stability properties must be the same.

■

## 5.3 Dispersion and Polarization

Finally let us look at some dispersion curves. The above theorem (or rather its proof) shows that there is no need to draw the dispersion curves for the second order system: they are exactly the same as in case of the scalar wave equation. However, the first order system is different (except that the error in the direction of the group speed remains the same) and so we show some pictures concerning this case. In addition, there appears a new phenomenon with vector valued functions: polarization. Consider the following eigenproblem. Try to find $V \in \mathbb{R}^3$, $\lambda$ and $k$ such that

$$\nabla \times V e^{ik\cdot x} = i\lambda V e^{ik\cdot x}$$

It is easy to see that $V \cdot k = 0$ must hold. Exactly in the same way in the Maxwell equations the electric and the magnetic fields are orthogonal to the direction of propagation $k$. The plane orthogonal to $k$ is then called the plane of polarization. Now the discretization will change this plane and we would like to measure the effect in some way.

Let us define the vector $S = -i\left(\begin{array}{ccc} \hat{\mathcal{S}}_x & \hat{\mathcal{S}}_y & \hat{\mathcal{S}}_z \end{array}\right)^t$ (see (3.5) and note that $S \in \mathbb{R}^3$). Then considering the problem

$$\mathcal{T}_x V e^{ik\cdot x} = i\lambda V e^{ik\cdot x}$$

we find that $V \cdot S = 0$. So a natural way to estimate the change in polarization is to use the angle $\gamma$ which is defined by

$$\cos\gamma = \frac{S \cdot k}{|S|\,|k|} \tag{5.6}$$

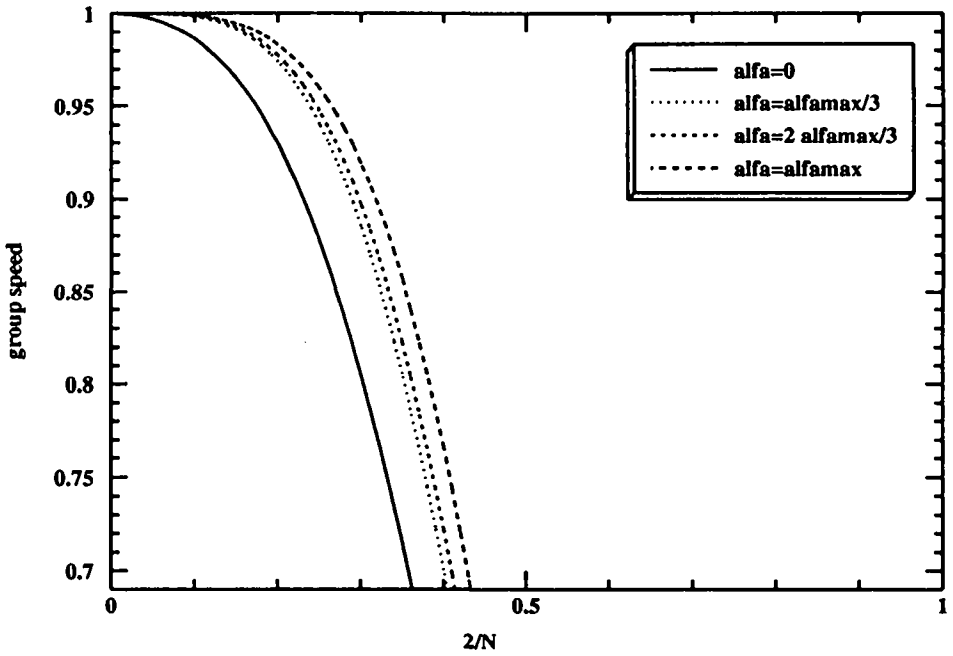



Figure 5.1: Group speed in the direction $\varphi = 20$, $\theta = 30$.

Note that $S$ is a symbol of a second order operator; however, the expansion shows that

$$\gamma = O(h^4)$$

This is a same kind of result as (4.2), and the explanation is also the same: the operator has some nice symmetries which make it possible to have more accuracy than expected, see [TU].

In figure 5.1 we show the group speed as a function of $2/N$ for different $\alpha$; other values for $\theta$ and $\varphi$ giving essentially similar pictures. Next in figure 5.2 the variation of the group speed is shown as a function of the direction, using the optimal $\alpha$ and $N = 5$. Figure 5.3 is similar but with $N = 9$. Comparing to the scalar wave equation we see that the error is a bit smaller, but essentially the 'same'. Finally in figure 5.4 and 5.5 there are the polarization error first as a function of $N$ and then in different directions with $N = 5$. The error is seen to be quite small. Note that the angle does not 'explode' when $N \to 2$, as the error angle of the group speed. This is because this error depends only on the first order difference operators.

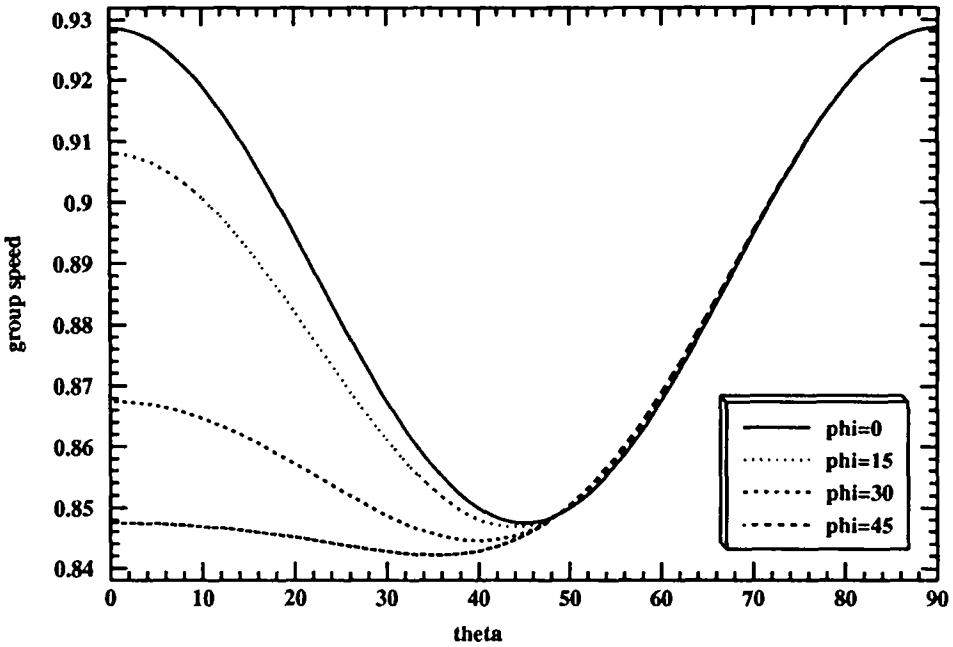



Figure 5.2: Group speed with $N = 5$.

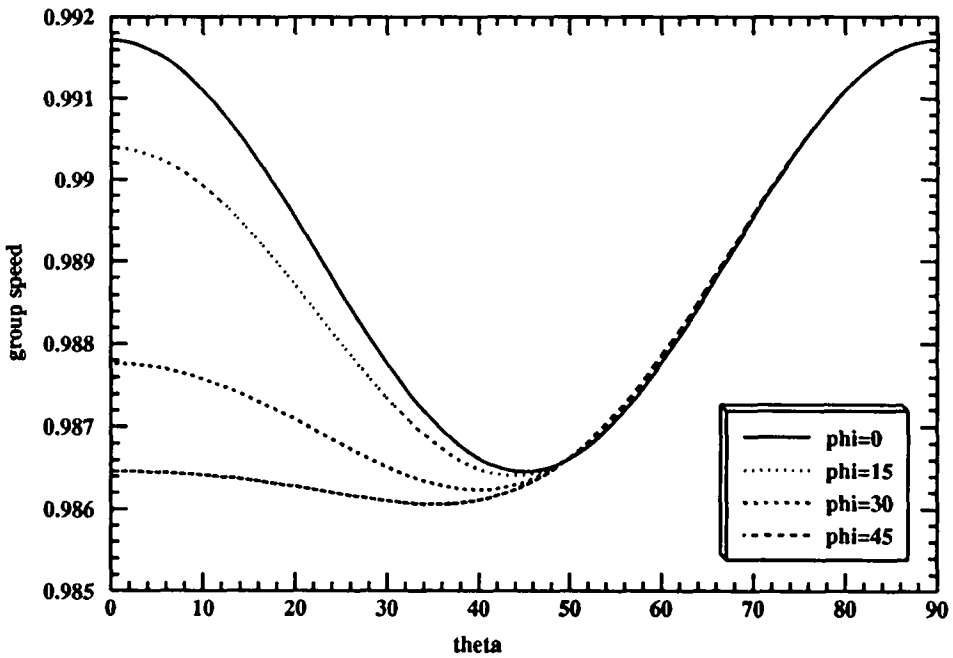



Figure 5.3: Group speed with $N = 9$.

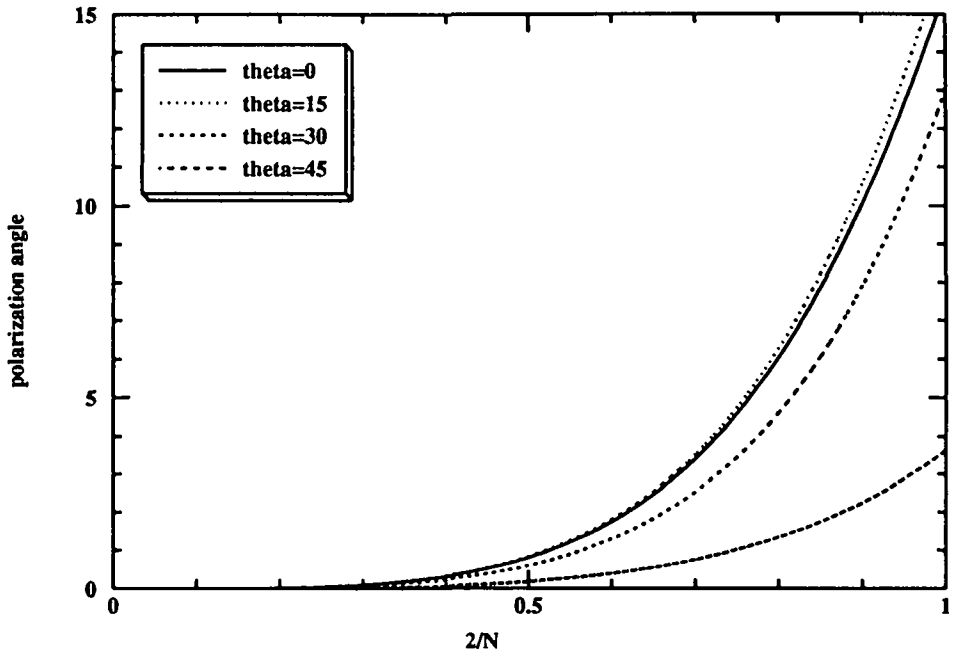



Figure 5.4: Polarization error with $\varphi = 20$.

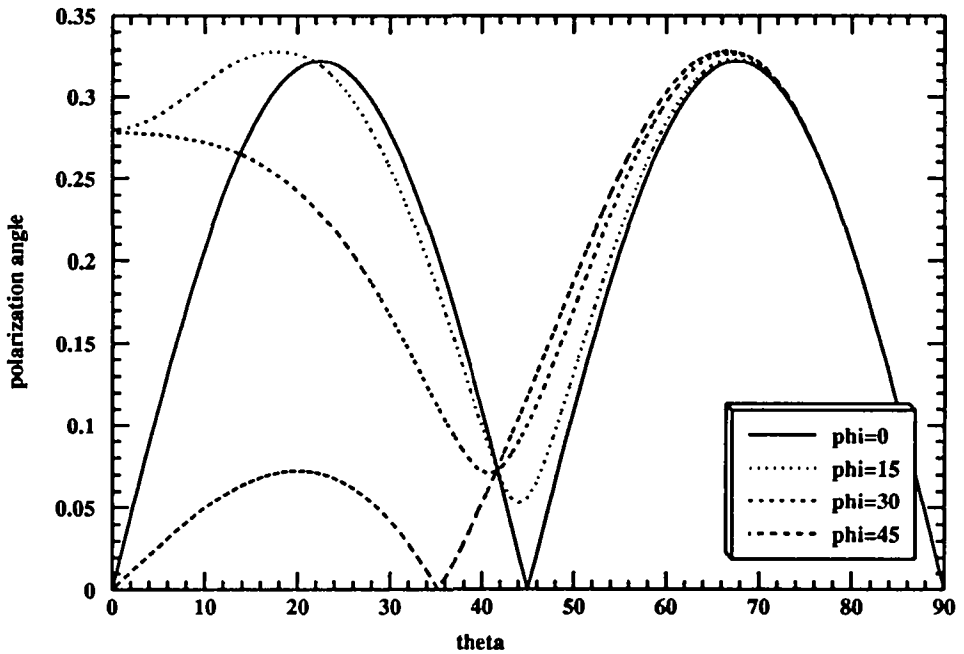



Figure 5.5: Polarization error with $N = 5$.

# 6 Linearized Elastodynamic Equations

## 6.1 Constant Coefficients

Finally we apply our scheme to linearized elastodynamic equations. We start with the simplest case, that is isotropic and homogeneous material. In this case our basic equation can be written in three different forms.

$$\begin{aligned} u_{tt} - (\lambda+2\mu)\Delta u - (\lambda+\mu)\nabla\times\nabla\times u &= 0 \\ u_{tt} - (\lambda+2\mu)\nabla\nabla\cdot u + \mu\nabla\times\nabla\times u &= 0 \\ u_{tt} - \mu\Delta u - (\lambda+\mu)\nabla\nabla\cdot u &= 0 \end{aligned}$$

where $\lambda$ and $\mu$ are the coefficients of Lamé and $u$ is a vector valued function. We use (3.2) as our basic operator (as well as associated operators $\mathcal{S}^+$ and $\mathcal{S}^-$). It does not matter which form we choose, because if we 'expand' the discretizations in terms of $\mathcal{S}$'s the resulting scheme will be exactly the same in the three cases.

Let us take the first equation; recall that using the operators $\mathcal{T}$ defined in (3.3) and $\mathcal{T}_\times$ defined in (5.1) we have

$$\begin{aligned} \frac{1}{h^2}\mathcal{T}u &= -\Delta u - \frac{h^2}{4}\Delta^2 u + O(h^4) \\ \frac{1}{h^2}{\mathcal{T}_\times}^2 u &= \nabla\times\nabla\times u - \frac{h^2}{4}(\nabla\times)^4 u + O(h^4) \end{aligned}$$

Note that because $u$ is vector valued $\mathcal{T}$ is here interpreted as a diagonal operator, in the same way as $\Delta u = (\Delta u_1, \Delta u_2, \Delta u_3)$. Then analysing the error in time gives

$$\begin{aligned} \frac{u^{n+1} - 2u^n + u^{n-1}}{\delta t^2} &= u_{tt} + \frac{\delta t^2}{12}u_{tttt} + O(\delta t^4) = u_{tt} + \\ &\frac{\delta t^2}{12}((\lambda+2\mu)^2\Delta^2 u - (\lambda+\mu)(\lambda+3\mu)(\nabla\times)^4 u) + O(\delta t^4) \end{aligned}$$

Let us introduce the following notation: $\alpha_1 = \sqrt{\lambda+2\mu}\,\delta t/h$, $b_1 = (\lambda+\mu)/(\lambda+2\mu)$ and $b_2 = (\lambda+3\mu)/(\lambda+2\mu)$. Then the fourth order scheme can be conveniently written as

$$\begin{aligned} &u^{n+1} - 2u^n + u^{n-1} + \alpha_1^2(\mathcal{T}u^n - b_1\mathcal{T}_\times^2 u^n + \\ &\frac{3-\alpha_1^2}{12}\mathcal{T}^2 u^n - \frac{b_1}{12}(3 - b_2\alpha_1^2)\mathcal{T}_\times^4 u^n) = 0 \end{aligned} \tag{6.1}$$

Then we look for the plane wave solutions $u = u_0\exp(i(k\cdot x - \omega t))$ as usual which leads to the following eigenvalue problem

$$4\sin^2(\omega\delta t/2)u_0 = \alpha_1^2\left((F + \frac{(3-\alpha_1^2)}{12}F^2)I + b_1(T_\times^2 - \frac{(3-b_2\alpha_1^2)}{12}T_\times^4)\right)u_0 \tag{6.2}$$

where $F$ is as in (5.3). But we have already seen that eigenvalues of $T_x^2$ are zero and double eigenvalue $-F$. The zero eigenvalue corresponds to the pressure waves and their dispersion relation is then exactly the same as for the wave equation. The other eigenvalue is more interesting; the corresponding solutions are called the shear waves. Denoting $\sqrt{\mu}\delta t/h$ by $\alpha_2$ we get

**Proposition 13** *The dispersion relation for the shear waves is*

$$4\sin^2(\omega\delta t/2) = \alpha_2^2\left(F + \frac{(3-\alpha_2^2)}{12}F^2\right) \tag{6.3}$$

**Proof** It is sufficient to substitute $-F$ for $T_x^2$ and $F^2$ for $T_x^4$ in (6.2) and simplify the expression.

■

The above result is interesting for the following reason: (6.3) does not depend on $\lambda$ so that the pressure waves cannot 'disturb' the shear waves. This is not automatic; for instance for some methods the accuracy is poor when $\lambda/\mu$ is large. Moreover, putting $\mu = 0$ makes the right hand side of (6.3) (and consequently the phase speed) exactly zero (and not just 'small'), which is another way to say that $\lambda$ does not interfere. Another advantage of the scheme is that by construction $T_x$ has eigenvalues $\pm i\sqrt{F}$ and zero, so that the operator on right hand side of (6.2) has an exact double eigenvalue (and not two eigenvalues close to each other). As regards the polarization we obtain

**Proposition 14** *The eigenvectors associated to the shear waves are exactly orthogonal to the eigenvector associated to the pressure wave.*

**Proof** This is clear from (6.2).

■

For completeness let us state

**Proposition 15** *The stability condition for the scheme (6.1) is*

$$\alpha_1 \leq \sqrt{\frac{39-3\sqrt{61}}{8}} \simeq 1.39$$

**Proof** It is sufficient to recall (2.3) which is valid also in three dimensions and note that $\alpha_2 \leq \alpha_1$

■

So in the constant coefficient case the stability condition is the same as for the wave equation

$$u_{tt} - (\lambda + 2\mu)\Delta u = 0$$

## 6.2 Variable Coefficients

Then how should we implement the scheme in the general case? In fact it is rather easy: even when the coefficients are not constant we can start from any of three forms. However, these are not anymore completely equivalent. To play it safe, let us take the second equation whose variational formulation gives

$$(\rho u_{tt}, v) + ((\lambda + 2\mu)\nabla \cdot u, \nabla \cdot v) + (\mu \nabla \times u, \nabla \times v) = 0$$

where for the sake of generality we have added $\rho$. Note that the second and the third term define positive bilinear forms and consequently one can associate to them positive definite operators in suitable Sobolev spaces. Recall that we can define discrete divergence $\mathcal{D}$ and gradient $\mathcal{G}$ operators as follows

$$\begin{aligned} \mathcal{G} &= \begin{pmatrix} \mathcal{S}_x^- & \mathcal{S}_y^- & \mathcal{S}_z^- \end{pmatrix}^t \\ \mathcal{D} &= \begin{pmatrix} \mathcal{S}_x^+ & \mathcal{S}_y^+ & \mathcal{S}_z^+ \end{pmatrix} \end{aligned}$$

where $\mathcal{S}_x^+$ etc are as in (3.3). In addition let us define the following operators.

$$\begin{aligned} \mathcal{L}_1 u &= \{(\lambda_{i+1,j,k} + \lambda_{i,j,k} + \lambda_{i+1,j+1,k} + \lambda_{i,j+1,k} + \\ &\quad \lambda_{i+1,j,k+1} + \lambda_{i,j,k+1} + \lambda_{i+1,j+1,k+1} + \lambda_{i,j+1,k+1} + \\ &\quad 2\mu_{i+1,j,k} + 2\mu_{i,j,k} + 2\mu_{i+1,j+1,k} + 2\mu_{i,j+1,k} + \\ &\quad 2\mu_{i+1,j,k+1} + 2\mu_{i,j,k+1} + 2\mu_{i+1,j+1,k+1} + 2\mu_{i,j+1,k+1})u_{ijk}/8\} \\ \mathcal{L}_2 u &= \{(\mu_{i+1,j,k} + \mu_{i,j,k} + \mu_{i+1,j+1,k} + \mu_{i,j+1,k} + \\ &\quad \mu_{i+1,j,k+1} + \mu_{i,j,k+1} + \mu_{i+1,j+1,k+1} + \mu_{i,j+1,k+1})u_{ijk}/8\} \\ \mathcal{R} u &= \{\rho_{ijk} u_{ijk}\} \\ \mathcal{T}_1 &= -\mathcal{G}\mathcal{L}_1\mathcal{D} \\ \mathcal{T}_2 &= \mathcal{T}_x^- \mathcal{L}_2 \mathcal{T}_x^+ \\ \mathcal{W}_1 &= \mathcal{T}_1 \mathcal{R}^{-1} \mathcal{T}_1 \\ \mathcal{W}_2 &= \mathcal{T}_2 \mathcal{R}^{-1} \mathcal{T}_2 \\ \mathcal{W}_3 &= (\mathcal{T}_1 + \mathcal{T}_2)\mathcal{R}^{-1}(\mathcal{T}_1 + \mathcal{T}_2) \end{aligned}$$

Further let us denote the speed of the pressure waves by $c_1^2 = (\lambda + 2\mu)/\rho$ and the speed of the shear waves by $c_2^2 = \mu/\rho$. Then the scheme can be written as

$$\mathcal{R}(u^{n+1} - 2u^n + u^{n-1}) + \alpha^2 \Big(\mathcal{T}_1 + \mathcal{T}_2 + \frac{1}{12c_1^2c_2^2}(3c_2^2\mathcal{W}_1 + 3c_1^2\mathcal{W}_2 - \alpha^2 c_1^2 c_2^2 \mathcal{W}_3)\Big)u^n = 0 \quad (6.4)$$

Then we have

**Proposition 16** *The operators $\mathcal{T}_i$ and $\mathcal{W}_i$ are positive and symmetric. The correction operator*

$$\mathcal{C} = 3c_2^2\mathcal{W}_1 + 3c_1^2\mathcal{W}_2 - \alpha^2 c_1^2 c_2^2 \mathcal{W}_3$$

*is positive when $(c_1^2 + c_2^2)\alpha^2 \leq 3$.*

**Proof** The positivity and the symmetry of operators $\mathcal{T}_i$ and $\mathcal{W}_i$ follow from the fact that $-\mathcal{G}$ is the transpose of $\mathcal{D}$ and that $\mathcal{T}_x^-$ is the transpose of $\mathcal{T}_x^+$.
To prove the second assertion we first obtain

$$(\mathcal{C}u, u) = 3c_2^2(\mathcal{R}^{-1}\mathcal{T}_1 u, \mathcal{T}_1 u) + 3c_1^2(\mathcal{R}^{-1}\mathcal{T}_2 u, \mathcal{T}_2 u) - \alpha^2 c_1^2 c_2^2 (\mathcal{R}^{-1}(\mathcal{T}_1 + \mathcal{T}_2)u, (\mathcal{T}_1 + \mathcal{T}_2)u)$$

Evidently this is positive for $\alpha$ sufficiently small. Now a sum is positive if all of its terms are positive, so using the notations

$$\begin{aligned} v_1 &= \{\mathcal{T}_1 u\}_{ijk} \\ v_2 &= \{\mathcal{T}_2 u\}_{ijk} \end{aligned}$$

we see that the sufficient condition for positivity is

$$3c_2^2|v_1|^2 + 3c_1^2|v_2|^2 - \alpha^2 c_1^2 c_2^2 |v_1 + v_2|^2 \geq 0$$

Then recalling the simple fact that

$$2\,|v_1 \cdot v_2| \leq 2\,|v_1|\,|v_2| \leq \eta|v_1|^2 + \frac{1}{\eta}|v_2|^2$$

for any $\eta > 0$, the condition $(c_1^2 + c_2^2)\alpha^2 \leq 3$ follows easily by taking $\eta = c_2^2/c_1^2$, which is the optimal choice.

■

Note that the positivity condition is less strict than the constant coefficient stability condition, which was $c_1\alpha \leq 1.39 < \sqrt{3}$. Intuitively we might say that while positivity depends on the 'local average' of the wave speeds the stability depends on the 'local maximum' of the wave speeds.

**Theorem 4** *A sufficient condition for the stability of the scheme (6.4) is*

$$4\alpha^2\Big(c_1^{*2}\Big(1 + \frac{4c_1^{*2}}{9c_{*1}^2}\Big) + c_2^{*2}\Big(1 + \frac{4c_2^{*2}}{9c_{*2}^2}\Big)\Big) \leq 9$$

*where lower (resp. upper) star indicates the minimum (resp. maximum) as usual.*

**Proof** Let us start by estimating the norm of $\mathcal{C}$. Supposing that the positivity condition is satisfied we obtain

$$(\mathcal{C}u, u) \leq \frac{3}{\rho_*}\Big(c_2^2(\mathcal{T}_1 u, \mathcal{T}_1 u) + c_1^2(\mathcal{T}_2 u, \mathcal{T}_2 u)\Big)$$

Now evidently we have

$$\begin{aligned} \|\mathcal{T}_1\| &\leq (\lambda^* + 2\mu^*)\,\|\mathcal{D}\|^2 \\ \|\mathcal{T}_2\| &\leq \mu^*\,\|\mathcal{T}_\times\|^2 \end{aligned}$$

This implies

$$\frac{1}{12c_1^2c_2^2}\,\|\mathcal{C}\| \leq \frac{1}{4}\Big(\frac{(\lambda^* + 2\mu^*)^2}{\rho_* c_{*1}^2}\,\|\mathcal{D}\|^4 + \frac{\mu^{*2}}{\rho_* c_{*2}^2}\,\|\mathcal{T}_\times\|^4\Big)$$

We have already seen that $\|\mathcal{T}_\times\|^2 = 16/9$. Next let us calculate $\|\mathcal{D}\|^2$. By Fourier transform we have

$$(\mathcal{D}u, \mathcal{D}u) = (SS^t\hat{u}, \hat{u})$$

where $S$ is as in (5.6). The eigenvalues of $SS^t$ are a double eigenvalue zero and $S \cdot S$. But $S \cdot S$ is just another notation for $F$ in (3.6) so we can conclude that $\|\mathcal{D}\|^2 = 16/9$. Then putting all these estimations together and using (2.4) we get

$$\frac{16\alpha^2}{9}\Big(c_1^{*2} + c_2^{*2} + \frac{4}{9}\Big(\frac{c_1^{*4}}{c_{*1}^2} + \frac{c_2^{*4}}{c_{*2}^2}\Big)\Big) \leq 4$$

Then simplifying a little we get the result.

■

Note that the above condition is far from optimal for the following reasons. First, various minima and maxima of $\lambda$, $\mu$ and $\rho$ are not necessarily attained at the same points. However, in any particular case it would be easy to modify the above proof to take into account the additional information on the parameters. Second, as can readily be verified by Fourier analysis, $(\mathcal{D}u, \mathcal{D}u)$ and $(\mathcal{T}_\times^+ u, \mathcal{T}_\times^+ u)$ do not attain their maxima with the same $u$. In fact when the other takes the maximum value the other is zero! Third, there is the problem of getting better estimates for

$$(\mathcal{T}_1 u, \mathcal{T}_2 u)$$

which is zero in the constant coefficient case because $\mathcal{T}_\times^+ \mathcal{G} = 0$ and $\mathcal{D}\mathcal{T}_\times^- = 0$. Here also we should make more precise assumptions of the coefficients to proceed. In any case we see that the condition depends on $c_1$ and $c_2$, and not only on $c_1$ as in the constant coefficient case.

## 6.3 Final Dispersion Curves

The same analysis of polarization carries over to the elastodynamic equations: instead of electric and magnetic fields we have two vectors corresponding to the shear waves, and in addition a vector corresponding to the pressure wave. It is sufficient to consider the plane of polarization as before, because for our method the pressure and shear

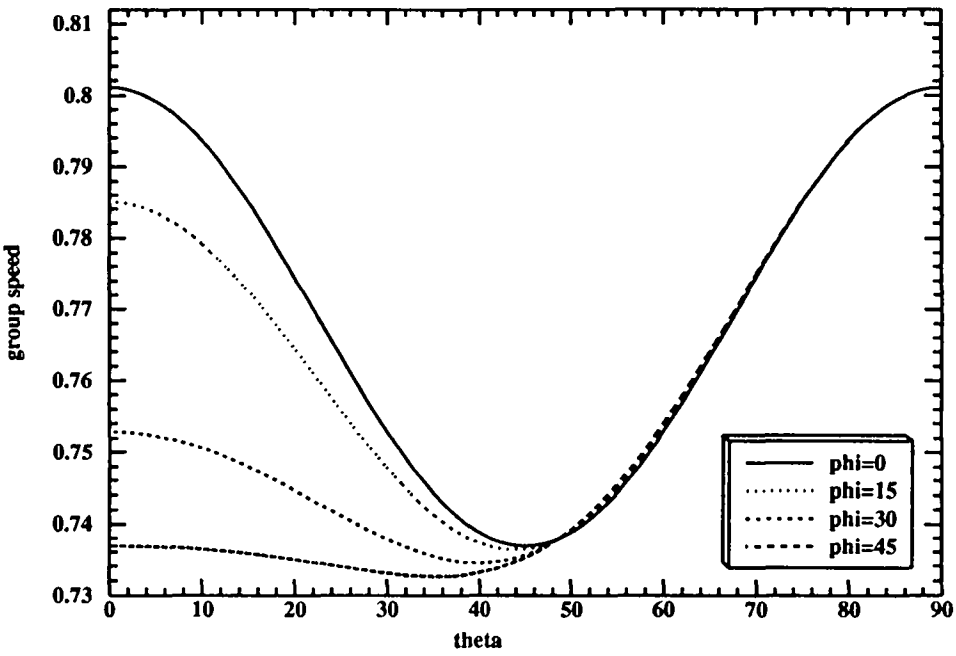



Figure 6.1: Group speed with $N = 5$ and $\lambda/\mu = 0$.

vectors are exactly orthogonal. Now the error is exactly the same as in case of the Maxwell equations, because in the constant coefficient case this depends only on the discretization of $\nabla\times$ which is the same in both cases.

So what remains? Only the effect of 'non optimal' $\alpha$. By this we mean the following: the wavelength $\lambda_s$ of the shear waves being smaller than that of the pressure waves we have to choose $h$ such that it is sufficiently small with respect to $\lambda_s$. On the other hand the stability condition is stricter for the pressure waves so that $\delta t$ and consequently $\alpha$ must be smaller (which makes the error bigger) than would be the case if there were only shear waves. Note, however, that this has no effect on the direction of the propagation. Now simple calculations show that

$$\alpha = \frac{\alpha_{opt}}{\sqrt{2 + \lambda/\mu}}$$

so that $\alpha \in (0, \alpha_{opt}/\sqrt{2}]$. In figures 6.1, 6.2, 6.3 and 6.4 we show the dispersion curves with $N = 5$ and $N = 9$ as in the scalar case, both cases with $\lambda/\mu = 0$ and $\lambda/\mu = 50$. The error is bigger than in the scalar case, but the difference is rather small. In particular to achieve the 2 % accuracy we need $N = 10.5$ when $\lambda/\mu = 0$ ('small') and $N = 11.5$ when $\lambda/\mu = 50$ ('close to infinity'), as shown in figures 6.5 and 6.6.

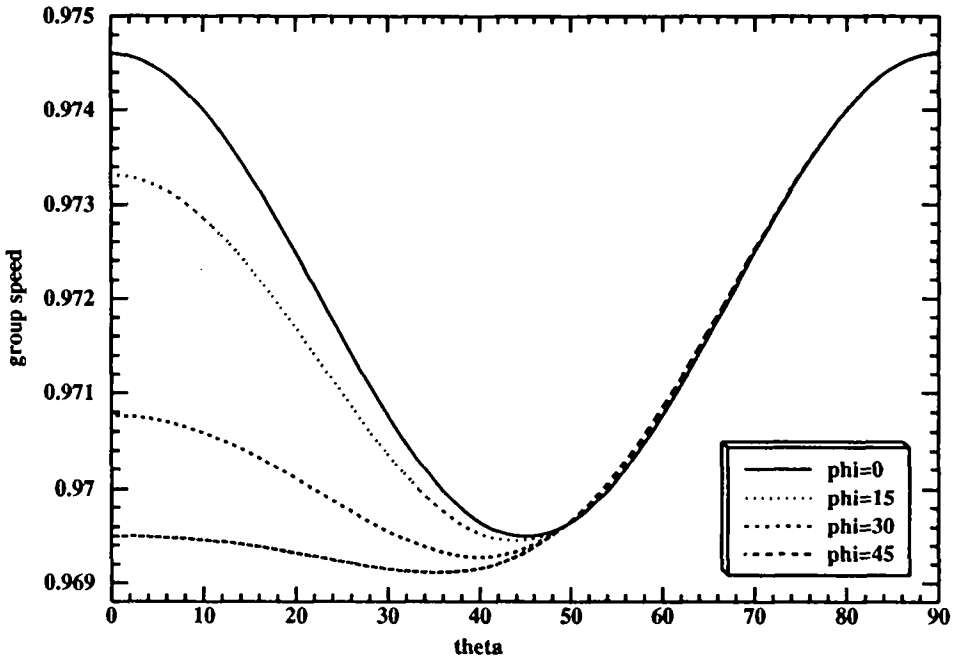



Figure 6.2: Group speed with $N = 9$ and $\lambda/\mu = 0$.

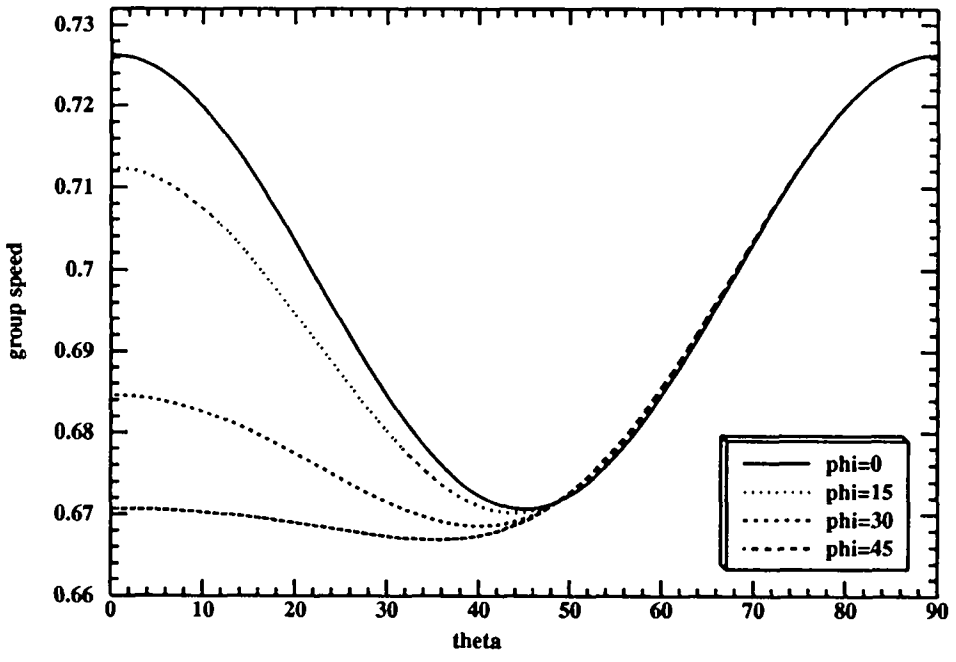



Figure 6.3: Group speed with $N = 5$ and $\lambda/\mu = 50$.

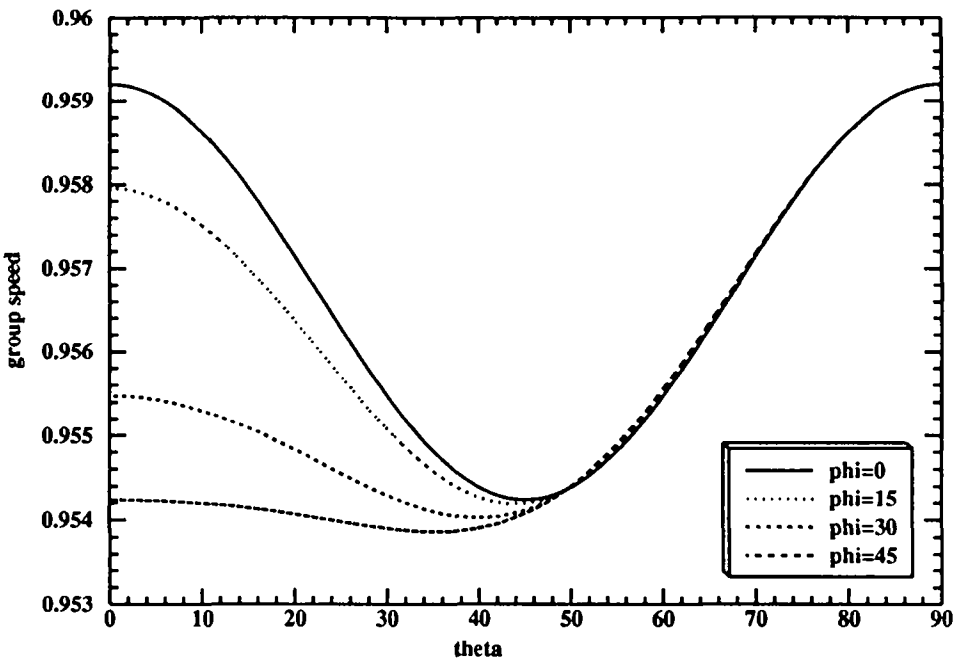



Figure 6.4: Group speed with $N = 9$ and $\lambda/\mu = 50$.

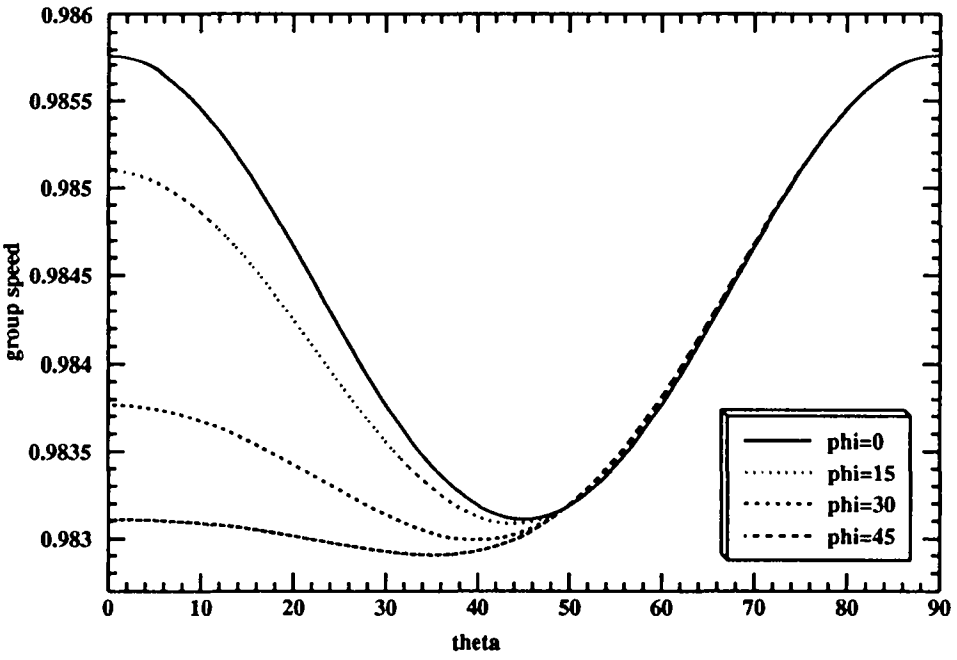



Figure 6.5: Group speed with $N = 10.5$ and $\lambda/\mu = 0$.

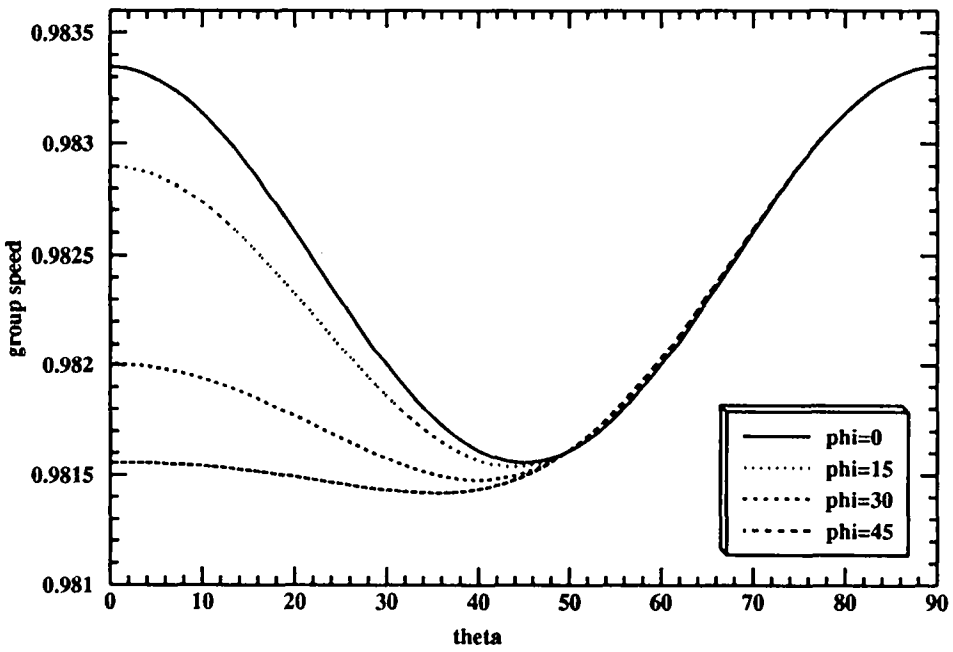



Figure 6.6: Group speed with $N = 11.5$ and $\lambda/\mu = 50$.

# 7 Conclusion

With our simple basic operator (2.1) we have constructed fourth order schemes in a large number of situations. It was not sufficiently general to handle arbitrary tensors, as we saw, but evidently with same kind of techniques such situations can be treated if desired.

The interesting points of these kind of schemes seem to be the following. First using formally the 'intermediate' points make the resulting difference operators shorter which is important from the practical point of view. Also when doing the Fourier stability analysis one can freely use these points as if they were 'real'.

The second important feature is that we explicitly construct the difference operator and its transpose, so that we have immediately a discrete variational formulation of the problem, and the corresponding coercivity result. In particular variable coefficients are as easy to treat as when using finite elements. In fact there seems to be no the most 'natural' way to take into account the variable coefficients: any reasonable choice leads to a stable and consistent scheme. This means that when one has some specific information on the coefficients, it is possible to try to optimize the scheme and take into account this information. Of course, by construction, the operator and its transpose are simply related (typically by a simple shift) so that in practice there are no additional costs. So all in all, the implementation is rather straightforward, because basically there is just one (rather simple) operator which is used a couple of times at each time step.

As regards the accuracy, the error in the group speed is less than 2 %, if $N \simeq 9$ (with the elastodynamic equation $N \in [10.5, 11.5]$, depending on the parameters). With Holberg's method (see [HO], [SPKV]) taking $N$ about five seems to be sufficient; however, in [HO] only the standard wave equation is treated and [SPKV] (as well as [BCL] for second order schemes) give results only for two dimensional elastodynamic equation. In addition, Holberg's optimal difference operators are quite long so one should really compare the overall computational cost instead of the values of $N$. Anyway, we think that our method is also interesting because of its generality: almost any 'reasonable' linear hyperbolic problem can be dealt with.

**Acknowledgements** I would like to thank Patrick Joly for many illuminating discussions on the wave phenomena. The Mathematica (see [WO]) has been of a great help for doing various tedious calculations.

# References

[BCL] A. Bamberger, G. Chavent, P. Lailly: *Etude de schémas numériques pour les équations de l'élastodynamique linéaire*; INRIA, rapport de recherche, 41, 1980.

[BO] A. Bossavit: *Un nouveau point de vue sur les éléments mixtes*; Matapli, SMAI, no 20, Octobre, 1989.

[CJ] G. Cohen, P. Joly: *Schémas d'ordre quatre en temps et en espace pour l'équation des ondes acoustiques*; INRIA, rapport de recherche, to appear.

[CO] G. Cohen: *Fourth order schemes for elastic wave propagation*; INRIA, rapport de recherche, to appear.

[HO] O. Holberg: *Computational aspects of the choice of operator and sampling interval for numerical differentiation in large-scale simulation of wave phenomena*; Geophysical Prospecting, 35, 629-655, 1987.

[SB] G. R. Shubin, J. B. Bell: *A modified equation approach to constructing fourth order methods for acoustic wave propagation*; SIAM J. Sci. Stat. Comp. 8, 135-151, 1987.

[SPKV] P. Sguazzero, A. Parisi, A. Kamel, A. Vesnaver: *Implementation of some explicit dispersion-bounded staggered schemes for the numerical integration of the elastodynamic equations*; in G. Cohen, L. Halpern, P. Joly (eds): *Mathematical and numerical aspects of wave propagation phenomena*; 35-43, SIAM Proceedings, 1991.

[TR] L. N. Trefethen: *Group velocity in finite difference schemes*; SIAM Review, 24, 1982.

[TU] J. Tuomela: *Fourth order schemes for wave equation, Maxwell's equations and linearized elastodynamic equations*; INRIA, rapport de recherche 1337, 1990.

[WO] S. Wolfram: *Mathematica: a system for doing mathematics by computer*; Addison-Wesley, 1988.



Imprimé en France
par
l'Institut National de Recherche en Informatique et en Automatique

ISSN 0249 - 6399