Chapter 1

Linear Analysis of Numerical Methods

The most important property of a numerical method is convergence. Convergence guarantees that as we increase the spatial and temporal resolution of a numerical method, its numerical solution approaches the solution of the partial differential equations. For linear partial differential equations the Lax-equivalence theorem states that a numerical method is convergent if and only if it is stable and consistent [5]. Where consistency means that the error introduced by the numerical method at every time step approaches zero as the spatial and temporal resolution is increased. While stability means that the errors from all previous time steps are not amplified by each evolution step.

The convergence of the FEVM introduced in Chapter [] has not been shown for the linearised Serre equations and so we investigate it here. The consistency of the FEVM follows from our use of well tested approximations and so we instead focus on demonstrating its stability. In particular we performed a Von Neumann stability [1] analysis on the FEVM applied to the linearised Serre equations for waves on quiescent water. At the end we will also present the results of the Von Neumann stability analysis for all methods described in this thesis for quiescent water and for the finite difference methods for flowing water.

Another attractive property of the Serre equations is that it possesses a dispersion relation that well approximates the dispersion relation for the Euler equations. For this reason we wish to know how well the dispersion relation for the FEVM approximates the dispersion relation of the Serre equations. We will demonstrate how the dispersion relationship is derived for the second-order FEVM for waves on quiescent water. At the end we will also present the results of the dispersion relation analysis for the FDVM.

The aim of both analyses is to produce a relationship between the values of the quantities h, u and G at the current time level with their value at the next time level. Since any two of these quantities completely determines the other one, we only need to describe the relationship for two quantities. For example for h and G we would derive an equation of the form

$$\begin{bmatrix} h \\ G \end{bmatrix}_j^{n+1} = \mathbf{E} \begin{bmatrix} h \\ G \end{bmatrix}_j^{n} \tag{1.1}$$

where $\mathbf{E}$ is the evolution matrix. The evolution matrix $\mathbf{E}$ is obtained in the analyses by propagating Fourier modes through the numerical scheme applied to the linearised Serre equations with horizontal beds. We neglect bed terms because variations in the bed have no effect on the dispersion relation. We begin by giving the linearised Serre equations, performing the dispersion relation analysis and then performing the stability analysis.

1.1 Linearised Serre equations with horizontal bed

The Serre equations with a horizontal bed (**??**) are linearised by considering waves as small perturbations $\delta\eta$ and $\delta\upsilon$ on a flow with a mean height H and a mean velocity U respectively. So we have that

$$h(x,t) = H + \delta\eta(x,t) + \mathcal{O}\left(\delta^2\right), \tag{1.2a}$$

$$u(x,t) = U + \delta\upsilon(x,t) + \mathcal{O}\left(\delta^2\right), \tag{1.2b}$$

where $\delta \ll 1$. These waves are relatively small so terms of order δ^2 are negligible. We substitute (1.2) into the Serre equations and neglect terms of order δ^2 to obtain

$$\frac{\partial\left(\delta\eta\right)}{\partial t} + H\frac{\partial\left(\delta\upsilon\right)}{\partial x} + U\frac{\partial\left(\delta\eta\right)}{\partial x} = 0, \tag{1.3a}$$

$$H\frac{\partial\left(\delta\upsilon\right)}{\partial t} + gH\frac{\partial\left(\delta\eta\right)}{\partial x} + UH\frac{\partial\left(\delta\upsilon\right)}{\partial x} - \frac{H^3}{3}\left(U\frac{\partial^3\left(\delta\upsilon\right)}{\partial x^3} + \frac{\partial^3\left(\delta\upsilon\right)}{\partial x^2\partial t}\right) = 0 \tag{1.3b}$$

and for G

$$G = UH + U\delta\eta + H\delta\upsilon - \frac{H^3}{3}\frac{\partial^2\left(\delta\upsilon\right)}{\partial x^2}. \tag{1.3c}$$

For waves on quiescent water we have that $U = 0$, so (1.3) reduces to

$$\frac{\partial \eta}{\partial t} + H\frac{\partial \upsilon}{\partial x} = 0, \tag{1.4a}$$

$$H\frac{\partial \upsilon}{\partial t} + gH\frac{\partial \eta}{\partial x} - \frac{H^3}{3}\left(\frac{\partial^3 \upsilon}{\partial x^3 \partial t}\right) = 0 \tag{1.4b}$$

with

$$G = H\upsilon - \frac{H^3}{3}\frac{\partial^2 \upsilon}{\partial x^2}. \tag{1.4c}$$

Where we have absorbed the δ factor into the corresponding terms η and υ to ease notation. The linearised equations (1.4) can be reformulated

$$\frac{\partial \eta}{\partial t} + H\frac{\partial \upsilon}{\partial x} = 0, \tag{1.5a}$$

$$\frac{\partial G}{\partial t} + gH\frac{\partial \eta}{\partial x} = 0. \tag{1.5b}$$

1.2 Dispersion Relation Analysis

To study the error in the dispersion relation caused by the numerical methods we will follow the work of [3]. We will demonstrate this analysis for one example; the second order FEVM. We will then show how the analysis extends to the second-order FDVM and then present the results for every FDVM and the FEVM. We will then compare the dispersion errors of all FDVM and the FEVM.

As in [3], we will study the dispersion of waves on quiescent water and so we are using equation (1.5). This is a reasonable simplification because for most ocean wave applications we are interested in modelling waves on quiescent water.

We will assume that η and υ are periodic functions in both space and time. In particular, we assume that these quantities are Fourier modes, which for a general quantity q means

$$q(x,t) = q(0,0)e^{i(\omega t + kx)}. \tag{1.6}$$

This is precisely the assumption made to derive the analytical dispersion relation of the linearised Serre equations []. A consequence of a quantity q being a Fourier mode represented on uniform temporal and spatial grid is that for any real numbers m and l we have

$$q_{j+l}^{n+m} = q_j^n e^{i(m\omega\Delta t + lk\Delta x)}. \tag{1.7}$$

Because η and υ are Fourier modes then so is G. Furthermore the cell averages of these quantities $\overline{\eta}$, $\overline{\upsilon}$ and $\overline{G}$ are Fourier modes as well.

1.2.1 Overview of the analysis

We will now present a brief overview of the analysis for a single evolution step of the second-order FEVM. In Subsection 1.2.6 we will extend this analysis to the Runge-Kutta steps which use multiple evolution steps to increase the temporal order of accuracy.

For the second-order FEVM the evolution step progresses in the following way

1. Given the vectors of the cell averages $\overline{\boldsymbol{\eta}}$ and $\overline{\boldsymbol{G}}$ at the current time.

2. We calculate η and G at the cell midpoint x_j from the cell averages using $\mathcal{M}$. We also reconstruct η and G at the cell interface $x^-_{j+1/2}$ and $x^+_{j+1/2}$ from the cell average values using $\mathcal{R}^-$ and $\mathcal{R}^+$ respectively. So that

$$\begin{aligned} \eta_j &= \mathcal{M}\left(\overline{\boldsymbol{\eta}}\right), & G_j &= \mathcal{M}\left(\overline{\boldsymbol{G}}\right), \\ \eta^-_{j+1/2} &= \mathcal{R}^-\left(\overline{\boldsymbol{\eta}}\right), & G^-_{j+1/2} &= \mathcal{R}^-\left(\overline{\boldsymbol{G}}\right), \\ \eta^+_{j+1/2} &= \mathcal{R}^+\left(\overline{\boldsymbol{\eta}}\right), & G^+_{j+1/2} &= \mathcal{R}^+\left(\overline{\boldsymbol{G}}\right). \end{aligned}$$

3. We use the map $\mathcal{G}$ given by the elliptic equation between G and υ to calculate $\upsilon_{j+1/2}$ from $\boldsymbol{G}$ and H

$$\upsilon_{j+1/2} = \mathcal{G}\left(H, \boldsymbol{G}\right).$$

4. We calculate the average flux across the cell boundary $x_{j+1/2}$ over time; $F_{j+1/2}$ using $\mathcal{F}$

$$F_{j+1/2} = \mathcal{F}\left(\eta^-_{j+1/2}, G^-_{j+1/2}, \eta^+_{j+1/2}, G^+_{j+1/2}, \upsilon_{j+1/2}\right).$$

5. We repeat this process for each cell edge and then apply the update formula (**??**) to evolve the vectors $\overline{\boldsymbol{\eta}}$ and $\overline{\boldsymbol{G}}$ from the current time level to the next time level.

This analysis uses the fact that υ, η and G are Fourier modes and so we have relations between our quantities at different grid points (1.7). Together with the fact that these operators are just linear combinations of the quantities at different grid points, to derive factors for the operators $\mathcal{M}$, $\mathcal{R}^-$, $\mathcal{R}^+$ and $\mathcal{G}$ that are the same for every time step and grid point.

For example we have in the case of the map between cell averages and nodal values $\mathcal{M}$ that this process for η leads to

$$\eta_j = \mathcal{M}\overline{\eta}_j.$$

Where $\mathcal{M}$ is the same for every cell and time level.

These operators are combined to give the matrix $\mathbf{F}$ that calculates the flux at the current time allowing us to write the update formula (**??**) as

$$\begin{bmatrix} \overline{\eta} \\ \overline{G} \end{bmatrix}_j^{n+1} = (\mathbf{I} - \Delta t \mathbf{F}) \begin{bmatrix} \overline{\eta} \\ \overline{G} \end{bmatrix}_j^{n} = \mathbf{E} \begin{bmatrix} \overline{\eta} \\ \overline{G} \end{bmatrix}_j^{n}$$

so we obtain an equation of the form (1.1) for a single forward Euler step. From this equation the dispersion relation of the method can be calculated for a single forward Euler step. We shall now derive the factors for each of the operators in the order in which they appear in our outline of a single evolution step for the second-order FEVM.

1.2.2 Step 2: Reconstruction

Given $\overline{\boldsymbol{\eta}}$ and $\overline{\boldsymbol{G}}$ at t^n the second step of our numerical method is to calculate η and G at x_j using $\mathcal{M}$ and at $x^-_{j+1/2}$ and $x^+_{j+1/2}$ using $\mathcal{R}^-$ and $\mathcal{R}^+$ respectively. The derivation of the factors for these operators is given in terms of a general quantity q, as they are the same for η and G.

Cell average values to nodal values: $\mathcal{M}$

For the second-order FEVM we use the fact that

$$\overline{q}_j = q_j + \mathcal{O}\left(\Delta x^2\right).$$

So to attain second-order accuracy we use

$$q_j = \overline{q}_j = \mathcal{M}\overline{q}_j. \tag{1.8}$$

Therefore we have a factor $\mathcal{M}$ representing the map between cell averages and nodal values for our numerical method that is the same for every grid point and time step, as desired.

Cell average values to interface values: $\mathcal{R}^-$ and $\mathcal{R}^+$

We reconstruct η and G at $x^-_{j+1/2}$ and $x^+_{j+1/2}$ as we allow these quantities to be discontinuous across the cell interfaces in our finite volume method. However, since we are assuming that these quantities are Fourier modes and therefore smooth we do not require non-linear limiters to ensure our scheme is TVD. Without limiters our reconstruction scheme for η and G written for a general quantity q is

$$\begin{aligned} q^-_{j+\frac{1}{2}} &= \overline{q}_j + \frac{-\overline{q}_{j-1} + \overline{q}_{j+1}}{4}, \\ q^+_{j+\frac{1}{2}} &= \overline{q}_{j+1} + \frac{-\overline{q}_j + \overline{q}_{j+2}}{4}. \end{aligned}$$

Using (1.7) and (1.8) these equations become

$$q^-_{j+\frac{1}{2}} = \overline{q}_j + \frac{-\overline{q}_j e^{-ik\Delta x} + \overline{q}_j e^{ik\Delta x}}{4} = \left(1 + \frac{i\sin(k\Delta x)}{2}\right)\overline{q}_j = \mathcal{R}^-\overline{q}_j, \tag{1.9a}$$

$$q^+_{j+\frac{1}{2}} = \frac{\overline{q}_j e^{ik\Delta x} + \overline{q}_j + \overline{q}_j e^{2ik\Delta x}}{4} = e^{ik\Delta x}\left(1 - \frac{i\sin(k\Delta x)}{2}\right)\overline{q}_j = \mathcal{R}^+\overline{q}_j. \tag{1.9b}$$

These are the reconstruction factors for both η and G.

1.2.3 Step 3: Solving The Elliptic Equation

We begin our FEM for (1.4c) with its weak formulation, obtained by multiplying by a test function τ and integrating over the domain Ω

$$\int_\Omega G\tau \, dx = H\int_\Omega \upsilon\tau \, dx + \frac{H^3}{3}\int_\Omega \frac{\partial \upsilon}{\partial x}\frac{\partial \tau}{\partial x} \, dx.$$

For G we use the basis functions $\psi^+_{j-1/2}$ and $\psi^-_{j+1/2}$ defined in Chapter [], which means G is linear inside a cell with discontinuous jumps at the cell edges. For τ and υ we use the basis functions $\phi_{j-1/2}$, ϕ_j and $\phi_{j+1/2}$ defined in Chapter [], so that τ and υ are quadratic functions inside a cell that are continuous across the cell edges. Substituting in the approximations to our quantities based on these basis functions and breaking our integration up into the sum of the integrals over a cell as we did in Chapter [], we get that

$$\sum_j \int_{x_{j-1/2}}^{x_{j+1/2}} \left(G^+_{j-1/2}\psi^+_{j-1/2} + G^-_{j+1/2}\psi^-_{j+1/2}\right) \begin{bmatrix}\phi_{j-1/2}\\ \phi_j \\ \phi_{j+1/2}\end{bmatrix} dx =$$

$$\sum_j H \int_{x_{j-1/2}}^{x_{j+1/2}} \left(\upsilon_{j-1/2}\phi_{j-1/2} + \upsilon_j\phi_j + \upsilon_{j+1/2}\phi_{j+1/2}\right) \begin{bmatrix}\phi_{j-1/2}\\ \phi_j \\ \phi_{j+1/2}\end{bmatrix} dx$$

$$+ \sum_j \frac{H^3}{3} \int_{x_{j-1/2}}^{x_{j+1/2}} \left(\upsilon_{j-1/2}\frac{\partial \phi_{j-1/2}}{\partial x} + \upsilon_j \frac{\partial \phi_j}{\partial x} + \upsilon_{j+1/2}\frac{\partial \phi_{j+1/2}}{\partial x}\right) \begin{bmatrix}\frac{\partial \phi_{j-1/2}}{\partial x}\\ \frac{\partial \phi_j}{\partial x} \\ \frac{\partial \phi_{j+1/2}}{\partial x}\end{bmatrix} dx.$$

By calculating all the integrals of the appropriate basis function combinations we get that

$$\sum_j \frac{\Delta x}{6} \begin{bmatrix} G^+_{j-1/2} \\ 2G^+_{j-1/2} + 2G^-_{j+1/2} \\ G^-_{j+1/2}\end{bmatrix} =$$

$$\sum_j \left(H\frac{\Delta x}{30}\begin{bmatrix} 4 & 2 & -1 \\ 2 & 16 & 2 \\ -1 & 2 & 4\end{bmatrix} + \frac{H^3}{9\Delta x}\begin{bmatrix} 7 & -8 & 1 \\ -8 & 16 & -8 \\ 1 & -8 & 7\end{bmatrix}\right) \begin{bmatrix}\upsilon_{j-1/2} \\ \upsilon_j \\ \upsilon_{j+1/2}\end{bmatrix}.$$

Using our relations from the periodic nature of υ and $\overline{G}$ (1.7), and the reconstructions $\mathcal{R}^+$ and $\mathcal{R}^-$ used on $\overline{G}$ to obtain $G^+_{j+1/2}$ and $G^-_{j+1/2}$ respectively, we obtain

$$\sum_j \frac{\Delta x}{6} \begin{bmatrix} e^{-ik\Delta x}\mathcal{R}^+\overline{G}_j \\ 2e^{-ik\Delta x}\mathcal{R}^+\overline{G}_j + 2\mathcal{R}^-\overline{G}_j \\ \mathcal{R}^-\overline{G}_j\end{bmatrix} =$$

$$\sum_j \left(H\frac{\Delta x}{30}\begin{bmatrix} 4 & 2 & -1 \\ 2 & 16 & 2 \\ -1 & 2 & 4\end{bmatrix} + \frac{H^3}{9\Delta x}\begin{bmatrix} 7 & -8 & 1 \\ -8 & 16 & -8 \\ 1 & -8 & 7\end{bmatrix}\right) \begin{bmatrix} e^{-ik\frac{\Delta x}{2}}\upsilon_j \\ \upsilon_j \\ e^{ik\frac{\Delta x}{2}}\upsilon_j\end{bmatrix},$$

$$\sum_j \frac{\Delta x}{6}\begin{bmatrix} e^{-ik\Delta x}\mathcal{R}^+ \\ 2e^{-ik\Delta x}\mathcal{R}^+ + 2\mathcal{R}^- \\ \mathcal{R}^- \end{bmatrix}\overline{G}_j =$$

$$\sum_j \left(H\frac{\Delta x}{30}\begin{bmatrix} 4e^{-ik\frac{\Delta x}{2}} + 2 - e^{ik\frac{\Delta x}{2}} \\ 2e^{-ik\frac{\Delta x}{2}} + 16 + 2e^{ik\frac{\Delta x}{2}} \\ -e^{-ik\frac{\Delta x}{2}} + 2 + 4e^{ik\frac{\Delta x}{2}} \end{bmatrix} + \frac{H^3}{9\Delta x}\begin{bmatrix} 7e^{-ik\frac{\Delta x}{2}} - 8 + e^{ik\frac{\Delta x}{2}} \\ -8e^{-ik\frac{\Delta x}{2}} + 16 - 8e^{ik\frac{\Delta x}{2}} \\ e^{-ik\frac{\Delta x}{2}} - 8 + 7e^{ik\frac{\Delta x}{2}} \end{bmatrix}\right)\upsilon_j.$$

The first element of the vector corresponds to the jth cell's contribution to the equation for $\upsilon_{j-1/2}$, the second element corresponds to the equation for υ_j and the last element corresponds to the jth cells contribution to the equation for $\upsilon_{j+1/2}$. Since our flux calculation (1.13) only requires $\upsilon_{j+1/2}$ our FEM can neglect the other terms and focus on solving the equation represented by the last element of the vectors. However, so far we have only given the contribution to $\upsilon_{j+1/2}$ from the jth cell, but $\upsilon_{j+1/2}$ will also get a contribution from the $(j+1)$th cell as $\phi_{j+1/2}$ is also non-zero there. Taking this into account we get that the equation represented by the last element of all the vectors is

$$\begin{aligned}
&\frac{\Delta x}{6}\left(\mathcal{R}^- + \mathcal{R}^+\right)\overline{G}_j = \\
&\quad\left(H\frac{\Delta x}{30}\left(-e^{-ik\frac{\Delta x}{2}} + 2 + 4e^{ik\frac{\Delta x}{2}} + e^{ik\Delta x}\left(4e^{-ik\frac{\Delta x}{2}} + 2 - e^{ik\frac{\Delta x}{2}}\right)\right)\right. \\
&\quad\left. + \frac{H^3}{9\Delta x}\left(e^{-ik\frac{\Delta x}{2}} - 8 + 7e^{ik\frac{\Delta x}{2}} + e^{ik\Delta x}\left(7e^{-ik\frac{\Delta x}{2}} - 8 + e^{ik\frac{\Delta x}{2}}\right)\right)\right)\upsilon_j. \\
&\quad = \left(H\frac{\Delta x}{30}\left(4\cos\left(\frac{k\Delta x}{2}\right) - 2\cos\left(k\Delta x\right) + 8\right)\right. \\
&\qquad\left. + \frac{H^3}{9\Delta x}\left(-16\cos\left(\frac{k\Delta x}{2}\right) + 2\cos\left(k\Delta x\right) + 14\right)\right)e^{ik\frac{\Delta x}{2}}\upsilon_j.
\end{aligned}$$

Using (1.7) we have that

$$
\begin{aligned}
\upsilon_{j+1/2} = &\left(\frac{\Delta x}{6}\left(\mathcal{R}^- + \mathcal{R}^+\right)\right) \\
&\div \left(H\frac{\Delta x}{30}\left(2\left(2\cos\left(\frac{k\Delta x}{2}\right) - \cos\left(k\Delta x\right) + 4\right)\right)\right. \\
&\left. + \frac{H^3}{9\Delta x}\left(-16\cos\left(\frac{k\Delta x}{2}\right) + 2\cos\left(k\Delta x\right) + 14\right)\right)\overline{G}_j = \mathcal{G}\overline{G}_j. \quad (1.10)
\end{aligned}
$$

1.2.4 Step 4: Flux calculation

To calculate the average flux $F_{j+1/2}$ we use Kurganov's method [4]. For the linearised Serre equations we have the wave speed bounds (**??**), so that

$$a^-_{j+1/2} = -\sqrt{gH} \qquad \text{and} \qquad a^+_{j+1/2} = \sqrt{gH}. \quad (1.11)$$

Substituting these into our average flux approximation (**??**) we obtain $F_{j+\frac{1}{2}}$, for η from (1.5a) we have

$$F^{\eta}_{j+\frac{1}{2}} = \frac{H\upsilon^-_{j+\frac{1}{2}} + H\upsilon^+_{j+\frac{1}{2}}}{2} - \frac{\sqrt{gH}}{2}\left[\eta^+_{j+\frac{1}{2}} - \eta^-_{j+\frac{1}{2}}\right]. \quad (1.12)$$

Since υ is continuous across the cell interface we have that $\upsilon_{j+1/2} = \upsilon^-_{j+\frac{1}{2}} = \upsilon^+_{j+\frac{1}{2}}$. By using this and substituting the appropriate factors (1.9) and (1.10) into (1.12) we obtain

$$F^{\eta}_{j+\frac{1}{2}} = H\mathcal{G}\overline{G}_j - \frac{\sqrt{gH}}{2}\left[\mathcal{R}^+ - \mathcal{R}^-\right]\overline{\eta}_j = \mathcal{F}^{\eta,G}\overline{G}_j + \mathcal{F}^{\eta,\eta}\overline{\eta}_j. \quad (1.13)$$

Doing the same for G we get that the Kurganov approximation to the flux average of (1.5b) is

$$F^{G}_{j+\frac{1}{2}} = \frac{gH\eta^-_{j+\frac{1}{2}} + gH\eta^+_{j+\frac{1}{2}}}{2} - \frac{\sqrt{gH}}{2}\left[G^+_{j+\frac{1}{2}} - G^-_{j+\frac{1}{2}}\right]. \quad (1.14)$$

Substituting in the appropriate reconstruction coefficients (1.9) into (1.14) we get that

$$F^{G}_{j+\frac{1}{2}} = gH\frac{\mathcal{R}^- + \mathcal{R}^+}{2}\overline{\eta}_j - \frac{\sqrt{gH}}{2}\left[\mathcal{R}^+ - \mathcal{R}^-\right]\overline{G}_j = \mathcal{F}^{G,\eta}\overline{\eta}_j + \mathcal{F}^{G,G}\overline{G}_j. \quad (1.15)$$

1.2.5 Step 5: Evolution Matrix

By substituting our flux approximations (1.13) and (1.15) into our update scheme (??) and making use of (1.8) our second order finite difference volume method can be written as

$$\begin{aligned}\overline{\eta}_j^{n+1} &= \overline{\eta}_j^{n} - \frac{\Delta t}{\Delta x}\left[\left(\mathcal{F}^{\eta,\eta}\overline{\eta}_j + \mathcal{F}^{\eta,G}\overline{G}_j\right) - \left(\mathcal{F}^{\eta,\eta}\overline{\eta}_{j-1} + \mathcal{F}^{\eta,G}\overline{G}_{j-1}\right)\right],\\ \overline{G}_j^{n+1} &= \overline{G}_j^{n} - \frac{\Delta t}{\Delta x}\left[\left(\mathcal{F}^{G,\eta}\overline{\eta}_j + \mathcal{F}^{G,G}\overline{G}_j\right) - \left(\mathcal{F}^{G,\eta}\overline{\eta}_{j-1} + \mathcal{F}^{G,G}\overline{G}_{j-1}\right)\right].\end{aligned}$$

Furthermore by making use of (1.7) we obtain

$$\begin{aligned}\overline{\eta}_j^{n+1} &= \overline{\eta}_j^{n} - \frac{\Delta t}{\Delta x}\left[\left(1 - e^{-ik\Delta x}\right)\left(\mathcal{F}^{\eta,\eta}\overline{\eta}_j + \mathcal{F}^{\eta,G}\overline{G}_j\right)\right],\\ \overline{G}_j^{n+1} &= \overline{G}_j^{n} - \frac{\Delta t}{\Delta x}\left[\left(1 - e^{-ik\Delta x}\right)\left(\mathcal{F}^{G,\eta}\overline{\eta}_j + \mathcal{F}^{G,G}\overline{G}_j\right)\right].\end{aligned}$$

This can be written in matrix form as

$$\begin{aligned}\begin{bmatrix}\overline{\eta}\\ \overline{G}\end{bmatrix}_j^{n+1} &= \begin{bmatrix}\overline{\eta}\\ \overline{G}\end{bmatrix}_j^{n} - \frac{\left(1 - e^{-ik\Delta x}\right)\Delta t}{\Delta x}\begin{bmatrix}\mathcal{F}^{\eta,\eta} & \mathcal{F}^{\eta,G}\\ \mathcal{F}^{G,\eta} & \mathcal{F}^{G,G}\end{bmatrix}\begin{bmatrix}\overline{\eta}\\ \overline{G}\end{bmatrix}_j^{n}\\ &= \left(\mathbf{I} - \Delta t\mathbf{F}\right)\begin{bmatrix}\overline{\eta}\\ \overline{G}\end{bmatrix}_j^{n} \qquad (1.16)\end{aligned}$$

for a single Euler step as desired.

1.2.6 SSP Runge-Kutta Time Stepping

Since we have demonstrated this process for a single evolution step the analysis will now proceed to the SSP Runge-Kutta time stepping that allows our FEVM to be temporally higher order accurate.

The second-order SSP Runge Kutta time stepping proceeds as follows

$$\begin{bmatrix}\overline{\eta}\\ \overline{G}\end{bmatrix}_j^{\prime} = \left(\mathbf{I} - \Delta t\mathbf{F}\right)\begin{bmatrix}\overline{\eta}\\ \overline{G}\end{bmatrix}_j^{n}, \qquad (1.17a)$$

$$\begin{bmatrix}\overline{\eta}\\ \overline{G}\end{bmatrix}_j^{\prime\prime} = \left(\mathbf{I} - \Delta t\mathbf{F}\right)\begin{bmatrix}\overline{\eta}\\ \overline{G}\end{bmatrix}_j^{\prime}, \qquad (1.17b)$$

$$\begin{bmatrix}\overline{\eta}\\\overline{G}\end{bmatrix}_j^{n+1} = \frac{1}{2}\left(\begin{bmatrix}\overline{\eta}\\\overline{G}\end{bmatrix}_j^{n} + \begin{bmatrix}\overline{\eta}\\\overline{G}\end{bmatrix}_j^{\prime\prime}\right). \tag{1.17c}$$

Substituting (1.17a) and (1.17b) into (1.17c) we can write this in terms of the flux matrix $\mathbf{F}$ and our cell averages at t^n as

$$\begin{bmatrix}\overline{\eta}\\\overline{G}\end{bmatrix}_j^{n+1} = \frac{1}{2}\left(\begin{bmatrix}\overline{\eta}\\\overline{G}\end{bmatrix}_j^{n} + (\mathbf{I} - \Delta t\mathbf{F})^2\begin{bmatrix}\overline{\eta}\\\overline{G}\end{bmatrix}_j^{n}\right).$$

Expanding $(\mathbf{I} - \Delta t\mathbf{F})^2$ we get

$$\begin{bmatrix}\overline{\eta}\\\overline{G}\end{bmatrix}_j^{n+1} = \frac{1}{2}\left(2\mathbf{I} - 2\Delta t\mathbf{F} + \Delta t^2\mathbf{F}^2\right)\begin{bmatrix}\overline{\eta}\\\overline{G}\end{bmatrix}_j^{n} = \mathbf{E}\begin{bmatrix}\overline{\eta}\\\overline{G}\end{bmatrix}_j^{n}. \tag{1.18}$$

Provided that the eigenvalue decomposition $\mathbf{F} = \mathbf{P\Lambda P}^{-1}$ exists, then (1.18) can be rewritten as

$$\begin{bmatrix}\overline{\eta}\\\overline{G}\end{bmatrix}_j^{n+1} = \frac{1}{2}\left(2\mathbf{I} - 2\Delta t\mathbf{P\Lambda P}^{-1} + \Delta t^2\mathbf{P\Lambda}^2\mathbf{P}^{-1}\right)\begin{bmatrix}\overline{\eta}\\\overline{G}\end{bmatrix}_j^{n}.$$

Multiplying both sides by $\mathbf{P}^{-1}$ on the left we obtain

$$\mathbf{P}^{-1}\begin{bmatrix}\overline{\eta}\\\overline{G}\end{bmatrix}_j^{n+1} = \frac{1}{2}\left(2 - 2\Delta t\mathbf{\Lambda} + \Delta t^2\mathbf{\Lambda}^2\right)\mathbf{P}^{-1}\begin{bmatrix}\overline{\eta}\\\overline{G}\end{bmatrix}_j^{n}.$$

Since $\overline{\eta}$ and $\overline{G}$ are Fourier modes we have from (1.7) that

$$e^{i\omega\Delta t}\left(\mathbf{P}^{-1}\begin{bmatrix}\overline{\eta}\\\overline{G}\end{bmatrix}_j^{n}\right) = \left(1 - 1\Delta t\mathbf{\Lambda} + \frac{1}{2}\Delta t^2\mathbf{\Lambda}^2\right)\left(\mathbf{P}^{-1}\begin{bmatrix}\overline{\eta}\\\overline{G}\end{bmatrix}_j^{n}\right).$$

Since $\mathbf{\Lambda}$ is a diagonal matrix consisting of the eigenvalues λ_+ and λ_- we have that

$$e^{i\omega_\pm\Delta t} = 1 + \frac{1}{2}\Delta t^2\lambda_\pm^2 - \Delta t\lambda_\pm.$$

Where the subscript ± denotes the positive and negative branch of the dispersion relationship for the right and left travelling waves respectively. The dispersion relation for the second order FEVM is then

$$\omega_\pm = \frac{1}{i\Delta t}\ln\left(1 + \frac{1}{2}\Delta t^2\lambda_\pm^2 - \Delta t\lambda_\pm\right). \tag{1.19}$$

1.2.7 Derivation of $\mathcal{G}$ for the Finite Difference Method

The derivation of $\mathcal{G}$ for the FDVM is very different and so we would like to demonstrate this derivation using the second-order FDVM as an example. To calculate $\upsilon_{j+1/2}$ in the FDVM there are three steps: first we calculate the nodal values $\boldsymbol{G}$ from the cell averages $\overline{\boldsymbol{G}}$, second we calculate the nodal values $\boldsymbol{\upsilon}$ by solving the elliptic equation (1.4c) and then we reconstruct $\upsilon_{j+1/2}$ from $\boldsymbol{\upsilon}$.

For the second-order FDVM we use $\mathcal{M} = 1$ as we did for the second-order FEVM (1.8) to compute $\boldsymbol{G}$ from $\overline{\boldsymbol{G}}$.

For the second-order FDVM we use the second-order centred finite difference approximation to (1.4c)

$$G_j = H\upsilon_j - \frac{H^3}{3}\frac{\upsilon_{j-1} - 2\upsilon_j + \upsilon_{j+1}}{\Delta x^2}.$$

Using (1.7) this becomes

$$\begin{aligned} G_j &= H\upsilon_j - \frac{H^3}{3}\frac{e^{-ik\Delta x}\upsilon_j - 2\upsilon_j + e^{ik\Delta x}\upsilon_j}{\Delta x^2} \\ &= \left(H - \frac{H^3}{3}\frac{2\cos(k\Delta x) - 2}{\Delta x^2}\right)\upsilon_j. \end{aligned} \tag{1.20}$$

To reconstruct $\upsilon_{j+1/2}$ from the nodal values up to second-order accuracy we use

$$\upsilon_{j+1/2} = \frac{\upsilon_j + \upsilon_{j+1}}{2}.$$

Using (1.7) this becomes

$$\upsilon_{j+1/2} = \frac{\upsilon_j + e^{ik\Delta x}\upsilon_j}{2} = \frac{1 + e^{ik\Delta x}}{2}\upsilon_j. \tag{1.21}$$

Combining (1.8), (1.20) and (1.21) we have that

$$\mathcal{M}\overline{G}_j = \left(H - \frac{H^3}{3}\frac{2\cos(k\Delta x) - 2}{\Delta x^2}\right)\frac{2}{1 + e^{ik\Delta x}}\upsilon_{j+1/2}.$$

Therefore for the second-order FDVM we have

$$\upsilon_{j+1/2} = \frac{3\mathcal{M}\Delta x^2\left(\frac{1+e^{ik\Delta x}}{2}\right)}{3\Delta x^2 H - H^3\left(2\cos(k\Delta x) - 2\right)}\overline{G}_j = \mathcal{G}\overline{G}_j.$$

1.2.8 Derived Expressions for all Methods

In the following we present tables which give both the formula for the fundamental approximations and the lowest order term of the Taylor series for the error between the approximation and the analytic value. In particular we take the error to be the value of the approximation minus the analytic value. We also present the error for the elements of the flux matrix $\mathbf{F}$ and the error in the dispersion relation to demonstrate that when combined the steps of our numerical method do indeed provide us with the correct order of accuracy in both space and time.

Tables for Factors: $\mathcal{M}$ example

We will demonstrate how the tables for $\mathcal{M}$, $\mathcal{R}^-$, $\mathcal{R}^+$ and $\mathcal{G}$ are constructed by using $\mathcal{M}$ as an example and then just present the tables for the other factors. First we calculate the analytic value for $\mathcal{M}$. For a general quantity q we have by definition [] that

$$\overline{q}_j = \frac{1}{\Delta x}\int_{x_{j-1/2}}^{x_{j+1/2}} q \; dx.$$

Assuming q is a Fourier mode by (1.6) we have that

$$\begin{aligned}\overline{q}_j &= \frac{1}{\Delta x}\int_{x_{j-1/2}}^{x_{j+1/2}} q(0,0)e^{i(\omega t+kx)} \; dx = \frac{q(0,0)e^{i\omega t}}{\Delta x}\left[\frac{1}{ik}e^{ikx}\right]_{x_{j-1/2}}^{x_{j+1/2}} \\ &= \frac{q(0,0)e^{i\omega t}}{\Delta x}\frac{1}{ik}e^{ikx_j}\left[e^{ik\frac{\Delta x}{2}} - e^{-ik\frac{\Delta x}{2}}\right] = \frac{q(0,0)e^{i(\omega t+kx_j)}}{\Delta x}\frac{1}{ik}\left[2i\sin\left(k\frac{\Delta x}{2}\right)\right] \\ &= \frac{2}{k\Delta x}\sin\left(k\frac{\Delta x}{2}\right)q_j.\end{aligned}$$

Therefore,

$$q_j = \frac{k\Delta x}{2\sin\left(k\frac{\Delta x}{2}\right)}\overline{q}_j = \mathcal{M}\overline{q}_j. \tag{1.22}$$

This is the analytic value of $\mathcal{M}$ with which we want to compare the derived $\mathcal{M}$ from our numerical methods. To compare them we take their Taylor series expansion and compare those to get the lowest order term of the error. For the analytic value we have

$$\mathcal{M} = \frac{k\Delta x}{2\sin\left(k\frac{\Delta x}{2}\right)} = 1 + \frac{1}{24}k^2\Delta x^2 + \frac{7}{5760}k^4\Delta x^4 + \cdots . \tag{1.23}$$

Since our value for $\mathcal{M}$ for the second-order FEVM is 1 (1.8) we can see that the lowest order term of the error between the second-order FEVM and the analytical value is $-\frac{1}{24}k^2\Delta x^2$. These results have been summarised in Table 1.1 for all FDVM and the FEVM.

Scheme	Expression	Lowest Order Term of Error
FDVM$_1$	1	$-\frac{1}{24}k^2\Delta x^2$
FDVM$_2$ and FEVM$_2$	1	$-\frac{1}{24}k^2\Delta x^2$
FDVM$_3$	$\frac{26-2\cos(k\Delta x)}{24}$	$-\frac{3}{640}k^4\Delta x^4$

Table 1.1: Factor $\mathcal{M}$ from transformation between nodal and cell average values. Where the analytic value is $\mathcal{M} = \dfrac{k\Delta x}{2\sin\left(k\frac{\Delta x}{2}\right)}$.

Scheme	Formula	Lowest Order Term of Error
FDVM$_1$	$e^{ik\Delta x}$	$\frac{i}{2}k\Delta x$
FDVM$_2$ and FEVM$_2$	$e^{ik\Delta x}\left(1-\frac{i\sin(k\Delta x)}{2}\right)$	$\frac{1}{12}k^2\Delta x^2$
FDVM$_3$	$\frac{e^{ik\Delta x}}{6}\left(5+2e^{-ik\Delta x}-e^{ik\Delta x}\right)$	$\frac{i}{12}k^3\Delta x^3$

Table 1.2: Factor $\mathcal{R}^+$ from reconstruction of η and G at $x^+_{j+1/2}$. Where the analytic value is $\mathcal{R}^+ = e^{ik\Delta x/2}\dfrac{k\Delta x}{2\sin\left(\frac{k\Delta x}{2}\right)}$.

Scheme	Expression	Lowest Order Term of Error
FDVM$_1$	1	$-\frac{i}{2}k\Delta x$
FDVM$_2$ and FEVM$_2$	$1+\frac{i\sin\left(k\Delta x\right)}{2}$	$\frac{1}{12}k^2\Delta x^2$
FDVM$_3$	$\frac{1}{6}\left(5-e^{-ik\Delta x}+2e^{ik\Delta x}\right)$	$-\frac{i}{12}k^3\Delta x^3$

Table 1.3: Factor $\mathcal{R}^-$ from reconstruction of η and G at $x^-_{j+1/2}$. Where the analytic value is $\mathcal{R}^- = e^{ik\Delta x/2}\dfrac{k\Delta x}{2\sin\left(\frac{k\Delta x}{2}\right)}$.

Scheme	Expression	Lowest Order Term of Error
FDVM$_1$	$\dfrac{3\Delta x^2\left(\frac{1+e^{ik\Delta x}}{2}\right)}{3\Delta x^2 H - H^3\left(2\cos\left(k\Delta x\right)-2\right)}$	$-\dfrac{6+H^2k^2}{4H\left(3+H^2k^2\right)^2}k^2\Delta x^2$
FDVM$_2$	$\dfrac{3\Delta x^2\left(\frac{1+e^{ik\Delta x}}{2}\right)}{3\Delta x^2 H - H^3\left(2\cos\left(k\Delta x\right)-2\right)}$	$-\dfrac{6+H^2k^2}{4H\left(3+H^2k^2\right)^2}k^2\Delta x^2$
FEVM$_2$	$\left(\frac{\Delta x}{6}\left(1+\frac{i\sin(k\Delta x)}{2}+e^{ik\Delta x}\left(1-\frac{i\sin(k\Delta x)}{2}\right)\right)\right) \div \Bigg(H\frac{\Delta x}{30}\left(2\left(2\cos\left(\frac{k\Delta x}{2}\right)-\cos\left(k\Delta x\right)+4\right)\right) + \frac{H^3}{9\Delta x}\left(-16\cos\left(\frac{k\Delta x}{2}\right)+2\cos\left(k\Delta x\right)+14\right)\Bigg)$	$\dfrac{12+5H^2k^2}{40H\left(3+H^2k^2\right)^2}k^2\Delta x^2$
FDVM$_3$	$\dfrac{36\Delta x^2\left(\frac{-e^{-ik\Delta x}+9e^{ik\Delta x}-e^{2ik\Delta x}+9}{16}\right)}{36\Delta x^2 H - H^3\left(32\cos\left(k\Delta x\right)-2\cos\left(2k\Delta x\right)-30\right)}$	$-\dfrac{243+49H^2k^2}{960H\left(3+H^2k^2\right)^2}k^4\Delta x^4$

Table 1.4: Factor $\mathcal{G}$ from solving the elliptic equation (1.4c) for $\upsilon_{j+1/2}$. Where the analytic value is $\mathcal{G} = \dfrac{3}{3H+H^3k^2}\dfrac{1}{e^{-ik\Delta x/2}}\dfrac{k\Delta x}{2\sin\left(\frac{k\Delta x}{2}\right)}$.

Scheme	Variable			
	$\frac{\left(1-e^{-ik\Delta x}\right)}{\Delta x}\mathcal{F}^{\eta,\eta}$	$\frac{\left(1-e^{-ik\Delta x}\right)}{\Delta x}\mathcal{F}^{\eta,G}$	$\frac{\left(1-e^{-ik\Delta x}\right)}{\Delta x}\mathcal{F}^{G,\eta}$	$\frac{\left(1-e^{-ik\Delta x}\right)}{\Delta x}\mathcal{F}^{G,G}$
Exact	0	$\frac{3ik}{3+H^2k^2}$	$igkH$	0
Lowest Order Error Term from Taylor Series				
FDVM$_1$	$\frac{\sqrt{gH}}{2}k^2\Delta x$	$-\frac{i\left(6+H^2k^2\right)}{4(3+H^2k^2)^2}k^3\Delta x^2$	$-\frac{igH}{6}k^3\Delta x^2$	$\frac{\sqrt{gH}}{2}k^2\Delta x$
FDVM$_2$	$\frac{\sqrt{gH}}{8}k^4\Delta x^3$	$-\frac{i\left(6+H^2k^2\right)}{4H(3+H^2k^2)^2}k^2\Delta x^2$	$\frac{igH}{12}k^3\Delta x^2$	$\frac{\sqrt{gH}}{8}k^4\Delta x^3$
FEVM$_2$	$\frac{\sqrt{gH}}{8}k^4\Delta x^3$	$\frac{i\left(12+5H^2k^2\right)}{40(3+H^2k^2)^2}k^3\Delta x^2$	$\frac{igH}{12}k^3\Delta x^2$	$\frac{\sqrt{gH}}{8}k^4\Delta x^3$
FDVM$_3$	$\frac{\sqrt{gH}}{12}k^4\Delta x^3$	$-\frac{i\left(243+49H^2k^2\right)}{960(3+H^2k^2)^2}k^5\Delta x^4$	$-\frac{igH}{30}k^5\Delta x^4$	$\frac{\sqrt{gH}}{12}k^4\Delta x^3$

Table 1.5: Spatial error for elements of **F** for all FDVM and the FEVM.

Flux matrix and Dispersion relation Tables

To calculate the elements of **F** we substitute the appropriate expression given in Tables 1.1, 1.2, 1.3 and 1.4 into the equations (1.13) and (1.15) and then compare its Taylor series to the analytic values to get the lowest order error terms. The results of this are given in Table 1.5.

Having calculated **F** for all the methods we can then find its eigenvalues and substitute them into the appropriate dispersion relations given by the Runge-Kutta time stepping method below

$$\text{First-Order: } \omega_{\pm} = \frac{1}{i\Delta t}\ln\left(1 - \Delta t\lambda_{\pm}\right), \tag{1.24a}$$

$$\text{Second-Order: } \omega_{\pm} = \frac{1}{i\Delta t}\ln\left(1 + \frac{1}{2}\Delta t^2\lambda_{\pm}^2 - \Delta t\lambda_{\pm}\right), \tag{1.24b}$$

$$\text{Third-Order: } \omega_{\pm} = \frac{1}{i\Delta t}\ln\left(1 - \frac{1}{6}\Delta t^3\lambda_{\pm}^3 + \frac{1}{2}\Delta t^2\lambda_{\pm}^2 - \Delta t\lambda_{\pm}\right). \tag{1.24c}$$

We then compare its Taylor series to the analytic values to get the lowest order error terms in Δx and Δt respectively. The results of this are given in Table 1.6.

Scheme	Lowest Order Term of Error	
	Δx	Δt
FDVM$_1$	$\frac{i\sqrt{gH}}{2}k^2\Delta x$	$-\frac{3igH}{6+2H^2k^2}k^2\Delta t$
FDVM$_2$	$-\frac{\sqrt{3gH}}{8\left(3+H^2k^2\right)^{3/2}}k^3\Delta x^2$	$\frac{\sqrt{3}}{2}\left(\frac{gH}{3+H^2k^2}\right)^{3/2}k^3\Delta t^2$
FEVM$_2$	$\frac{\sqrt{3gH}\left(14+5H^2k^2\right)}{80\left(3+H^2k^2\right)^{3/2}}k^3\Delta x^2$	$\frac{\sqrt{3}}{2}\left(\frac{gH}{3+H^2k^2}\right)^{3/2}k^3\Delta t^2$
FDVM$_3$	$\frac{i\sqrt{gH}}{12}k^4\Delta x^3$	$\frac{3ig^2H^2}{8\left(3+H^2k^2\right)^2}k^4\Delta t^3$

Table 1.6: Table showing lowest order error term for approximating ω_+ for all FDVM and the FEVM. Where the analytic value is $\omega_+ = k\sqrt{gH}\sqrt{\frac{3}{3+k^2H^2}}$. All cross terms, $\Delta x^a \Delta t^b$ had $a+b$ larger than the reported terms here.

Discussion

Tables 1.1, 1.2, 1.3 and 1.4 demonstrate that the basic operators all have the correct spatial order of accuracy or better. Consequently our approximation to the flux matrix $\mathbf{F}$ in Table 1.5 also has the correction spatial order of accuracy or better, and thus our schemes all have the correct spatial order of accuracy. Finally Table 1.6 demonstrates that all methods approximated the analytic dispersion relation with the expected order of accuracy in both space and time.

All our methods introduce some diffusive error for $\mathcal{F}^{\eta,\eta}$ and $\mathcal{F}^{G,\upsilon}$. This is due to the Kurganov approximation containing both a flux averaging part and a diffusive part, which are split for these linearised equations with $U = 0$. So that the off diagonal terms $\mathcal{F}^{G,\eta}$ and $\mathcal{F}^{\eta,\upsilon}$ are the flux average part while the diagonal terms $\mathcal{F}^{\eta,\eta}$ and $\mathcal{F}^{G,\upsilon}$ are the diffusive part. This shows up in the errors for these terms as even powers of Δx for the dispersive errors and odd powers of Δx for the diffusive errors.

Table 1.4 demonstrates that the second-order FEVM performs better than the second-order FDVM in solving the elliptic equation (1.4c) for $\upsilon_{j+1/2}$. Since H, k and Δx are all real and greater than 0, the lowest order error term for the second-order FEVM is always lower than the corresponding error for the second-order FDVM due to

$$\left|\frac{6+H^2k^2}{4H}\right| > \left|\frac{12+5H^2k^2}{40H}\right|$$

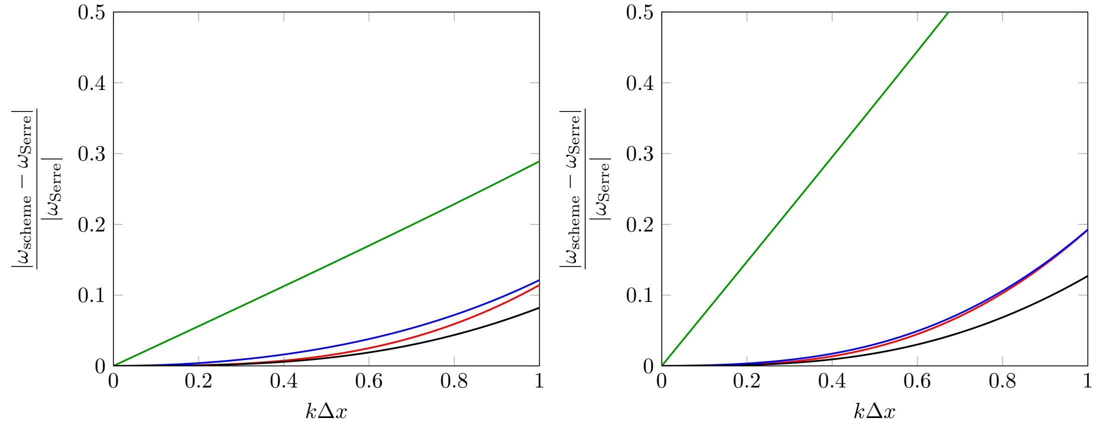
when $H > 0$ and $k > 0$. This leads to $\mathbf{F}$ being better approximated elementwise by the second-order FEVM than the second-order FDVM. However by performing the Runge-Kutta time stepping and calculating ω as we have done in Table 1.6 we observe that the second-order FDVM has a smaller error than the second-order FEVM. This is because the lowest order error term is the same for Δt but the error is smaller for Δx as

$$\left|\frac{14+5H^2k^2}{80}\right| > \left|\frac{1}{8}\right|$$

for all H and k values. Showing that even though the approximation of $\mathcal{G}$ for the second-order FEVM is better, this does not necessarily translate into a better overall method.

1.2.9 Results

From the basic factors presented in the Tables 1.1, 1.2, 1.3 and 1.4 the flux factors $\mathcal{F}^{\eta,\eta}$,$\mathcal{F}^{\eta,G}$, $\mathcal{F}^{G,\eta}$ and $\mathcal{F}^{G,G}$ can be calculated using (1.13) and (1.15) respectively.



$H = 1$ and $k = 0.5$ $H = 1$ and $k = 2.5$

Figure 1.1: Dispersion error for first-order FDVM (🟩), second-order FDVM(🟥), second-order FEVM (🟦) and third-order FDVM (⬛) with $r = 1/2$.

From there the matrix $\mathbf{F}$ can be computed using (1.16), and its eigenvalues found. Having found the eigenvalues we then substitute them into the appropriate equation given by the Runge-Kutta time stepping method which are all given in (1.24).

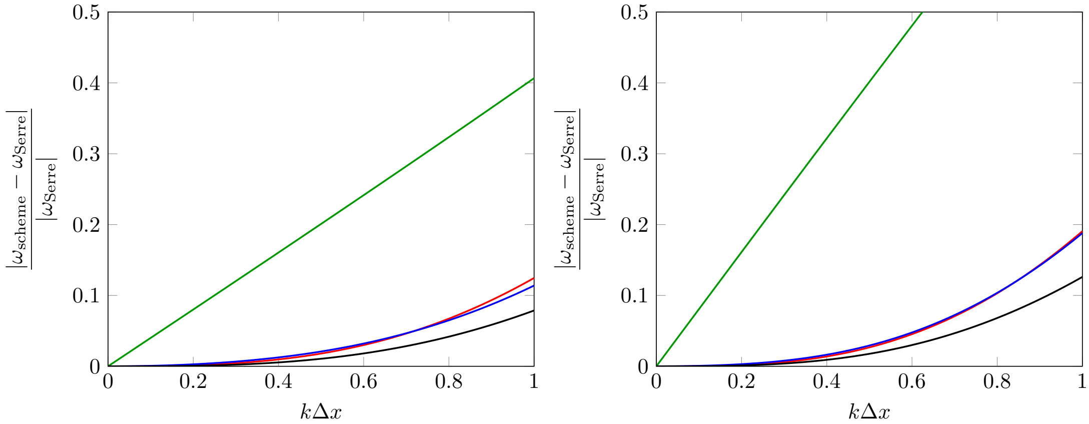
We did this numerically for various H and k values and observed the behaviour of the dispersion error as we varied Δx. With $\Delta t = \left(r/\sqrt{gH}\right)\Delta x$, so that we satisfy the CFL condition []. We present the results for $kH = 0.5$ and $kH = 2.5$ for $r = 1/2$ in Figure 1.1 and for $r = 1/4$ in Figure 1.2.

These values for kH were chosen because of their use in [3], because their results are representative for all kH values and finally because they cover the physical situations we will be interested in throughout this thesis.

From Figures 1.1 and 1.2 we can see that as expected increasing the resolution of our numerical methods decreases the dispersion error of the numerical scheme. While increasing the order of accuracy of the scheme decreases the dispersion error.

Comparing Figures 1.1 and 1.2 we can see that lower values of r actually lead to an increase in the dispersion error, most significantly for the first-order FDVM but also for the other methods.

When $r = 1/2$ the second-order FDVM consistently outperforms the FEVM for $k\Delta x \leq 1$. This matches well with what we expect given the lowest order error term for ω in Table 1.6, which tells us that when Δx is small, and thus the lowest order error term is dominant, that our second-order FDVM should have a smaller error than the FEVM.



$H = 1$ and $k = 0.5$

$H = 1$ and $k = 2.5$

Figure 1.2: Dispersion error for first-order FDVM (🟩), second-order FDVM(🟥), second-order FEVM (🟦) and third-order FDVM (⬛) with $r = 1/4$.

For other choices of Δt which satisfy the CFL condition such as $r = 1/4$ as in Figure 1.2, the dominance of the second-order FDVM no longer occurs for all $k\Delta x \le 1$. We found that for $1/2 \le r \le 1$ the second-order FDVM has a lower dispersion error than the FEVM for all $k\Delta x \le 1$. While for $0 < r \le 2/5$ the FEVM had a lower disperison error for $k\Delta x$ values close to 1 as in Figure 1.1. In our numerical experiments we most often choose $r = 1/2$ and so we expect the second-order FDVM to have better dispersion properties for our numerical experiments, although the difference is small.

Our results compare well with those of [3] who performed a similar analysis for their numerical method applied to the linearised Serre equations with $U = 0$. We have extended their results by combining the spatial and temporal contribution to the dispersion relation and performing it on a different numerical method.

For a fixed r these plots only depend on the parameter kH. This parameter kH is proportional to the shallowness parameter σ with $2\pi\sigma = kH$. So for $kH = 2.5$ the water is no longer shallow and the Serre equations are not an appropriate model for water waves, although our results demonstrate that our numerical methods will have a small dispersion error in this case. In general, we also find that as kH is increased our numerical methods perform worse generally although the dispersion error still converges to 0 as $\Delta x \to 0$.

1.3 Von Neumann Stability

A scheme is said to have Von Neumann stability if its evolution matrix has a spectral radius less than or equal to 1. In the dispersion relation analysis above we demonstrated how to calculate $\mathbf{E}$ for the second-order FEVM (1.18). Likewise the above work also demonstrated how to obtain the evolution matrix for the FDVM as well, to summarise we have

- First-order FDVM
$$\mathbf{E} = \mathbf{I} - \Delta t \mathbf{F}.$$
- Second-order FDVM and FEVM
$$\mathbf{E} = \mathbf{I} - \Delta t \mathbf{F} + \frac{1}{2} \Delta t^2 \mathbf{F}^2.$$
- Third-order FDVM
$$\mathbf{E} = \mathbf{I} - \Delta t \mathbf{F} + \frac{1}{2} \Delta t^2 \mathbf{F}^2 - \frac{1}{6} \Delta t^3 \mathbf{F}^3.$$

Where $\mathbf{F}$ is the appropriate flux matrix for the method. If the spectral radius of all these evolution matrices are less than or equal to one then all our FDVM and FEVM possess Von Neumann stability for the lienarised Serre equation with $U = 0$.

1.3.1 Evolution Matrix for Finite Difference Methods with $U \neq 0$

The other methods described and used in this thesis, the two finite difference methods $\mathcal{D}$ and $\mathcal{W}$ require a slightly different handling to obtain their evolution matrix. We also extended the stability analysis for these two methods to non-zero values of U.

Due to these differences we will demonstrate how we obtained the evolution matrix for the naive second-order finite difference method $\mathcal{D}$. Having done this we will then just present the evolution matrix we obtained for the second-order finite difference/Lax-Wendroff method $\mathcal{W}$.

Naive Second-Order Finite Difference Method $\mathcal{D}$

The numerical method $\mathcal{D}$ is obtained by replacing all the derivatives in (1.3) with their second-order finite difference approximations. For the linearised Serre equation (1.3) $\mathcal{D}$ is

$$\eta_j^{n+1} = \eta_j^{n-1} - \Delta t \left(U \frac{-\eta_{j-1}^n + \eta_{j+1}^n}{\Delta x} + H \frac{-\upsilon_{j-1}^n + \upsilon_{j+1}^n}{\Delta x} \right), \tag{1.25a}$$

$$\begin{aligned} \upsilon_j^{n+1} - \frac{H^2}{3} \frac{\upsilon_{j-1}^{n+1} - 2\upsilon_j^{n+1} + \upsilon_{j+1}^{n+1}}{\Delta x^2} &= \upsilon_j^{n-1} - \frac{H^2}{3} \frac{\upsilon_{j-1}^{n-1} - 2\upsilon_j^{n-1} + \upsilon_{j+1}^{n-1}}{\Delta x^2} \\ &+ \Delta t \Bigg(- g \frac{-\eta_{j-1}^n + \eta_{j+1}^n}{\Delta x} - U \frac{-\upsilon_{j-1}^n + \upsilon_{j+1}^n}{\Delta x} \\ &\qquad + \frac{H^2}{3} \left(U \frac{-\upsilon_{j-2}^n + 2\upsilon_{j-1}^n - 2\upsilon_{j+1}^n + \upsilon_{j+2}^n}{\Delta x^3} \right) \Bigg) \end{aligned} \tag{1.25b}$$

where again δ has been absorbed into the corresponding η and υ terms.

We will again assume both η and υ are Fourier modes (1.6). In $\mathcal{D}$ we only need to simplify 3 finite differences. For a general quantity q using (1.7) we have

$$\frac{-q_{j-1}^n + q_{j+1}^n}{2\Delta x} = \frac{i \sin(k\Delta x)}{\Delta x} q_j^n \tag{1.26a}$$

$$\frac{q_{j-1}^n - 2q_j^n + q_{j+1}^n}{\Delta x^2} = \frac{2\cos(k\Delta x) - 2}{\Delta x^2} q_j^n \tag{1.26b}$$

$$\begin{aligned} \frac{-q_{j-2}^n + 2q_{j-1}^n - 2q_{j+1}^n + q_{j+2}^n}{2\Delta x^3} &= -4i \sin(k\Delta x) \frac{\sin^2\left(\frac{k\Delta x}{2}\right)}{\Delta x^3} q_j^n \\ &= \frac{i \sin(k\Delta x)}{\Delta x} \frac{2\cos(k\Delta x) - 2}{\Delta x^2} q_j^n. \end{aligned} \tag{1.26c}$$

Substituting (1.26) into (1.25) we get that

$$\eta_j^{n+1} = \eta_j^{n-1} - \Delta t \left(U \frac{i \sin(k\Delta x)}{\Delta x} \eta_j^n + H \frac{i \sin(k\Delta x)}{\Delta x} \upsilon_j^n \right),$$

$$\begin{aligned} \upsilon_j^{n+1} = \upsilon_j^{n-1} - \frac{3\Delta x^2 \Delta t}{3\Delta x^2 - 2H^2(\cos(k\Delta x) - 1)} \left(g \frac{i \sin(k\Delta x)}{\Delta x} \right) \eta_j^n \\ + U \frac{i \Delta t \sin(k\Delta x)}{\Delta x} \upsilon_j^n. \end{aligned}$$

We can rewrite this in matrix form as

$$\begin{bmatrix} \eta_j^{n+1} \\ \upsilon_j^{n+1} \\ \eta_j^n \\ \upsilon_j^n \end{bmatrix} = \mathbf{E} \begin{bmatrix} \eta_j^n \\ \upsilon_j^n \\ \eta_j^{n-1} \\ \upsilon_j^{n-1} \end{bmatrix}, \tag{1.27}$$

where

$$\mathbf{E} = \begin{bmatrix} -\frac{2i\Delta t}{\Delta x}U\sin\left(k\Delta x\right) & -\frac{2i\Delta t}{\Delta x}H\sin\left(k\Delta x\right) & 1 & 0 \\ -\frac{6gi\Delta x\Delta t}{3\Delta x^2 - 2H^2\left(\cos\left(k\Delta x\right)-1\right)}\sin\left(k\Delta x\right) & -\frac{2i\Delta t}{\Delta x}U\sin\left(k\Delta x\right) & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}.$$

Lax-Wendroff Method $\mathcal{W}$

After performing the same process for the finite difference and Lax Wendroff method $\mathcal{W}$ (**??**) we get the following matrix equation

$$\begin{bmatrix} \eta_j^{n+1} \\ \upsilon_j^{n+1} \\ \eta_j^{n} \\ \upsilon_j^{n} \end{bmatrix} = \mathbf{E} \begin{bmatrix} \eta_j^{n} \\ \upsilon_j^{n} \\ \eta_j^{n-1} \\ \upsilon_j^{n-1} \end{bmatrix}. \tag{1.28}$$

where

$$\mathbf{E} = \begin{bmatrix} E^{0,0} & E^{0,1} & 0 & -\frac{\Delta t}{\Delta x}H^{\frac{i\sin(k\Delta x)}{2}} \\ E^{1,0} & -\frac{2i\Delta t}{\Delta x}U\sin\left(k\Delta x\right) & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \tag{1.29}$$

with

$$\begin{aligned} E^{0,0} &= 1 - \frac{\Delta t}{\Delta x}\left(-\frac{6gi\Delta x\Delta t}{3\Delta x^2 - 2H^2\left(\cos\left(k\Delta x\right)-1\right)}\sin\left(k\Delta x\right)\right)H\frac{i\sin\left(k\Delta x\right)}{2} \\ &- \frac{\Delta t}{\Delta x}U\left(\left(i\sin\left(k\Delta x\right)\right) - \frac{\Delta t}{\Delta x}U\left(\cos\left(k\Delta x\right)-1\right)\right), \\ E^{0,1} &= -\frac{\Delta t}{\Delta x}\left[H\frac{i\sin\left(k\Delta x\right)}{2}\left(1 - \frac{2i\Delta t}{\Delta x}U\sin\left(k\Delta x\right)\right) - U\left(\frac{\Delta t}{\Delta x}H\left(\cos\left(k\Delta x\right)-1\right)\right)\right], \\ E^{1,0} &= -\frac{6gi\Delta x\Delta t}{3\Delta x^2 - 2H^2\left(\cos\left(k\Delta x\right)-1\right)}\sin\left(k\Delta x\right). \end{aligned}$$

1.3.2 Results

We will demonstrate that these methods possess Von Neumann stability numerically. We do this by calculating the spectral radius of the evolution matrices numerically for fixed H and k values and demonstrate the behaviour of this spectral radius as Δx changes. We use the CFL condition to determine Δt given Δx, and in particular we again have $\Delta t = \left(r / \left(U + \sqrt{gH}\right)\right) \Delta x$ with $r = 1/2$. We first show the results for $U = 0$ where we demonstrate the stability of all the methods. We then allow various U values and therefore only present the results for the finite difference methods.

Quiescent Fluid $U = 0$

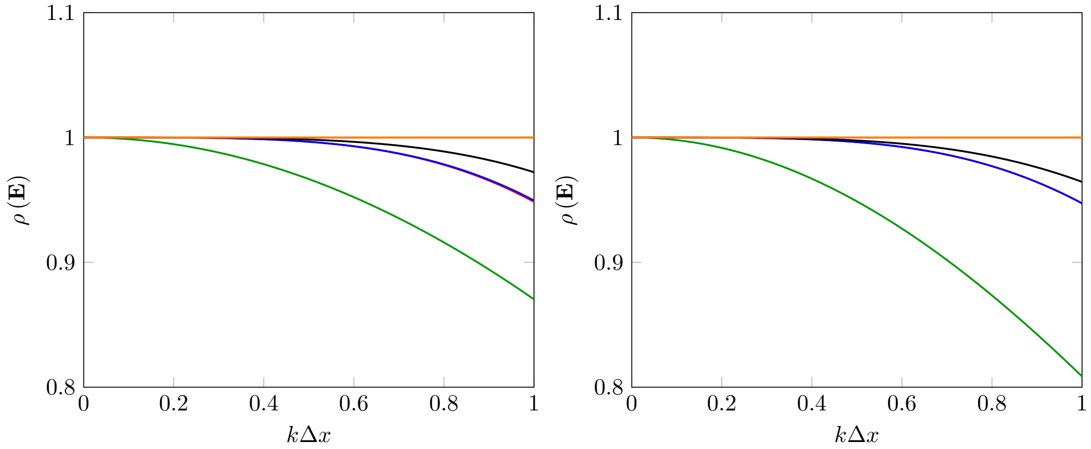
This is the situation in which we are most interested in for the purposes of ocean modelling. Most of the numerical experiments we perform later will occur in this region where the water is still with waves propagating on top. This is also the scenario in which the evolution matrices for the FDVM and FEVM were calculated and thus we can only demonstrate the stability of all methods in this region.

The spectral radius for a range of Δx values was plotted in Figure 1.3 for all numerical methods in this thesis. The representative values of $kH = 0.5$ and $kH = 2.5$ were chosen for the stability results plotted in Figure 1.1 due to their use in the dispersion error analysis. These values were representative of the results we observed for other scenarios and these values cover the range of physical scenarios we are interested in.

Our results demonstrate that all numerical methods satisfy the stability condition for a range of kH values with $U = 0$ as all methods have growth matrices with spectral radius less than or equal to 1. Indeed this is what we found generally for all these methods for all our investigated values of hK when $0 < r \leq 1$. This is the expected result given that when $0 < r \leq 1$ the CFL condition is satisfied.

We note that both the second-order FDVM (−) and FEVM (−) have very similar spectral radius values and their plots overlap so that only the curve for the FEVM (−) is visible. We also observe similar behaviour for the two second-order finite difference methods $\mathcal{D}$ (−) and $\mathcal{W}$ (−) so that only the curve for $\mathcal{W}$ (−) is visible.

We observe that the spectral radius for the second-order finite difference methods $\mathcal{D}$ and $\mathcal{W}$ are consistently 1 when $U = 0$ for various hK and $k\Delta x$ values with $r \leq 1$. This can be seen in Table 1.7, where the average of the spectral radius for $\mathcal{D}$ and $\mathcal{W}$ over various $k\Delta x$ values is 1 plus a number which is just the accumulation of round-off errors caused by performing this analysis numerically.



$U = 0$, $H = 1$ and $k = 0.5$ $U = 0$, $H = 1$ and $k = 2.5$

Figure 1.3: Spectral radius of growth matrix $\mathbf{E}$ for first-order FDVM (🟩), second-order FDVM(🟥), second-order FEVM (🟦), third-order FDVM (⬛), $\mathcal{D}$ (🟪) and $\mathcal{W}$ (🟧) .

Method	kH	Average
$\mathcal{D}$	0.5	$1 + 4 \times 10^{-16}$
$\mathcal{D}$	2.5	$1 + 4 \times 10^{-16}$
$\mathcal{W}$	0.5	$1 + 4 \times 10^{-16}$
$\mathcal{W}$	2.5	$1 + 4 \times 10^{-16}$

Table 1.7: Average of $\rho(\mathbf{E})$ over all Δx values for the second-order finite difference methods when $U = 0$.

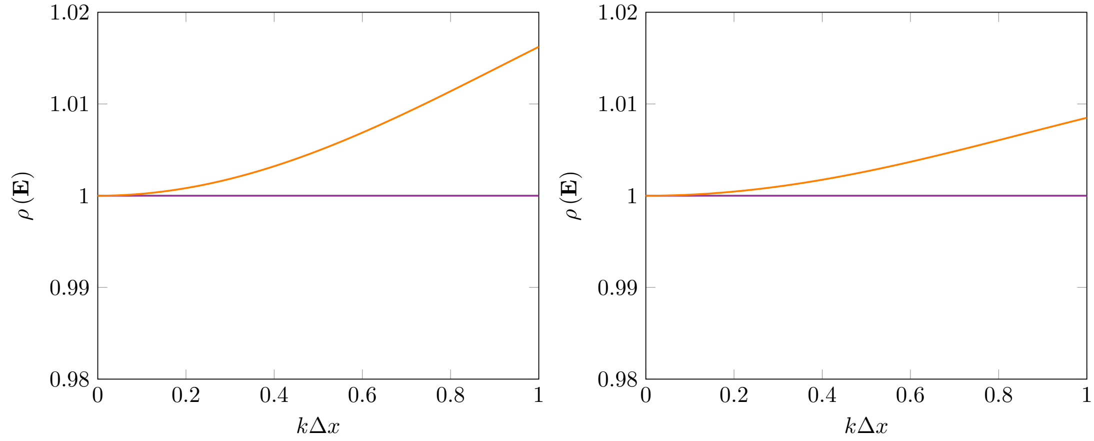
Non-zero Mean Flow

Only the finite difference methods $\mathcal{W}$ and $\mathcal{D}$ for the linearised Serre equations allow for nonzero values of U. We investigated the behaviour of the spectral radius of the growth matrix for various values of U, kH and Δx. Again we have chosen $\Delta t = \left(r / \left(U + \sqrt{gH}\right)\right) \Delta x$ with $r = 1/2$ to satisfy the CFL condition []. We present the results for $U = 1$ with $kH = 0.5$ and 2.5 in Figure 1.4. These values were chosen because they are representative of the behaviour for both methods for most values of U and kH, and because these values of kH match those used in the results above.

These results demonstrate that the naive second-order method $\mathcal{D}$ is still stable even with a background mean flow, with a spectral radius that is consistently 1. This is demonstrated in Table 1.8 as well where the average spectral radius is 1 plus a number that is just round-off error. This behaviour was consistent for various U, kH and Δx values provided $0 < r \leq 1$. Therefore this method is stable as desired for a range of flow scenarios.

Unfortunately the finite difference/Lax-Wendroff method $\mathcal{W}$ is no longer stable anywhere with growth factors that are consistently larger than 1 although it approaches stability as $\Delta x \to 0$. This is evident in Table 1.8 where the average spectral radius is larger than 1 by significantly more than round-off error. By modifying the parameters we can increase the spectral radius of the evolution matrix for $\mathcal{W}$ as desired. The largest $k\Delta x$ value appears to correspond to our largest spectral radius. However, the interaction between the spectral radius, hK and U is not so obvious. Since the spectral radius was consistently larger than 1 when $U \neq 0$ this means the Lax-Wendroff method is not stable unless $U = 0$. Although the growth factors are only marginally greater than 1 for most situations and so the instabilities may not be apparent when performing numerical experiments, as we demonstrate in Chapter [].

We also investigated different relationships between Δt and Δx, such as $\Delta t \propto \Delta x^a$ but could not find a reasonable a that gave us stability for $\mathcal{W}$ when $|U| > 0$ across a range of $k\Delta x$ values.



$U = 1,\ H = 1$ and $k = 0.5$

$U = 1,\ H = 1$ and $k = 2.5$

Figure 1.4: Spectral radius of growth matrix $\mathbf{E}$ for $\mathcal{D}$ (🟪) and $\mathcal{W}$ (🟧) .

Method	kH	Average
$\mathcal{D}$	0.5	$1 + 4 \times 10^{-16}$
$\mathcal{D}$	2.5	$1 + 4 \times 10^{-16}$
$\mathcal{W}$	0.5	$1 + 6 \times 10^{-3}$
$\mathcal{W}$	2.5	$1 + 3 \times 10^{-3}$

Table 1.8: Average of $\rho(\mathbf{E})$ over all Δx values for the second-order finite difference methods when $U = 1$.

Bibliography

[1] J. G. Charney, R. Fjörtoft, and J. v. Neumann. Numerical integration of the barotropic vorticity equation. *Tellus*, 2(4):237–254, 1950.

[2] G. El, R. H. J. Grimshaw, and N. F. Smyth. Unsteady undular bores in fully nonlinear shallow-water theory. *Physics of Fluids*, 18(2):027104, 2006.

[3] A. G. Filippini, M. Kazolea, and M. Ricchiuto. A flexible genuinely nonlinear approach for nonlinear wave propagation, breaking and run-up. *Journal of Computational Physics*, 310:381–417, 2016.

[4] A. Kurganov, S. Noelle, and G. Petrova. Semidiscrete central-upwind schemes for hyperbolic conservation laws and Hamilton-Jacobi equations. *Journal of Scientific Computing, Society for Industrial and Applied Mathematics*, 23(3):707–740, 2002.

[5] P. D. Lax and R. D. Richtmyer. Survey of the stability of linear finite difference equations. *Communications on pure and applied mathematics*, 9(2):267–293, 1956.