Chapter 1

Analysis of Numerical Methods

The most important property of a numerical method is convergence, because convergence guaranties that as we increase the spatial and temporal resolution of a numerical method, the numerical methods solutions better approximate the solutions of the differential equations. Which is precisely the desired property for a given numerical method. For linear partial differential equations the Lax-equivalence theorem demonstrates that a numerical method is convergent if and only if it is stable and consistent []. 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 the current time step.

The convergence of the finite difference volume methods and the finite element volume methods are inherited from the convergence of the finite volume methods at their core, which were shown to be convergent []. However, the convergence of the finite difference methods has not been shown for this particular equation and so we demonstrate that here. The consistency of the finite difference methods is very straightforward, given that we are using well tested approximations and so we instead focus on demonstrating stability for these methods. In particular we demonstrate von Nuemann stability [] for the linearised Serre equations.

Having demonstrated the convergence of a particular method there are other properties that can be used to discriminate between two convergent numerical methods in order to pick the most appropriate numerical method for a particular differential equation. Such properties of numerical methods are: the order of accuracy and the computationally efficiency. For our methods we demonstrate the order of accuracy in [] using both analytical and forced solutions.

The Serre equations can be written in conservation law form therefore their

solutions conserve mass h, momentum uh, irrotationality G [] and the Hamiltonian $\mathcal{H}$. The conservation of these quantities for the solutions of the numerical methods is investigated in chapter [].

The Serre equations are of particular interest for water wave modelling as they possess a dispersion relation that approximates the dispersion relation for the Euler equations well. Since the dispersion relation determines the speed of propagation for waves, approximating it well is of particular interest for tsunami modelling. Therefore knowing how well a given numerical methods dispersion relation approximates the dispersion relation of the Serre equations is an important question for our numerical methods as they are intended to be used for modelling tsunamis. For this reason we wish to know the error in the dispersion relation for all the numerical methods we are interested in; the finite difference volume methods and the finite element volume methods.

In this chapter we will focus on analysing the stability and the dispersion relation of our numerical methods. We group these two analyses together as they employ very similar methods and require some working. We shall leave the rest of the analyses of the numerical methods in this thesis to chapters [] where we investigate the solutions of the numerical methods.

Both the dispersion relation and stability analysis are performed on the linearised Serre equation with horizontal beds. We neglect bed terms because our finite difference methods neglect them too and because the dispersion relation has no contribution from the bed term. We also focus on the linearised Serre equations as a first step for the stability analysis and those are the equations from which the dispersion relation is derived. We begin by giving the linearised Serre equations with a horizontal bed.

1.0.1 Linearised Serre equations with horizontal bed

The Serre equations with a horizontal bed were given earlier (**??**) and we present them here to remind the reader

$$\frac{\partial h}{\partial t} + \frac{\partial (uh)}{\partial x} = 0$$

and

$$\frac{\partial (uh)}{\partial t} + \frac{\partial}{\partial x}\left(u^2 h + \frac{gh^2}{2} + \frac{h^3}{3}\Phi \right) = 0.$$

To linearise these equations we assume that we are modelling waves which are perturbations ontop of a flow with a mean height H and a mean flow velocity U so that

$$h(x,t) = H + \delta\eta(x,t) + \mathcal{O}\left(\delta^2\right), \tag{1.1}$$
$$u(x,t) = U + \delta\upsilon(x,t) + \mathcal{O}\left(\delta^2\right). \tag{1.2}$$

Where $\delta \ll 1$, so that we are modelling relatively small waves where terms of order δ^2 are negligible. We substitute this into (**??**) and (**??**) and neglect terms of order δ^2 to obtain the linearised Serre equations with horizontal beds

$$\frac{\partial\eta}{\partial t} + H\frac{\partial\upsilon}{\partial x} + U\frac{\partial\eta}{\partial x} + = 0 \tag{1.3a}$$

and

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

where the linearised irrotationality is

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

1.1 Dispersion Error

1.1.1 Background and Assumptions

To study the error in the dispersion relation caused by the numerical methods we will follow the work of [1] who used a range of numerical methods on a different reformulation of the Serre equations. Therefore as they do in that paper we will assume that $U = 0$, so that there is no mean flow velocity. This is a reasonable simplification because we are interested in modelling waves on quiescent water, such as tsunamis. Additionally the more complicated term in the dispersion relation to approximate is the term which determines the speed of waves on still water, and not the contribution from U.

By assuming that $U = 0$ the equations (1.3) and (1.3c) reduce to

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

and

$$h_0\frac{\partial\upsilon}{\partial t} + gH\frac{\partial\eta}{\partial x} - \frac{H^3}{3}\left(\frac{\partial^3\upsilon}{\partial x^3\partial t}\right) = 0 \tag{1.4b}$$

with

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

The linearised equations (1.4) can be reformulated into equations with η and G as conserved variables as in (**??**) to obtain

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

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

These will be the equations we will apply our numerical methods to in order to assess the error in dispersion that they introduce. It can be seen that these equations are linear conservation equations for η and G, and so our FDVM and FEVM methods are applicable to them.

The final assumption of this analysis is that we 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 is given by

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

This is precisely the assumption made to derive the dispersion relation of the linearised Serre equation as well. A consequence of these quantities being Fourier modes and our use of fixed temporal and spatial grids is that

$$q^n_{j\pm l} = q^n_j e^{\pm ikl\Delta x} \qquad \text{and} \qquad q^{n\pm l}_j = q^n_j e^{\pm i\omega l \Delta t} \tag{1.7}$$

1.1.2 Overview of the analysis

We will now present a brief overview of how this analysis progresses for a single evolution step of a FDVM. We do this because at the outset we wish to explain the process, so that the example we demonstrate is more illuminating. The dispersion analysis also extends to the Runge-Kutta steps used to make the schemes the appropriate temporal order of accuracy, however we believe the example provided in the relevant section demonstrates the process well without the need for an overview.

For the FDVM the evolution step progresses like so

1. We possess the vectors of the cell averages $\bar{\boldsymbol{\eta}}$ and $\bar{\boldsymbol{G}}$ at the current time

2. We use the inverse of the transformation $\mathcal{M}$ we calculate $\boldsymbol{\eta}$ and $\boldsymbol{G}$ from $\bar{\boldsymbol{\eta}}$ and $\bar{\boldsymbol{G}}$ where

$$\begin{aligned}\boldsymbol{\eta} &= \mathcal{M}^{-1}\left(\bar{\boldsymbol{\eta}}\right)\\ \boldsymbol{G} &= \mathcal{M}^{-1}\left(\bar{\boldsymbol{G}}\right)\end{aligned}$$

3. We use the elliptic solver $\mathcal{G}^{-1}$ to calculate $\boldsymbol{\upsilon}$ from H and $\boldsymbol{G}$

$$\boldsymbol{\upsilon} = \mathcal{G}^{-1}\left(H, \boldsymbol{G}\right)$$

4. We reconstruct η and G at $x^{-}_{j+1/2}$ and $x^{+}_{j+1/2}$ from the cell average values using $\mathcal{R}^{-}$ and $\mathcal{R}^{+}$ while reconstructing υ at $x_{j+1/2}$ from the nodal values using $\mathcal{R}^{\upsilon}$ for all cell edges.

$$\begin{aligned}\eta^{-}_{j+1/2} &= \mathcal{R}^{-}\left(\bar{\boldsymbol{\eta}}\right) & \qquad G^{-}_{j+1/2} &= \mathcal{R}^{-}\left(\bar{\boldsymbol{G}}\right)\\ \eta^{+}_{j+1/2} &= \mathcal{R}^{+}\left(\bar{\boldsymbol{\eta}}\right) & \qquad G^{+}_{j+1/2} &= \mathcal{R}^{+}\left(\bar{\boldsymbol{G}}\right)\\ \upsilon_{j+1/2} &= \mathcal{R}^{\upsilon}\left(\boldsymbol{\upsilon}\right)\end{aligned}$$

5. We calculate $F_{j+1/2}$ using $\mathcal{F}$ for each j

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

6. We use update formula (**??**) to calculate $\bar{\boldsymbol{\eta}}$ and $\bar{\boldsymbol{G}}$ at the next time

All these operators named in the above list are linear combinations of the quantities at different grid points. Together with (1.7) we can turn our operators into just constant coefficients, so that for example we have in the case of the elliptic operator

$$G_j = \mathcal{G} \times u_j.$$

All these coefficients are combined in the final step to give the constant matrix $\mathbf{F}$ that calculates the flux at the current time allowing us to write the update formula [] as

$$\begin{bmatrix}\eta\\ \upsilon\end{bmatrix}^{n+1}_{j} = \left(\mathbf{I} - \Delta t\mathbf{F}\right)\begin{bmatrix}\eta\\ \upsilon\end{bmatrix}^{n}_{j}$$

in terms of primitive variables. From this equation dispersion relation of the method can be calculated, accounting for the Runge-Kutta steps, which we demonstrate how to handle later [].

We will now present an example of this analysis for the second-order FDVM and FEVM. Although both methods are very similar, we present the FEVM as well as the FDVM because the finite element method requires a bit more work than the finite difference method and we wished to show how it was handled, so as to be repeatable. In this analysis we break the evolution step up into the three parts the elliptic equation which relates G to υ, the evolution equation we use to update η and G and the Runge-Kutta steps we use to increase the order of accuracy of the method in time. We begin with the elliptic equation.

1.1.3 Elliptic Equation

For both the finite difference and the finite element method we are trying to determine the coefficient $\mathcal{G}$ from the elliptic equation for the linearised Serre equation with $U = 0$ (1.4c) such that

$$G_j = \mathcal{G}\upsilon_j.$$

We will use subscripts to denote the method that generates a particular coefficient. For this analysis we will demonstrate how $\mathcal{G}_{FD2}$ and $\mathcal{G}_{FEM2}$ are generated by the finite difference method and the finite element method respectively. At the end we will also give $\mathcal{G}_{FD4}$ from the fourth order finite difference method and $\mathcal{G}_A$ which is the analytic value. We begin with the second-order finite difference method.

Finite Difference ($\mathcal{G}_2$)

The elliptic equation at a particular grid point x_j is

$$G_j = H\upsilon_j - \frac{H^3}{3}\left(\frac{\partial^2\upsilon}{\partial x^2}\right)_j.$$

For the second-order finite difference method the derivative $\frac{\partial^2\upsilon}{\partial x^2}$ is approximated by

$$\left(\frac{\partial^2\upsilon}{\partial x^2}\right)_j = \frac{\upsilon_{j-1} - 2\upsilon_j + \upsilon_{j+1}}{\Delta x^2}.$$

Making use of (1.7) this becomes

$$\left(\frac{\partial^2\upsilon}{\partial x^2}\right)_j = \frac{\upsilon_j e^{-ik\Delta x} - 2\upsilon_j + \upsilon_j e^{ik\Delta x}}{\Delta x^2}.$$

Which reduces to

$$\left(\frac{\partial^2 \upsilon}{\partial x^2}\right)_j = \frac{2\cos(k\Delta x) - 2}{\Delta x^2}\upsilon_j.$$

Substituting this approximation into our elliptic equation one obtains

$$G_j = \left(H - \frac{H^3}{3}\frac{2\cos(k\Delta x) - 2}{\Delta x^2}\right)\upsilon_j.$$

Therefore we have an equation which is independent of j which gives the second order finite differences transformation between G_j and υ_j as desired.

In particular for the centred second-order finite difference method we have

$$\mathcal{G}_{FD2} = \left(H - \frac{H^3}{3}\frac{2\cos(k\Delta x) - 2}{\Delta x^2}\right).$$

The process to calculate $\mathcal{G}_{FD4}$ is very similar and so we omit it for brevity.

Finite Element Method

Since finite difference methods are all very similar in how the error coefficient is found, it is sufficient to just show one example of the process used. However, because the process for the finite element method is different we present the working for it here. Here as with the finite difference method we desire a coefficient independent of j, which we call $\mathcal{G}_{FEM2}$. However, unlike the finite difference case this does not come from an equation relating G_j and u_j, but instead an equation between G_j and $u_{j+1/2}$. This is because our finite element method calculates $u_{j+1/2}$ directly, and we use this value without a reconstruction in the flux calculation.

To attain this equation we being with the matrix equation of the FEM for the linearised equations (1.4c), which is simpler to obtain than the full Serre equations presented earlier and is

$$\sum_j \frac{\Delta x}{6}\begin{bmatrix} G^+_{j-1/2} \\ 2G^+_{j-1/2} + 2G^-_{j+1/2} \\ G^-_{j+1/2} \end{bmatrix} = \sum_j \left(H\frac{\Delta x}{30}\begin{bmatrix} 4 & 2 & -1 \\ 2 & 16 & 2 \\ -1 & 2 & 4 \end{bmatrix} + \frac{H^3}{9\Delta x}\begin{bmatrix} 7 & -8 & 1 \\ -8 & 16 & -8 \\ 1 & -8 & 7 \end{bmatrix}\right)\begin{bmatrix} \upsilon_{j-1/2} \\ \upsilon_j \\ \upsilon_{j+1/2} \end{bmatrix}.$$

Because we have demonstrate how to calculate this matrix equation for the full Serre equations we omit the working from which this equation is derived.

Using our relations from the periodic nature of υ and G, and the minmod reconstruction used on G which will be given later [] we get that

$$\sum_j \frac{\Delta x}{6}\begin{bmatrix} e^{-ik\Delta x}\mathcal{R}_2^+ \\ 2e^{-ik\Delta x}\mathcal{R}_2^+ + 2\mathcal{R}_2^- \\ \mathcal{R}_2^- \end{bmatrix} G_j = \sum_j \left(H\frac{\Delta x}{30}\begin{bmatrix} 4e^{-ik\frac{\Delta x}{2}} + 2 - e^{ik\frac{\Delta x}{2}} \\ 2e^{-ik\frac{\Delta x}{2}} + 16 + 2e^{ik\frac{\Delta x}{2}} \\ -e^{-ik\frac{\Delta x}{2}} + 2 + 4e^{ik\frac{\Delta x}{2}} \end{bmatrix} + \frac{H^3}{9\Delta x}\begin{bmatrix} 7e^{-ik\frac{\Delta x}{2}} - 8 + e^{ik\frac{\Delta x}{2}} \\ -8e^{-ik\frac{\Delta x}{2}} + 16 - 8e^{ik\frac{\Delta x}{2}} \\ e^{-ik\frac{\Delta x}{2}} - 8 + 7e^{ik\frac{\Delta x}{2}} \end{bmatrix}\right)\upsilon_j$$

We can now add all the terms that overlap i.e the extra contributions from the functions $\phi_{j+1/2}$ and $\phi_{j-1/2}$ from outside the cell $\left[x_{j-1/2}, x_{j+1/2}\right]$, this then gives us a relation between the sub-vectors of the total vectors of the FEM. Doing this we can rewrite the matrix equation as []

$$\sum_j \frac{\Delta x}{6}\begin{bmatrix} 2 \\ \mathcal{R}_2^- + \mathcal{R}_2^+ \end{bmatrix}^T \begin{bmatrix} G_j \\ G_j \end{bmatrix} = \sum_j \left(H\frac{\Delta x}{30}\begin{bmatrix} 2e^{-ik\frac{\Delta x}{2}} + 16 + 2e^{ik\frac{\Delta x}{2}} \\ -e^{-ik\frac{\Delta x}{2}} + 2 + 4e^{ik\frac{\Delta x}{2}} + e^{ik\Delta x}\left(4e^{-ik\frac{\Delta x}{2}} + 2 - e^{ik\frac{\Delta x}{2}}\right) \end{bmatrix}^T + \frac{H^3}{9\Delta x}\begin{bmatrix} -8e^{-ik\frac{\Delta x}{2}} + 16 - 8e^{ik\frac{\Delta x}{2}} \\ e^{-ik\frac{\Delta x}{2}} - 8 + 7e^{ik\frac{\Delta x}{2}} + e^{ik\Delta x}\left(7e^{-ik\frac{\Delta x}{2}} - 8 + e^{ik\frac{\Delta x}{2}}\right) \end{bmatrix}^T\right)\begin{bmatrix} \upsilon_j \\ \upsilon_{j+1/2} \end{bmatrix}$$

Which reduces to

$$\sum_j \frac{\Delta x}{6}\begin{bmatrix} 2 \\ \mathcal{R}_2^- + \mathcal{R}_2^+ \end{bmatrix}^T \begin{bmatrix} G_j \\ G_j \end{bmatrix} = \sum_j \left(H\frac{\Delta x}{30}\begin{bmatrix} 16 + 4\cos\left(\frac{k\Delta x}{2}\right) \\ 2e^{ik\frac{\Delta x}{2}}\left(2\cos\left(\frac{k\Delta x}{2}\right) - \cos\left(k\Delta x\right) + 4\right) \end{bmatrix}^T + \frac{H^3}{9\Delta x}\begin{bmatrix} 16 - 16\cos\left(\frac{k\Delta x}{2}\right) \\ -8e^{ik\frac{\Delta x}{2}}\sin^2\left(\frac{k\Delta x}{4}\right)\left(\cos\left(\frac{k\Delta x}{2}\right) - 3\right) \end{bmatrix}^T\right)\begin{bmatrix} \upsilon_j \\ \upsilon_j \end{bmatrix}$$

So the equation for $\upsilon_{j+1/2}$ is

$$\frac{\Delta x}{6}\left(\mathcal{R}_2^+ + \mathcal{R}_2^-\right) G_j = \left(H\frac{\Delta x}{30}\left(2e^{ik\frac{\Delta x}{2}}\left(2\cos\left(\frac{k\Delta x}{2}\right) - \cos\left(k\Delta x\right) + 4\right)\right) + \frac{H^3}{9\Delta x}\left(-8e^{ik\frac{\Delta x}{2}}\sin^2\left(\frac{k\Delta x}{4}\right)\left(\cos\left(\frac{k\Delta x}{2}\right) - 3\right)\right)\right)\upsilon_j$$

so we have

$$G_j = \frac{6}{\Delta x}\frac{1}{\mathcal{R}_2^+ + \mathcal{R}_2^-} \times \left(H\frac{\Delta x}{30}\left(2e^{ik\frac{\Delta x}{2}}\left(2\cos\left(\frac{k\Delta x}{2}\right) - \cos\left(k\Delta x\right) + 4\right)\right) + \frac{H^3}{9\Delta x}\left(-8e^{ik\frac{\Delta x}{2}}\sin^2\left(\frac{k\Delta x}{4}\right)\left(\cos\left(\frac{k\Delta x}{2}\right) - 3\right)\right)\right)\upsilon_j$$

So we have

$$\mathcal{G}_{FEM2} = \frac{6}{\Delta x}\frac{1}{\mathcal{R}_2^+ + \mathcal{R}_2^-} \times \left(H\frac{\Delta x}{30}\left(2e^{ik\frac{\Delta x}{2}}\left(2\cos\left(\frac{k\Delta x}{2}\right) - \cos\left(k\Delta x\right) + 4\right)\right) + \frac{H^3}{9\Delta x}\left(-8e^{ik\frac{\Delta x}{2}}\sin^2\left(\frac{k\Delta x}{4}\right)\left(\cos\left(\frac{k\Delta x}{2}\right) - 3\right)\right)\right)$$

This is the numerical methods transformation coefficient between G_j and $\upsilon_{j+1/2}$, the only υ value we need for our numerical method.

1.1.4 Conservation Equation

Finite volume methods have the following update scheme to approximate equations in conservation law form [] for some quantity q

$$\bar{q}_j^{n+1} = \bar{q}_j^n - \frac{\Delta t}{\Delta x}\left[F_{j+1/2}^n - F_{j-1/2}^n\right]. \tag{1.8}$$

Where the bar denotes that it is the cell average of the quantity q and $F_{j+1/2}^n$ and $F_{j-1/2}^n$ are the approximations to the average fluxes across the cell boundary between the times t^n and t^{n+1}.

In our methods there is some transformation between the nodal value q_j and the cell average $\bar{q}_j$, which produces some factor $\mathcal{M}$. For first and second order methods $\mathcal{M}_1 = \mathcal{M}_2 = 1$, however for higher-order methods $\mathcal{M} \neq 1$. Because of this we will highlight the transformations between cell averages and nodal values for this second-order example with $\mathcal{M}_2$ even though it is simply unity, so as to guide the reader if they wish to replicate this work for higher-order methods.

To calculate the fluxes $F^n_{j+1/2}$ and $F^n_{j-1/2}$ we use Kurganov's method [2] which is

$$F_{j+\frac{1}{2}} = \frac{a^+_{j+\frac{1}{2}} f\left(q^-_{j+\frac{1}{2}}\right) - a^-_{j+\frac{1}{2}} f\left(q^+_{j+\frac{1}{2}}\right)}{a^+_{j+\frac{1}{2}} - a^-_{j+\frac{1}{2}}} + \frac{a^+_{j+\frac{1}{2}}\, a^-_{j+\frac{1}{2}}}{a^+_{j+\frac{1}{2}} - a^-_{j+\frac{1}{2}}} \left[q^+_{j+\frac{1}{2}} - q^-_{j+\frac{1}{2}}\right]$$

where $a^+_{j+\frac{1}{2}}$ and $a^-_{j+\frac{1}{2}}$ are given by the wave speed bounds [], so that

$$a^-_{j+1/2} = -\sqrt{gH}$$

$$a^+_{j+1/2} = \sqrt{gH}.$$

We have suppressed the superscripts denoting the time to simplify the notation as all times are now t^n in the flux calculation. Substituting these values into the flux approximation we obtain

$$F_{j+\frac{1}{2}} = \frac{f\left(q^-_{j+\frac{1}{2}}\right) + f\left(q^+_{j+\frac{1}{2}}\right)}{2} - \frac{\sqrt{gH}}{2}\left[q^+_{j+\frac{1}{2}} - q^-_{j+\frac{1}{2}}\right] \tag{1.9}$$

For η our Kurganov approximation to the flux of (1.5a) is then

$$F^\eta_{j+\frac{1}{2}} = \frac{H\upsilon^-_{j+\frac{1}{2}} + H\upsilon^+_{j+\frac{1}{2}}}{2} - \frac{\sqrt{gH}}{2}\left[\eta^+_{j+\frac{1}{2}} - \eta^-_{j+\frac{1}{2}}\right] \tag{1.10}$$

The missing pieces here are the factors introduced by reconstruction of the edge values $\upsilon^-_{j+\frac{1}{2}}$, $\upsilon^+_{j+\frac{1}{2}}$, $\eta^-_{j+\frac{1}{2}}$ and $\eta^+_{j+\frac{1}{2}}$ from the cell averages $\bar{\upsilon}_j$ and $\bar{\eta}_j$. Because our quantities are smooth the nonlinear limiters can be neglected so we have for the second-order reconstruction of η

$$\eta^-_{j+\frac{1}{2}} = \bar{\eta}_j + \frac{-\bar{\eta}_{j-1} + \bar{\eta}_{j+1}}{4},$$
$$\eta^+_{j+\frac{1}{2}} = \bar{\eta}_{j+1} + \frac{-\bar{\eta}_j + \bar{\eta}_{j+2}}{4}.$$

Using (1.7) these equations become

$$\eta^-_{j+\frac{1}{2}} = \mathcal{M}_2\eta_j + \frac{-\mathcal{M}_2\eta_j e^{-ik\Delta x} + \mathcal{M}_2\eta_j e^{ik\Delta x}}{4}$$

$$\eta^+_{j+\frac{1}{2}} = \mathcal{M}_2\eta_j e^{ik\Delta x} + \frac{-\mathcal{M}_2\eta_j + \mathcal{M}_2\eta_j e^{2ik\Delta x}}{4}.$$

For the second order case $\mathcal{M}_2 = 1$ and these equations can be reduced to

$$\eta^-_{j+\frac{1}{2}} = \left(1 + \frac{i\sin(k\Delta x)}{2}\right)\eta_j$$

$$\eta^+_{j+\frac{1}{2}} = e^{ik\Delta x}\left(1 - \frac{i\sin(k\Delta x)}{2}\right)\eta_j.$$

From these we introduce the second order reconstruction factors $\mathcal{R}^+_2 = e^{ik\Delta x}\left(1 - \frac{i\sin(k\Delta x)}{2}\right)$ and $\mathcal{R}^-_2 = 1 + \frac{i\sin(k\Delta x)}{2}$ for both η and G. [] So that we have

$$\eta^-_{j+\frac{1}{2}} = \mathcal{R}^-_2\eta_j$$

$$\eta^+_{j+\frac{1}{2}} = \mathcal{R}^+_2\eta_j.$$

In our numerical methods our reconstruction of υ is slightly different as $\upsilon^-_{j+\frac{1}{2}}$ and $\upsilon^+_{j+\frac{1}{2}}$ are equal as we assume υ is continuous. For the second order finite difference volume method we have

$$\upsilon^-_{j+1/2} = \upsilon^+_{j+1/2} = \frac{\upsilon_{j+1} + \upsilon_j}{2}$$

Using (1.7) and rearranging gives

$$\upsilon^-_{j+1/2} = \upsilon^+_{j+1/2} = \frac{e^{ik\Delta x} + 1}{2}\upsilon_j.$$

We therefore introduce the second order reconstruction error factor $\mathcal{R}^\upsilon_{FD2} = \frac{e^{ik\Delta x}+1}{2}$. Because our FEM for the elliptic equation [] calculates $\upsilon_{j+1/2}$ from G_j we do not need to reconstruct $\upsilon_{j+1/2}$ and so we have the $\mathcal{R}^\upsilon_{FEM2} = e^{ik\frac{\Delta x}{2}}$, which corresponds to calculating $\upsilon_{j+1/2}$ analytically given υ_j.

We will now suppress the superscripts and subscripts of the factors in the flux approximation, as the rest of the calculations are independent of the particular reconstructions used. Substituting these error coefficients into into (1.10) results in

$$F^\eta_{j+\frac{1}{2}} = \frac{H\mathcal{R}^u\upsilon_j + H\mathcal{R}^u\upsilon_j}{2} - \frac{\sqrt{gH}}{2}\left[\mathcal{R}^+\eta_j - \mathcal{R}^-\eta_j\right]$$

Which becomes

$$F^\eta_{j+\frac{1}{2}} = H\mathcal{R}^u\upsilon_j - \frac{\sqrt{gH}}{2}\left[\mathcal{R}^+ - \mathcal{R}^-\right]\eta_j$$

We then introduce the factors $\mathcal{F}^{\eta,\upsilon}$ and $\mathcal{F}^{\eta,\eta}$

$$\mathcal{F}^{\eta,\eta} = -\frac{\sqrt{gH}}{2}\left[\mathcal{R}^{+} - \mathcal{R}^{-}\right] \tag{1.11}$$

$$\mathcal{F}^{\eta,\upsilon} = H\mathcal{R}^{u} \tag{1.12}$$

so that

$$F^{\eta}_{j+\frac{1}{2}} = \mathcal{F}^{\eta,\upsilon}\upsilon_j + \mathcal{F}^{\eta,\eta}\eta_j. \tag{1.13}$$

For G our Kurganov approximation (1.9) to the flux of (1.5a) is then

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

Repeating this process for above for $F^{G}_{j+\frac{1}{2}}$ we get that

$$F^{G}_{j+\frac{1}{2}} = \mathcal{F}^{G,\eta}\eta_j + \mathcal{F}^{G,\upsilon}\upsilon_j \tag{1.14}$$

where

$$\mathcal{F}^{G,\eta} = gH\frac{\mathcal{R}^{-}_{2} + \mathcal{R}^{+}_{2}}{2}, \tag{1.15}$$

$$\mathcal{F}^{G,\upsilon} = -\frac{\sqrt{gH}}{2}\left[\mathcal{R}^{+} - \mathcal{R}^{-}\right]\mathcal{G}. \tag{1.16}$$

By substituting (1.13), (1.14) into (1.8) our finite volume method can be written as

$$\mathcal{M}_2\eta^{n+1}_j = \mathcal{M}_2\eta^{n}_j - \frac{\Delta t}{\Delta x}\left[\left(1 - e^{ik\Delta x}\right)\left(\mathcal{F}^{\eta,\eta}_2 h_j + \mathcal{F}^{\eta,\upsilon}_2\upsilon_j\right)\right],$$

$$\mathcal{M}_2 G^{n+1}_j = \mathcal{M}_2 G^{n}_j - \frac{\Delta t}{\Delta x}\left[\left(1 - e^{ik\Delta x}\right)\left(\mathcal{F}^{G,\eta}_2\eta_j + \mathcal{F}^{G,\upsilon}_2\upsilon_j\right)\right].$$

Furthermore by transforming the G's into υ's using our second order finite volume factor $\mathcal{G}_{FD2}$ we obtain

$$\eta^{n+1}_j = \eta^{n}_j - \frac{1}{\mathcal{M}_2}\frac{\Delta t}{\Delta x}\left[\left(1 - e^{ik\Delta x}\right)\left(\mathcal{F}^{\eta,\eta}_2\eta_j + \mathcal{F}^{\eta,\upsilon}_2\upsilon_j\right)\right],$$

$$\upsilon^{n+1}_j = \upsilon^{n}_j - \frac{1}{\mathcal{G}_{FD2}\mathcal{M}_2}\frac{\Delta t}{\Delta x}\left[\left(1 - e^{ik\Delta x}\right)\left(\mathcal{F}^{G,\eta}_2\eta_j + \mathcal{F}^{G,\upsilon}_2\upsilon_j\right)\right].$$

This can be written in matrix form as

$$\begin{bmatrix}\eta\\ \upsilon\end{bmatrix}^{n+1}_{j} = \begin{bmatrix}\eta\\ \upsilon\end{bmatrix}^{n}_{j} - \frac{\left(1 - e^{ik\Delta x}\right)\Delta t}{\Delta x}\begin{bmatrix}\frac{1}{\mathcal{M}_2}\mathcal{F}^{\eta,\eta}_2 & \frac{1}{\mathcal{M}_2}\mathcal{F}^{\eta,\upsilon}_2\\ \frac{1}{\mathcal{G}_{FD2}\mathcal{M}_2}\mathcal{F}^{\upsilon,\eta}_2 & \frac{1}{\mathcal{G}_{FD2}\mathcal{M}_2}\mathcal{F}^{\upsilon,\upsilon}_2\end{bmatrix}\begin{bmatrix}\eta\\ \upsilon\end{bmatrix}^{n}_{j}$$

Introducing the matrix

$$\mathbf{F}_2 = \frac{\left(1 - e^{ik\Delta x}\right)}{\Delta x} \begin{bmatrix} \mathcal{F}_2^{\eta,\eta} & \mathcal{F}_2^{\eta,\upsilon} \\ \frac{1}{\mathcal{G}}\mathcal{F}_2^{\upsilon,\eta} & \frac{1}{\mathcal{G}}\mathcal{F}_2^{\upsilon,\upsilon} \end{bmatrix} \tag{1.17}$$

this becomes

$$\begin{bmatrix} \eta \\ \upsilon \end{bmatrix}_j^{n+1} = (\mathbf{I} - \Delta t \mathbf{F}_2) \begin{bmatrix} \eta \\ \upsilon \end{bmatrix}_j^n$$

where we have used $\mathcal{M}_2 = 1$. So we have created the desired equation and derived all the factors introduced by the various operators of the second order FDVM and FEVM. We will now demonstrate how the analysis continues with the Runge-Kutta Time Stepping employed by our higher-order methods, again by taking the second-order case as an example.

1.1.5 Runge-Kutta Time Stepping

Since we have demonstrated this process for a single evolution step the analysis will now proceed for the Runge-Kutta time stepping that make allow our schemes to be temporally higher order accurate. We will again demonstrate this process for just the second-order FDVM and FEVM, and omit the analysis for the other methods, presenting the results at the end.

For second order time stepping the Runge Kutta time stepping proceeds as follows

$$\begin{bmatrix} \eta \\ \upsilon \end{bmatrix}_j^1 = (\mathbf{I} - \Delta t \mathbf{F}_2) \begin{bmatrix} \eta \\ \upsilon \end{bmatrix}_j^n, \tag{1.18a}$$

$$\begin{bmatrix} \eta \\ \upsilon \end{bmatrix}_j^2 = (\mathbf{I} - \Delta t \mathbf{F}_2) \begin{bmatrix} \eta \\ \upsilon \end{bmatrix}_j^1 \tag{1.18b}$$

$$\begin{bmatrix} \eta \\ \upsilon \end{bmatrix}_j^{n+1} = \frac{1}{2} \left(\begin{bmatrix} \eta \\ \upsilon \end{bmatrix}_j^n + \begin{bmatrix} \eta \\ \upsilon \end{bmatrix}_j^2 \right) \tag{1.18c}$$

Substituting (1.18a) and (1.18b) into (1.18c) we can write this in terms of the second-order operator $\mathbf{F}_2$ and the primitive variables at t^n

$$\begin{bmatrix} \eta \\ \upsilon \end{bmatrix}_j^{n+1} = \frac{1}{2} \left(\begin{bmatrix} \eta \\ \upsilon \end{bmatrix}_j^n + (\mathbf{I} - \Delta t \mathbf{F}_2)^2 \begin{bmatrix} \eta \\ \upsilon \end{bmatrix}_j^n \right).$$

Expanding $(\mathbf{I} - \Delta t\mathbf{F}_2)^2$ we get

$$\begin{bmatrix}\eta\\ \upsilon\end{bmatrix}_j^{n+1} = \frac{1}{2}\left(2\mathbf{I} - 2\Delta t\mathbf{F}_2 + \Delta t^2\mathbf{F}_2^2\right)\begin{bmatrix}\eta\\ \upsilon\end{bmatrix}_j^{n}.$$

Provided that an eigenvalue decomposition $\mathbf{F}_2 = \mathbf{P}\mathbf{\Lambda}\mathbf{P}^{-1}$ exists our equation can be rewritten as

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

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

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

Since η and υ are Fourier modes we have by (1.7) that

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

Since $\mathbf{\Lambda}$ is a diagonal matrix consisting of the eigenvalues λ_1 and λ_2 we have that

$$e^{i\omega\Delta t} = 1 + \frac{1}{2}\Delta t^2\lambda_1^2 - \Delta t\lambda_1,$$
$$e^{i\omega\Delta t} = 1 + \frac{1}{2}\Delta t^2\lambda_2^2 - \Delta t\lambda_2.$$

So that the dispersion relation for the second order finite difference finite volume method is

$$\omega = \frac{1}{i\Delta t}\ln\left(1 + \frac{1}{2}\Delta t^2\lambda_1^2 - \Delta t\lambda_1\right), \tag{1.19}$$
$$\omega = \frac{1}{i\Delta t}\ln\left(1 + \frac{1}{2}\Delta t^2\lambda_2^2 - \Delta t\lambda_2\right). \tag{1.20}$$

This is the dispersion relation for the second-order FDVM which generates $\mathbf{F}_2$ and uses second-order Runge-Kutta time stepping. By comparing this with the dispersion relation (**??**) of the Serre equations we can determine the error in dispersion caused by the particular method.

For the first-order Runge Kutta steps we get that

$$\omega = \frac{1}{i\Delta t} \ln\left(1 - \Delta t \lambda_1\right), \tag{1.21}$$

$$\omega = \frac{1}{i\Delta t} \ln\left(1 - \Delta t \lambda_2\right), \tag{1.22}$$

where λ_1 and λ_2 are the eigenvalues of $\mathbf{F}_1$ which uses the appropriate first-order FDVM.

For the third-order Runge Kutta steps we get that

$$\omega = \frac{1}{i\Delta t} \ln\left(1 - \frac{1}{6}\Delta t^3 \lambda_1^3 + \frac{1}{2}\Delta t^2 \lambda_1^2 - \Delta t \lambda_1\right), \tag{1.23}$$

$$\omega = \frac{1}{i\Delta t} \ln\left(1 - \frac{1}{6}\Delta t^3 \lambda_2^3 + \frac{1}{2}\Delta t^2 \lambda_2^2 - \Delta t \lambda_2\right), \tag{1.24}$$

where λ_1 and λ_2 are the eigenvalues of $\mathbf{F}_3$ which uses the appropriate third-order FDVM.

We now proceed to present the tables of the factors generated by the operators $\mathcal{M}$, $\mathcal{G}$, $\mathcal{R}^+$, $\mathcal{R}^-$, $\mathcal{R}^v$. As the other factors of interest can be calculated directly just using these basic factors. We also provide the lowest order of the Taylor expansion of the error between the approximation to the operator and its true value to demonstrate that the methods are using appropriate order approximations. Finally we also present the error for the elements of the flux matrix $\mathbf{F}$ to demonstrate that when combined the pieces of our numerical method do indeed provide us with the correct spatial order of accuracy.

1.1.6 Tables of Factors

In the following we present tables which give both the formula for the 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. As shown above we use subscripts to denote the order of accuracy of the approximation, and where necessary specify the specific method for the two different second-order methods. We use subscript A to denote that it is the true factor if we calculated the operator analytically.

From Tables 1.1, 1.2, 1.3, 1.4 and 1.5 we can see that the basic operators all have the correct spatial order of accuracy or better. Therefore we expect the higher order methods to outperform lower order methods, due to the smaller errors associated with the improved spatial order of accuracy.

Variable	Formula	Lowest Order of Error
$\mathcal{M}_A$	$\frac{2}{k\Delta x}\sin\left(\frac{k\Delta x}{2}\right)$	–
$\mathcal{M}_1$	1	$\frac{1}{24}k^2\Delta x^2$
$\mathcal{M}_2$	1	$\frac{1}{24}k^2\Delta x^2$
$\mathcal{M}_3$	$\frac{24}{26-2\cos(k\Delta x)}$	$\frac{3}{640}k^4\Delta x^4$

Table 1.1: Factor $\mathcal{M}$ from transformation between nodal and cell average values

Variable	Formula	Lowest Order of Error
$\mathcal{R}_A^+$	$\exp\left(ik\frac{\Delta x}{2}\right)$	–
$\mathcal{R}_1^+$	$\exp(ik\Delta x)$	$\frac{i}{2}k\Delta x$
$\mathcal{R}_2^+$	$\exp(ik\Delta x)\left(1-\frac{i\sin(k\Delta x)}{2}\right)$	$\frac{1}{8}k^2\Delta x^2$
$\mathcal{R}_3^+$	$\frac{2\exp(2ik\Delta x)-10\exp(ik\Delta x)-4}{\cos(k\Delta x)-13}$	$\frac{i}{12}k^3\Delta x^3$

Table 1.2: Factor $\mathcal{R}^+$ from reconstruction at $x_{j+1/2}$ from cell C_{i+1} for η and G

Variable	Formula	Lowest Order of Error
$\mathcal{R}_A^-$	$\exp\left(ik\frac{\Delta x}{2}\right)$	–
$\mathcal{R}_1^-$	1	$-\frac{i}{2}k\Delta x$
$\mathcal{R}_2^-$	$1+\frac{i\sin(k\Delta x)}{2}$	$\frac{1}{8}k^2\Delta x^2$
$\mathcal{R}_3^-$	$\frac{2\exp(-ik\Delta x)-4\exp(ik\Delta x)-10}{\cos(k\Delta x)-13}$	$-\frac{i}{12}k^3\Delta x^3$

Table 1.3: Factor $\mathcal{R}^-$ from reconstruction at $x_{j+1/2}$ from cell C_i for η and G

Variable	Formula	Lowest Order of Error
$\mathcal{R}_A^v$	$\exp\left(ik\frac{\Delta x}{2}\right)$	–
$\mathcal{R}_1^v$	$\frac{\exp(ik\Delta x)+1}{2}$	$-\frac{1}{8}k^2\Delta x^2$
$\mathcal{R}_{FD2}^v$	$\frac{\exp(ik\Delta x)+1}{2}$	$-\frac{1}{8}k^2\Delta x^2$
$\mathcal{R}_{FEM2}^v$	$\exp\left(ik\frac{\Delta x}{2}\right)$	0
$\mathcal{R}_3^v$	$\frac{-\exp(-ik\Delta x)+9\exp(ik\Delta x)-\exp(2ik\Delta x)+9}{16}$	$-\frac{3}{128}k^4\Delta x^4$

Table 1.4: Factor $\mathcal{R}^v$ from reconstruction at $x_{j+1/2}$ for v

Variable	Formula	Lowest Order of Error
$\mathcal{G}_A$	$H + \frac{H^3}{3}k^2$	$-$
$\mathcal{G}_1$	$H - \frac{H^3}{3}\frac{2\cos(k\Delta x) - 2}{\Delta x^2}$	$-\frac{H^3}{36}k^4\Delta x^2$
$\mathcal{G}_{2FD}$	$H - \frac{H^3}{3}\frac{2\cos(k\Delta x) - 2}{\Delta x^2}$	$-\frac{H^3}{36}k^4\Delta x^2$
$\mathcal{G}_{2FEM}$	$*$	$-\frac{3H}{40}k^2\Delta x^2 - \frac{H^3}{36}k^4\Delta x^2$
$\mathcal{G}_3$	$H - \frac{H^3}{3}\frac{32\cos(k\Delta x) - 2\cos(2k\Delta x) - 30}{12\Delta x^2}$	$-\frac{H^3}{270}k^6\Delta x^4$

Table 1.5: Factor $\mathcal{G}$ from solving the elliptic equation (1.4c)

$$\mathcal{G}_{2FEM} = \left(\frac{2H^3}{3\Delta x^2}\left(\exp\left(ik\frac{3\Delta x}{2}\right) + 14\exp\left(ik\frac{\Delta x}{2}\right) - 8\exp(ik\Delta x) - 8 + \exp\left(-ik\frac{\Delta x}{2}\right)\right)\right.$$
$$\left. + \frac{H}{5}\left(-\exp\left(ik\frac{3\Delta x}{2}\right) + 8\exp\left(ik\frac{\Delta x}{2}\right) + 2\exp(ik\Delta x) + 2 - \exp\left(-ik\frac{\Delta x}{2}\right)\right)\right) \div$$
$$\left(-\frac{1}{4}\exp(2i\Delta xk) + \exp(i\Delta xk) + \frac{i}{2}\sin(k\Delta x) + \frac{5}{4}\right)$$

Variable	Exact	Lowest Order Truncation Term			
		FDVM$_1$	FDVM$_2$	FEVM$_2$	FDVM$_3$
$\frac{\mathcal{D}}{\mathcal{M}\Delta x}\mathcal{F}^{\eta,\eta}$	0	$\frac{\sqrt{gH}}{2}k^2\Delta x$	$\frac{\sqrt{gH}}{8}k^4\Delta x^3$	$\frac{\sqrt{gH}}{8}k^4\Delta x^3$	$\frac{\sqrt{gH}}{12}k^4\Delta x^3$
$\frac{\mathcal{D}}{\mathcal{M}\Delta x}\mathcal{F}^{\eta,v}$	ikH	$-\frac{iH}{6}k^3\Delta x^2$	$-\frac{iH}{6}k^3\Delta x^2$	$-\frac{iH}{24}k^3\Delta x^2$	$-\frac{9iH}{320}k^5\Delta x^4$
$\frac{\mathcal{D}}{\mathcal{GM}\Delta x}\mathcal{F}^{G,\eta}$	$\frac{3ikgH}{H^2k^2+3}$	$-\frac{ig\left(H^2k^2+6\right)}{4\left(H^2k^2+3\right)^2}k^3\Delta x^2$	$\frac{ig\left(2H^2k^2+3\right)}{4\left(H^2k^2+3\right)^2}k^3\Delta x^2$	$\frac{ig\left(20H^2k^2+57\right)}{40\left(H^2k^2+3\right)^2}k^3\Delta x^2$	$-\frac{ig\left(2H^2k^2+9\right)}{30\left(H^2k^2+3\right)^2}k^5\Delta x^4$
$\frac{\mathcal{D}}{\mathcal{GM}\Delta x}\mathcal{F}^{G,v}$	0	$\frac{\sqrt{gH}}{2}k^2\Delta x$	$\frac{\sqrt{gH}}{8}k^4\Delta x^3$	$\frac{\sqrt{gH}}{8}k^4\Delta x^3$	$\frac{\sqrt{gH}}{12}k^4\Delta x^3$

Table 1.6: Elements of **F**. Due to the length of the expression for these values we now only show the lowest order term in the Taylor series of the error and the analytic value.

The most notable thing about these tables is the difference between the second-order FDVM and FEVM. For most operators both have the same factor, in particular for $\mathcal{M}$ (Table 1.1), $\mathcal{R}^+$ (Table 1.2) and $\mathcal{R}^-$ (Table 1.3). The main difference between these methods comes in factors $\mathcal{G}$ (Table 1.1) and $\mathcal{R}^\upsilon$ (Table 1.2). In particular we can see that for $\mathcal{G}$ the FDVM has a smaller error than the FEVM, which is caused by the FEM approximations to the integrals of G and υH, whereas the FDVM just takes the point values of these quantities without integral approximations. While the derivative approximation for both methods introduces the same error. For $\mathcal{R}^\upsilon$ we can see that the FEVM performs better than the FDVM because no reconstruction is required as the FEM calculated $\upsilon_{j+1/2}$ directly.

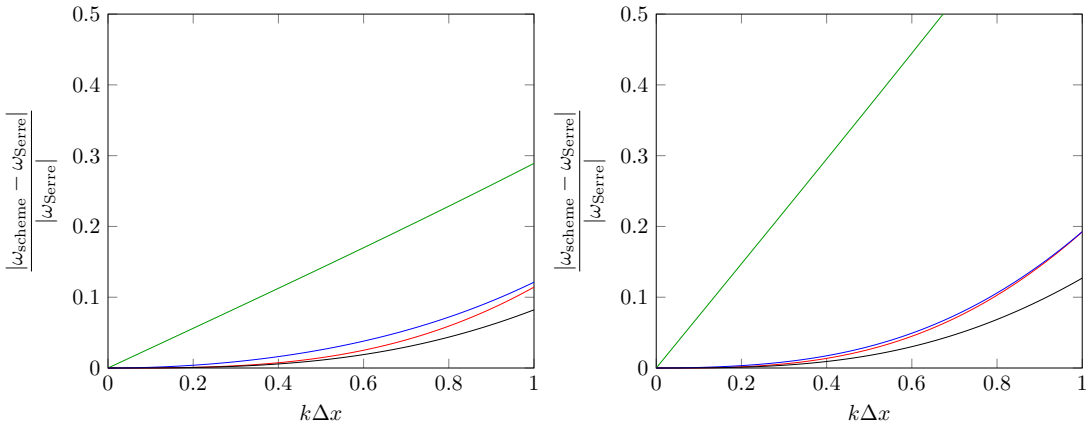
When these errors are combined in the construction the flux matrix in Table 1.6 these lead to slightly larger errors for the second-order FEVM compared to the FDVM for $\frac{\mathcal{D}}{\mathcal{G}\mathcal{M}\Delta x}\mathcal{F}^{G,\eta}$ while the error is larger for the second-order FDVM compared to the FEVM for $\frac{\mathcal{D}}{\mathcal{M}\Delta x}\mathcal{F}^{\eta,\upsilon}$. For the other methods we see that the flux matrix elements have the correct spatial order of accuracy, and therefore high-order methods will outperform lower-order ones as expected.

We can also see that all these methods introduce some diffusive error for $\frac{\mathcal{D}}{\mathcal{M}\Delta x}\mathcal{F}^{\eta,\eta}$ and $\frac{\mathcal{D}}{\mathcal{G}\mathcal{M}\Delta x}\mathcal{F}^{G,\upsilon}$. This is due to the Kurganov approximation (1.9) containing both a flux averaging part and a diffusive part, which are nicely split for these linearised equations. In particular we have the terms $\frac{\mathcal{D}}{\mathcal{G}\mathcal{M}\Delta x}\mathcal{F}^{G,\eta}$ and $\frac{\mathcal{D}}{\mathcal{M}\Delta x}\mathcal{F}^{\eta,\upsilon}$ as the flux average part while $\frac{\mathcal{D}}{\mathcal{M}\Delta x}\mathcal{F}^{\eta,\eta}$ and $\frac{\mathcal{D}}{\mathcal{G}\mathcal{M}\Delta x}\mathcal{F}^{G,\upsilon}$ are the diffusive part. This shows up in the errors for these terms as even powers of Δx for the dispersive errors of the flux averaging and odd powers of Δx for the diffusive errors introduced by the dispersive term.

Having demonstrate how to compute these factors and given expressions for the basic ones, from which the more complicated factors can be obtained we will now present our results for the dispersion error of the FDVM and FEVM numerical methods.

1.1.7 Results

From the basic factors presented in the Tables 1.1, 1.2, 1.3, 1.4 and 1.5 the flux factors $\mathcal{F}^{\eta,\eta}$,$\mathcal{F}^{\eta,\upsilon}$, $\mathcal{F}^{G,\eta}$ and $\mathcal{F}^{G,\upsilon}$ can be calculated using (1.11) and (1.15) respectively. From there the matrix $\mathbf{F}$ can be computed using (1.17), and its eigenvalues found. Having found the eigenvalues we then substitute them into the appropriate equation given by the Runge-Kutta time stepping method; (1.21)



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

Figure 1.1: Dipserion error for first-order FDVM (–), second-order FDVM(–), second-order FEVM (–) and third-order FDVM (–).

for first-order, (1.19) for second-order or (1.23) for third-order.

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

From these plots we can see that as expected as you increase the order of accuracy of the scheme your dispersion relation error decreases everywhere. This is most evident for the different order FDVM. Also we have as expected that increasing the resolution of our numerical methods decreases the dispersion error of the numerical scheme.

More interesting is that the second-order FDVM performs better than the FEVM consistently across various numerical resolutions indicating that the FEVM will have a larger phase error than its FDVM counterpart. Thus it appears that error in approximating $\mathcal{G}$ was more significant than the error in approximating $\mathcal{R}^v$ for the dispersion relation.

Our results compare well with those of [1] 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.

These plots only depend on the parameter kH and do not vary independently for H or k. This parameter kH is proportional to the shallowness parameter σ with $2\pi\sigma = kH$. So for kH we have left shallow water and the Serre equations are

no longer an appropriate model for water waves, although our results demonstrate that our numerical methods are still well approximating the dispersion relation of the Serre equations in this case. In general we also find that as kH is increased our numerical methods perform worse generally although the dispersion relation still converges to 0 as $\Delta x \to 0$.

We will now turn our attention to demonstrating Von Neumann stability for the two finite difference methods described in this thesis [], and used for the purposes of comparison. We shall also see that we can use some of the previous working on the dispersion relation error to demonstrate stability for our FDVM and FEVM when $U = 0$.

1.2 Von Neumann Stability

To demonstrate Von Neumann stability we again begin with the linearised Serre equations (1.3) however will no longer require $U = 0$, so we have

$$\frac{\partial \eta}{\partial t} + H\frac{\partial \upsilon}{\partial x} + U\frac{\partial \eta}{\partial x} = 0,$$

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

As in the above dispersion relation analysis we will demonstrate the working for one problem and then just present the results for the other one to be both repeatable and brief. Our example will be the naive second-order finite difference method $\mathcal{D}$ (**??**).

We will again begin by replacing both η and υ by Fourier nodes (1.6). Because our approximations to derivatives is constant for $\mathcal{D}$ we will provide all the factors for the second order centred finite difference approximations to derivatives of some quantity q generated by making use of (1.7). Having all the factors associated with all the spatial derivatives we are really only left with the task of rearranging the equations and writing it in matrix form. The factors for each derivative approximation are

$$\left(\frac{\partial q}{\partial x}\right)_j^n = \frac{-q_{j-1}^n + q_{j+1}^n}{2\Delta x} = \frac{i\sin(k\Delta x)}{\Delta x} q_j^n \tag{1.25a}$$

$$\left(\frac{\partial^2 q}{\partial x^2}\right)_j^n = \frac{q_{j-1}^n - 2q_j^n + q_{j+1}^n}{\Delta x^2} = \frac{2\cos(k\Delta x) - 2}{\Delta x^2} q_j^n \tag{1.25b}$$

$$\left(\frac{\partial^3 q}{\partial x^3}\right)_j^n = \frac{-q_{j-2}^n + 2q_{j-1}^n - 2q_{j+1}^n + q_{j+2}^n}{2\Delta x^3} = -4i\sin(k\Delta x)\frac{\sin^2\left(\frac{k\Delta x}{2}\right)}{\Delta x^3}q_j^n \quad (1.25c)$$

The numerical method $\mathcal{D}$ is just attained from replacing all the derivatives in (1.3) with the approximations in (1.25). For the linearised equations the update formulas for $\mathcal{D}$ becomes

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

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

Since we have assumed that η and υ are fourier nodes, we can just replace the finite difference approximations with the appropriate factors from (1.25). After some rearranging 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 vector form as

$$\begin{bmatrix}\eta_j^{n+1}\\ \upsilon_j^{n+1}\\ \eta_j^n\\ \upsilon_j^n\end{bmatrix} = \mathbf{E}\begin{bmatrix}\eta_j^n\\ \upsilon_j^n\\ \eta_j^{n-1}\\ \upsilon_j^{n-1}\end{bmatrix}, \quad (1.26)$$

where

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

This is the evolution matrix $\mathbf{E}_{\mathcal{D}}$ and if its spectral radius is at most 1 then $\mathcal{D}$ is stable.

After following through with the same process for $\mathcal{W}$ we get the following matrix equation

$$\begin{bmatrix} \eta_j^{n+1} \\ \upsilon_j^{n+1} \\ \eta_j^{n} \\ \upsilon_j^{n} \end{bmatrix} = \mathbf{E}_{\mathcal{W}} \begin{bmatrix} \eta_j^{n} \\ \upsilon_j^{n} \\ \eta_j^{n-1} \\ \upsilon_j^{n-1} \end{bmatrix}. \tag{1.27}$$

where

$$\mathbf{E}_{\mathcal{W}} = \begin{bmatrix} E_{\mathcal{W}}^{0,0} & E_{\mathcal{W}}^{0,1} & 0 & -\frac{\Delta t}{\Delta x}H\frac{i\sin(k\Delta x)}{2} \\ -\frac{6gi\Delta x\Delta t}{3\Delta x^2 - 2H^2(\cos(k\Delta x)-1)}\sin(k\Delta x) & -\frac{2i\Delta t}{\Delta x}U\sin(k\Delta x) & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \tag{1.28}$$

with

$$E_{\mathcal{W}}^{0,0} = 1 - \frac{\Delta t}{\Delta x}\left(-\frac{6gi\Delta x\Delta t}{3\Delta x^2 - 2H^2(\cos(k\Delta x)-1)}\sin(k\Delta x)\right)H\frac{i\sin(k\Delta x)}{2}$$
$$-\frac{\Delta t}{\Delta x}U\left((i\sin(k\Delta x)) - \frac{\Delta t}{\Delta x}U(\cos(k\Delta x)-1)\right),$$
$$E_{\mathcal{W}}^{0,1} = -\frac{\Delta t}{\Delta x}\left[H\frac{i\sin(k\Delta x)}{2}\left(1 - \frac{2i\Delta t}{\Delta x}U\sin(k\Delta x)\right) - U\left(\frac{\Delta t}{\Delta x}H(\cos(k\Delta x)-1)\right)\right].$$

If the evolution matrix $\mathbf{E}_{\mathcal{W}}$ has a spectral radius less than or equal to 1 then $\mathcal{W}$ is stable.

It can be seen that we generated this evolution matrix for each of the FDVM and FEVM. In particular we have that

$$\mathbf{E}_1 = \mathbf{I} - \Delta t \mathbf{F}_1$$
$$\mathbf{E}_{FD2} = \mathbf{I} - \Delta t \mathbf{F}_{FD2} + \frac{1}{2}\Delta t^2 \mathbf{F}_{FD2}^2$$
$$\mathbf{E}_{FEM2} = \mathbf{I} - \Delta t \mathbf{F}_{FEM2} + \frac{1}{2}\Delta t^2 \mathbf{F}_{FEM2}^2$$
$$\mathbf{E}_3 = \mathbf{I} - \Delta t \mathbf{F}_3 + \frac{1}{2}\Delta t^2 \mathbf{F}_3^2 - \frac{1}{6}\Delta t^3 \mathbf{F}_3^3.$$

For the FDVM and FEVM the matrix equation is a bit simpler

$$\begin{bmatrix} \eta \\ v \end{bmatrix}_j^{n+1} = \mathbf{E} \begin{bmatrix} \eta \\ v \end{bmatrix}_j^n. \tag{1.29}$$

As with the finite difference methods if the spectral radius of the evolution matrix is at most 1 then the scheme is stable.

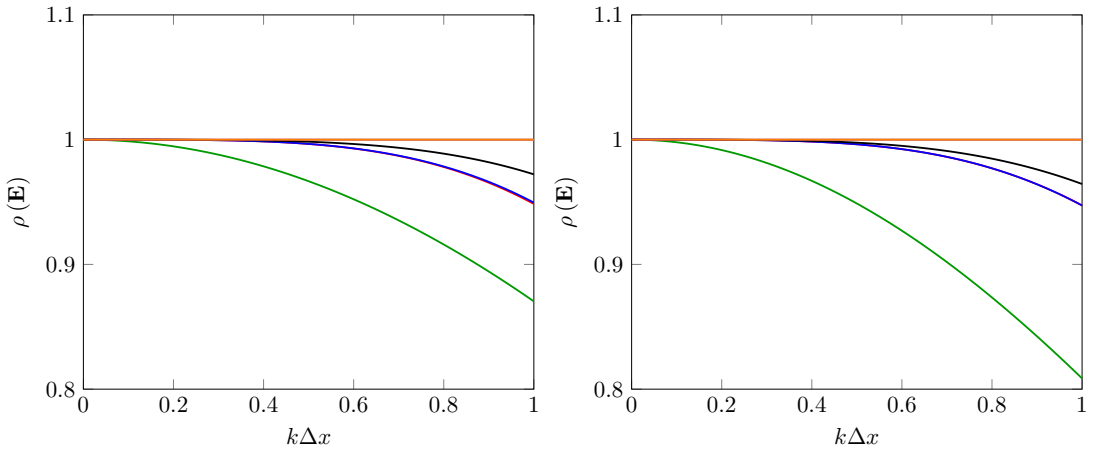
We will now present the results of our analysis of the stability of all these methods for various U, H and k values.

1.2.1 Results

We will demonstrate that these finite difference methods possess Von Neumann stability numerically. We do this by calculating the spectral radius of the growth matrices numerically for various 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 $\Delta t = \left(0.5/\sqrt{gH}\right)\Delta x$. We first show the results for $U = 0$ so that we can demonstrate the stability of the FDVM and FEVM as well and then allow various U values and we therefore do not include the FDVM and FEVM.

Quiescent Fluid

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 growth matrices for the FDVM and FEVM was calculated and thus we can only demonstrate the stability of all methods in this region. Although of course as mentioned previously the FDVM and FEVM inherit their stability from the FVM at their core.



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

Figure 1.2: Spectral radius of growth matrix $\mathbf{E}$ for first-order FDVM (–), second-order FDVM(–), second-order FEVM (–), third-order FDVM (–), $\mathcal{D}$ (–) and $\mathcal{W}$ (–) .

The spectral radius for a range of Δx values were plotted in Figure 1.2 for all numerical methods in this thesis. The representative values of $kH = 0.5$ and $kH = 2.5$ were the chosen due to their use in the dispersion error analysis for the stability results plotted in Figure 1.1.

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 other values of hK as well.

We note that both the second-order FDVM (–) and FEVM (–) have very similar spectral radius values such that their plots overlap and 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. This can be seen in Table 1.7, where the average of the spectral radius for $\mathcal{D}$ and $\mathcal{D}$ over various $k\Delta x$ values is 1 plus a number which is just round-off errors accumulated by doing this process numerically.

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 $U = 0$.

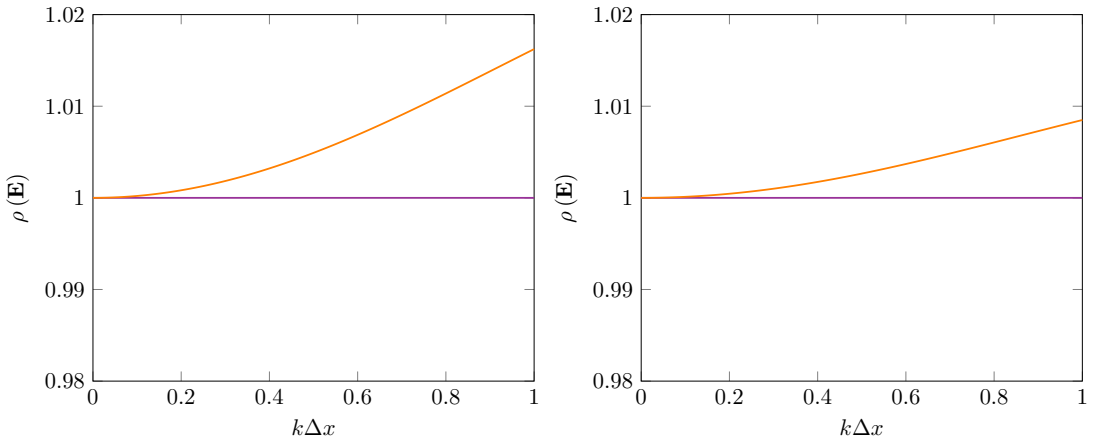
Non-zero Mean Flow

The situation of waves on still water is certainly the most common situation in ocean modelling, however there are scenarios that can arise where certain regions have a mean flow and we are interested in the waves on top, such as undular bores. Therefore we must also demonstrate stability in this case fo the two finite difference methods. Since the dispersion relation analysis that derived the evolution matrix for the FDVM and FEVM assumed $U = 0$ we do not demonstrate their stability here.

We have investigated the behaviour of the spectral radius of the growth matrix for various values of U, kH and Δx. We present the results for $U = 1$ with $kH = 0.5$ and 2.5 in Figure 1.3. These values were chosen because they are representative of the behaviour for both methods for most values of U and kH, and because they match the previous values in the results[].

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 at round-off error. This behaviour was consistent for various U, kH and Δx values provided the CFL condition was used to determine Δt. Therefore this method is stable as desired for a range of flow scenarios.

Unfortunately the Lax-Wendroff finite difference 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 as desired. However, the parameters do not follow some simple rule and their behaviour is quite



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

Figure 1.3: 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 $U = 1$.

complicated, it is however that case that our largest Δx value did 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 still very close to 1 for most situations and so the instabilities may not be apparent when performing numerical experiments, as we demonstrate in chapter [].

Bibliography

[1] 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.

[2] 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.