Numerical solution of conservation laws applied to the Shallow Water Wave Equations

Stephen G Roberts
Mathematical Sciences Institute, Australian National University
Updated January 4, 2013
(based on notes by SGR and Chris Zoppou, and report by Angus Griffith)

AMSI Summer School, 2013

Contents

1 Shallow Water Wave Equations **4**

1.1 Geometry of the problem 4
1.2 Eulerian Formulation 5
1.3 Lagrangian Formulation 6
1.4 Conservation of mass 8
1.5 1D Saint Venant equations 10
1.6 Conservation and Balance Laws 10

2 Non-Linear Conservation Laws **12**

2.1 Shock-wave Example 12
2.2 Advection Equations 13
2.3 Integral Conservation Laws 13
2.4 PDE Conservation Laws 16
2.5 Rarefaction Fan 20
2.6 Problems 21

3 Dam Break Problem **25**

3.1 Stoker's Solution 25
3.2 Ritter's Dry Bed Solution 30

4 Conservative Numerical Schemes **33**

5 Modified Equivalent Partial Differential Equation **35**

5.1 Advective-diffusion equation 38
5.2 Advective-dispersion equation 40
5.3 Higher-order equations 43
5.4 Accuracy in the presence of a discontinuity 45
5.5 First-order schemes 46
5.6 Second-order schemes 47

6 Naive Finite Difference Schemes **49**

6.1 Lax-Friedrichs 65
6.2 Richtmyer Scheme 65
6.3 Lax-Wendroff Scheme 65
6.4 Beam-Warming Scheme 69
6.5 MacCormack Scheme 69

7 Artificial Viscosity and Limited Schemes **70**
7.1 Limiters . . . 76
7.2 Flux Corrected Transport . . . 87

8 Upwind Schemes **90**
8.1 Bermudez and Vazquez Scheme . . . 91
8.2 Steger and Warming Scheme . . . 94
8.3 Yang et al. . . . 96
8.4 Harten Scheme . . . 97

9 Godunov-Type Schemes **99**
9.1 Averaging . . . 99
9.2 Re-averaging . . . 100
9.3 Evolution Step . . . 102
9.4 Exact Riemann Solver . . . 107
9.5 The Riemann Problem for Constant Coefficient System . . . 119
9.6 Roe's Solver . . . 121
9.7 Osher's scheme . . . 126
9.8 Harten, Lax and van Leer Scheme . . . 128
9.9 Weighted Average Flux Scheme . . . 132
9.10 Piecewise Parabolic Method . . . 135

10 Splitting Techniques **141**
10.1 Fractional splitting . . . 142
10.2 Strang splitting . . . 142

11 Boundary Conditions **144**
11.1 Transmissive Boundary Condition . . . 144
11.2 Dirichlet Boundary Condition . . . 147
11.3 Non-Reflective Boundary Condition . . . 148

12 Treatment of the Source Term using Fractional Steps **154**
12.1 The Ordinary Differential Equation . . . 160

13 The Finite Volume Method **162**
13.1 Basic Scheme . . . 162
13.2 Side Channel Friction . . . 166
13.3 Wetting and Drying of Cells . . . 167

1 Shallow Water Wave Equations

There are several ways to derive the shallow water wave equations (SWWEs). Often they are derived by assuming incompressible, inviscid, irrotational flow and applying depth averaging to the Navier-Stokes equations; see for example [54, pp. 12–21].

The methods presented here are more natural in that the SWEE is derived from first principles using conservation of mass and conservation of momentum.

Unlike other equations describing fluid flow such as Navier-Stokes equation, the shallow water wave equations are two dimensional equations. The solutions to the shallow water wave equations give the height of water $h(x, y)$ above the ground level, along with the velocity field $(u(x, y), v(x, y))$.

1.1 Geometry of the problem

Begin by considering water flowing over some terrain, which has been discretised into two-dimensional cells. Let $h(x, y)$ be the height of water above some terrain of height $b(x, y)$ Functions, $u(x, y)$ and $v(x, y)$ describe the rate of flow of water in the horizontal x and y directions respectively.

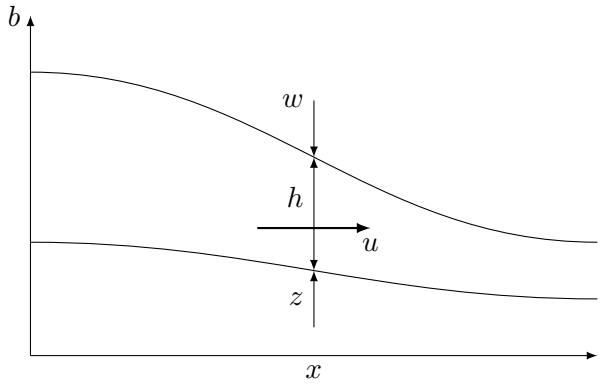


Figure 1: Nomenclature for shallow water wave equations; depth of water, h; elevation of the bed, b; the stage, $w = h + b$; and the depth averaged horizontal velocity, u.

The total height (stage) w of the water above some reference point is hence given by

$$w \equiv h + b$$

1.2 Eulerian Formulation

The Eulerian formulation considers the problem from the perspective of a single cell. The volume of water in a particular cell $((x_0, x_1), (y_0, y_1))$ at time t is given by

$$\int_{x_0}^{x_1} \int_{y_0}^{y_1} h(x, y, t)\, dy\, dx \tag{1}$$

As time passes the amount of water changes according to the flows coming in and out of the cell (see Figure 2). When discretising in time, the volume of water in a cell will evolve according to the

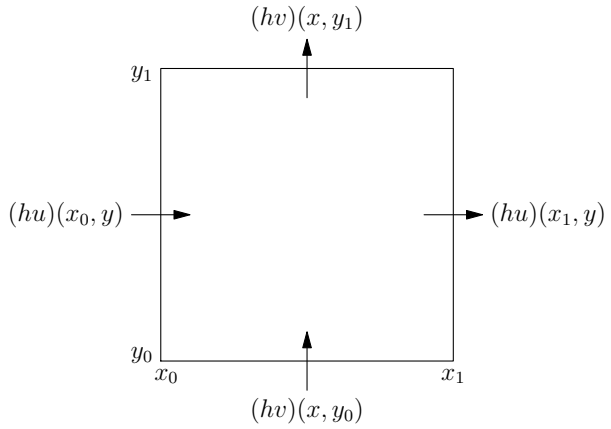


Figure 2: Flow in and out of a cell.

flows u and v,

$$\begin{aligned}\int_{x_0}^{x_1} \int_{y_0}^{y_1} h(x, y, t_1)\, dx\, dy &= \int_{x_0}^{x_1} \int_{y_0}^{y_1} h(x, y, t_0)\, dx\, dy \\ + \int_{t_0}^{t_1} \int_{y_0}^{y_1} h(x_0, y, t) u(x_0, y, t)\, dy\, dt &- \int_{t_0}^{t_1} \int_{y_0}^{y_1} h(x_1, y, t) u(x_1, y, t)\, dy\, dt \\ + \int_{t_0}^{t_1} \int_{x_0}^{x_1} h(x, y_0, t) v(x, y_0, t)\, dx\, dt &- \int_{t_0}^{t_1} \int_{x_0}^{x_1} h(x, y_1, t) v(x, y_1, t)\, dx\, dt\end{aligned} \tag{2}$$

Applying the fundamental theorem of calculus[1] in variables t, x and y,

$$\begin{aligned}\int_{t_0}^{t_1} \int_{x_0}^{x_1} \int_{y_0}^{y_1} h_t\, dy\, dx\, dt + \int_{t_0}^{t_1} \int_{x_0}^{x_1} \int_{y_0}^{y_1} (uh)_x\, dy\, dx\, dt \\ + \int_{t_0}^{t_1} \int_{x_0}^{x_1} \int_{y_0}^{y_1} (vh)_y\, dy\, dx\, dt = 0.\end{aligned} \tag{3}$$

Since this is true for any choice of cell in space and time,

$$h_t + (uh)_x + (vh)_y = 0. \tag{4}$$

Equation (4) is the mass conservation equation of shallow water wave equations.

[1] An equivalent derivation can be formulated by applying the divergence theorem to the domain in Figure 2.

1.3 Lagrangian Formulation

In contrast to the Eulerian formulation, the Lagrangian formulation takes the perspective of a domain of water which propagates in time (see Figure 3). We will use a notation which is dimension independent.

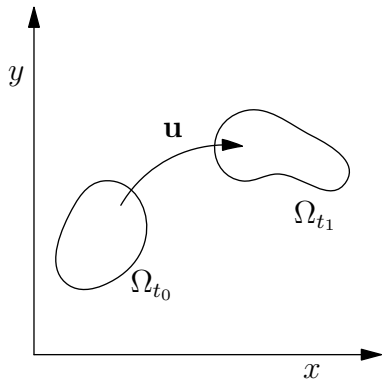


Figure 3: Propagation of a domain Ω of water.

Let $\mathbf{x} \in \mathbb{R}^2$ define an arbitrary position, and let $\mathbf{u} : \mathbb{R}^2 \times \mathbb{R}^+ \to \mathbb{R}^2$ define the velocity field of fluid. That is,

$$\mathbf{u}(\mathbf{x}, t) := (u(\mathbf{x}, t), v(\mathbf{x}, t)) \tag{5}$$

Now consider the evolution of particles under the influence of the velocity field $\mathbf{u}$. The position $\mathbf{x}(\mathbf{s}, t) \in \mathbb{R}^2$ of a particle flowing in the velocity field satisfies

$$\frac{d\mathbf{x}}{dt}(\mathbf{s}, t) = \mathbf{u}(\mathbf{x}(\mathbf{s}, t), t). \tag{6}$$

Theorem 1.1 (Flowing Integrals)**.** *Let $f(\mathbf{x}, t)$ be the density of some arbitrary quantity (for example, the concentration of a dissolved material). Consider the amount of the quantity in the region Ω_t, moving with the fluid, i.e. consider*

$$F(t) = \int_{\Omega_t} f(\mathbf{x}, t)\, d\mathbf{x}. \tag{7}$$

The rate of change of this quantity is given by,

$$\frac{d}{dt}F(t) \;\; = \;\; \int_{\Omega_t} \left(f_t + \operatorname{div}(f\mathbf{u})\right)\, d\mathbf{x}.$$

Proof

$$\frac{dF}{dt} = \frac{d}{dt}\int_{\Omega_t} f(\mathbf{x},t),t)d\mathbf{x} = \int_{\Omega_0} \frac{d}{dt}\left(f(\mathbf{x}(\mathbf{s},t),t)\det \mathbf{x_s}\right) d\mathbf{s}$$

$$= \int_{\Omega_0} \frac{d}{dt}\left(f(\mathbf{x}(\mathbf{s},t),t)\right)\det \mathbf{x_s} d\mathbf{s} \tag{8}$$

$$+ \int_{\Omega_0} f(\mathbf{x}(\mathbf{s},t),t)\frac{d}{dt}\left(\det \mathbf{x_s}\right) d\mathbf{s} \tag{9}$$

Well we need to differentiate the determinant of a matrix. Jacobi's formula provides the result we need.

Theorem 1.2 (Jacobi's formula: differential of a determinant)**.**

$$d\left(\det A\right) = \text{Trace}\left(\text{adj}(A)\ dA\right)$$

Proof When considering the determinant as a function of all the matrix elements, it is clear that the chain rule can be applied to yield,

$$d(\det A) = \sum_i \sum_j \frac{\partial \det(A)}{\partial a_{ij}} da_{ij}$$

where a_{ij} denotes the ij-th component of the matrix A.

By Laplace's formula

$$\det A = \sum_k (-1)^{i+k} a_{ik} \det(A_{ik}) = \sum_k a_{ik}\,\text{adj}(A)_{ki}$$

where A_{ij} is the block matrix obtained by removing row i and column j. Note that $\text{adj}(A)_{ji}$ are elements in the transpose of the adjoint of A. We have,

$$\frac{\partial \det(A)}{\partial a_{ij}} = \sum_k \frac{\partial a_{ik}}{\partial a_{ij}}\,\text{adj}(A)_{ki} = \text{adj}(A)_{ji}$$

since $\text{adj}(A)_{ki}$ does not depend on elements of A in the i^{th} row. Therefore,

$$\begin{aligned} d\left(\det A\right) &= \textstyle\sum_i \sum_j \text{adj}(A)_{ji}\ da_{ij} \\ &= \text{Trace}\left(\text{adj}(A)\ dA\right). \end{aligned}$$

□

We can now apply Jacobi's formula, to equation (9).

$$\frac{d}{dt}\left(\det \mathbf{x_s}\right) = \text{Trace}\left(\text{Adj}(\mathbf{x_s})\ \frac{d\mathbf{x_s}}{dt}\right).$$

Now $\frac{d\mathbf{x}}{dt} = \mathbf{u}$ and so $\frac{d\mathbf{x_s}}{dt} = \mathbf{u_x x_s}$. So, using Jacobi's formula yields

$$\begin{aligned}\frac{d}{dt}\left(\det \mathbf{x_s}\right) &= \text{Trace}\left(\text{adj}(\mathbf{x_s})\mathbf{u_x x_x}\right)\\ &= \text{Trace}\left(\det \mathbf{x_s} I \mathbf{u_x}\right)\\ &= \det \mathbf{x_s}\,\text{Trace}(\mathbf{u_x})\\ &= \det \mathbf{x_s}\,\text{div}\,\mathbf{u}.\end{aligned}$$

So the rate of change of area form is proportional to the divergence of the velocity field.

This can be used to simplify equation (9), namely,

$$\int_{\Omega_0} f(\mathbf{x}(\mathbf{s},t),t)\,\text{div}(\mathbf{u})\det(\mathbf{x_s})\;d\mathbf{s} = \int_{\Omega_t} f(\mathbf{x},t)\,\text{div}(\mathbf{u})\;d\mathbf{x}.$$

Equation (8) satisfies

$$\begin{aligned}\int_{\Omega_0}\frac{d}{dt}\left(f(\mathbf{x}(\mathbf{s},t),t)\right)\det \mathbf{x_s}\,d\mathbf{s} &= \int_{\Omega_0}\left(f_t(\mathbf{x}(\mathbf{s},t),t) + \nabla_{\mathbf{x}} f(\mathbf{x}(\mathbf{s},t),t)\cdot\frac{d\mathbf{x}(\mathbf{s},t)}{dt}\right)\det \mathbf{x_s}\,d\mathbf{s}\\ &= \int_{\Omega_t}\left(f_t(\mathbf{x},t) + \nabla_{\mathbf{x}} f(\mathbf{x},t)\cdot\mathbf{u}(\mathbf{x},t)\right)\;d\mathbf{x}.\end{aligned}$$

Taken together we have

$$\begin{aligned}\frac{d}{dt}\int_{\Omega_t} f(\mathbf{x},t)d\mathbf{x} &= \int_{\Omega_t}\left(f_t(\mathbf{x},t) + \nabla_{\mathbf{x}} f(\mathbf{x},t)\mathbf{u}(\mathbf{x},t) + f(\mathbf{x},t)\,\text{div}(\mathbf{u})\right)\;d\mathbf{x}\\ &= \int_{\Omega_t}\left(f_t(\mathbf{x},t) + \text{div}(f(\mathbf{x},t)\mathbf{u}(\mathbf{x},t))\right)\;d\mathbf{x}\end{aligned}$$

.

1.4 Conservation of mass

Conservation of mass can be expressed as

$$\frac{d}{dt}\int_{\Omega_t} h(\mathbf{x},t)dt = 0$$

By Theroem 1.1

$$\frac{d}{dt}\int_{\Omega_t} h(\mathbf{x},t)\,d\mathbf{x} = \int_{\Omega_t}\left(h_t + \nabla\cdot(\mathbf{u}h)\right)\;d\mathbf{x}.$$

Conservation of mass hence takes the form

$$\int_{\Omega_t}\left(h_t + \nabla\cdot(\mathbf{u}h)\right)d\mathbf{x} = 0. \tag{10}$$

This conservation equation holds for all Ω_t. In differential form,

$$\begin{aligned}&h_t + \nabla\cdot(\mathbf{u}h) = 0\\ &h_t + (uh)_x + (vh)_y = 0\end{aligned} \tag{11}$$

This is exactly the conservation of mass equation (4) introduced earlier.

Conservation of momentum and the hydrostatic assumption

The Shallow water wave equations are completed by considering so called conservation of momentum, actually Newton's second law, the rate of change of momentum with respect to time is equal to the applied forces.

The applied forces that we need to worry about are the pressure forces. First note that the pressure force on a unit surface area is normal to that surface.

The shallow water assumption contends that the vertical flows are negligible and that the magnitude of the pressure depends linearly with water depth (the hydrostatic assumption).

Consider the momentum in a moving region Ω_t,

$$\int_{\Omega_t} \mathbf{p}\, d\mathbf{x}$$

where $\mathbf{p} = \mathbf{u}h = (uh, vh)$.

The horizontal pressure term applied to the region Ω_t is given by

$$-\int_{\partial\Omega_t} \frac{1}{2} gh\mathbf{n}\, d\sigma - \int_{\Omega_t} gh\nabla b\, d\mathbf{x},$$

the first term is the horizontal pressure applied around the boundary of the domain Ω and second term is the horizontal component of the pressure due to the sloping bed.

Newton's second law then gives

$$\frac{d}{dt}\int_{\Omega_t} \mathbf{p}\, d\mathbf{x} = -\int_{\partial\Omega_t} \frac{1}{2} gh\mathbf{n}\, d\sigma - \int_{\Omega_t} gh\nabla b\, d\mathbf{x}.$$

Applying Theorem 1.1 to the first term and Stokes theorem to the first term on the LHS gives

$$\int_{\Omega_t} (\mathbf{p}_t + \nabla\cdot(\mathbf{u}\otimes\mathbf{p}))\, d\mathbf{x} = -\int_{\Omega_t} \nabla(\frac{1}{2}gh)\, d\mathbf{x} - \int_{\Omega_t} gh\nabla b\, d\mathbf{x}$$

and since this integral equation is true for all Ω_t), we conclude

$$\mathbf{p}_t + \nabla\cdot(\mathbf{u}\otimes\mathbf{p}) + \nabla(\frac{1}{2}gh) = -gh\nabla b.$$

Specifically, for the x component,

$$(uh)_t + \left(u\,uh + \frac{1}{2}gh^2\right)_x + (v\,uh)_y = -ghb_x \tag{12}$$

and for the y component,

$$(vh)_t + (u\,vh)_x + \left(v\,vh + \frac{1}{2}gh^2\right)_y = -ghb_y \tag{13}$$

Equation (11), together with equations (12) and (13) constitute the Shallow Water Wave Equations (SWWE).

1.5 1D Saint Venant equations

Let's consider a one-dimensional flow problem. That is, a flow which only has one component. A rotation of coordinates can always make it so that this component is x, such that the mass conservation equation reduces to

$$h_t + (uh)_x = 0 \tag{14}$$

and the momentum equations reduce to just one equation,

$$(uh)_t + \left(uvh + \frac{1}{2}gh^2\right)_x = -ghb_x \tag{15}$$

Together these two equations are know as the one dimensional shallow Water Wave Equations, or as the Saint Venant equations.

Often the momentum equation is written in primitive variables (u instead of uh), as

$$u_t + \left(\frac{1}{2}u^2 + gh\right)_x = -gb_x \tag{16}$$

Often equations (14) and (16) are easier to work with, particularly when finding analytic solutions to the shallow water wave equation.

1.6 Conservation and Balance Laws

The SWW equations in one (spatial) dimension has the form

$$\mathbf{U}_t + \mathbf{F}(\mathbf{U})_x = \mathbf{S}$$

where

$$\mathbf{U} = \begin{bmatrix} h \\ uh \end{bmatrix}, \quad \mathbf{F}(\mathbf{U}) = \begin{bmatrix} u\,h \\ u\,uh + \frac{1}{2}gh^2 \end{bmatrix} \quad \text{and} \quad \mathbf{S} = \begin{bmatrix} 0 \\ -ghb_x \end{bmatrix}.$$

$\mathbf{U}$ is the conserved quantity vector, $\mathbf{F}(\mathbf{U})$ is the associated flux vector, and $\mathbf{S}$ is the source term.

The SWW equations in two (spatial) dimensions has the form

$$\mathbf{U}_t + \mathbf{F}(\mathbf{U})_x + \mathbf{G}(\mathbf{U})_y = \mathbf{S}$$

where now

$$\mathbf{U} = \begin{bmatrix} h \\ uh \\ vh \end{bmatrix}, \quad \mathbf{F}(\mathbf{U}) = \begin{bmatrix} u\,h \\ u\,uh + \frac{1}{2}gh^2 \\ u\,vh \end{bmatrix}, \quad \mathbf{G}(\mathbf{U}) = \begin{bmatrix} v\,h \\ v\,uh \\ v\,vh + \frac{1}{2}gh^2 \end{bmatrix}, \text{and} \quad \mathbf{S} = \begin{bmatrix} 0 \\ -ghb_x \\ -ghb_y \end{bmatrix}.$$

where $\mathbf{F}(\mathbf{U})$ is the flux vector in x direction, $\mathbf{G}(\mathbf{U})$ is the flux vector in y direction.

These equations are written in conservation form, and the equations are called conversation laws (well actually conservation laws if the RHS is zero and called balance laws if the RHS is non-zero).

Provided the solutions to these equations are smooth, then the equations can be written in the equivalent non-conservation form in terms of primitive variables

$$\begin{bmatrix} h \\ u \end{bmatrix}_t + \begin{bmatrix} u & h \\ g & u \end{bmatrix} \begin{bmatrix} h \\ u \end{bmatrix}_x = \begin{bmatrix} 0 \\ -gb_x \end{bmatrix}$$

We want to understand the properties of the solutions of the SWW equations, but in the next section we study methods for solving simple non-linear scalar conservation laws.

2 Non-Linear Conservation Laws

2.1 Shock-wave Example

Burgers' equation models the time evolution of the concentration u of a quantity.

Consider the shock-wave equation, Burgers equation

$$u_t + uu_x = 0$$

$$u(x,0) = \begin{cases} 1, & x < -1, \\ -x, & -1 \leq x \leq 1, \\ -1, & x > 1. \end{cases}$$

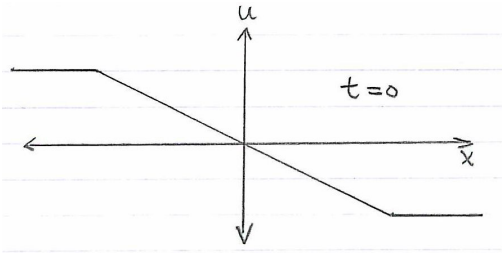


Figure 4: First example of initial condition.

General solution of problem with $u(x,0) = u_0(x)$.

Use characteristics

$$\frac{dx}{ds} = u(x,t)$$
$$\frac{dt}{ds} = 1$$
$$\frac{du}{ds} = 0$$

Initial conditions for ODE $(x(s,r), t(s,r), u(s,r))$

$$x(0,r) = r$$
$$y(0,r) = 0$$
$$u(0,r) = u_0(r)$$

The r parameter is being used as a parameter in the x direction (at least when $s = 0$.

Along a characteristic u is constant.

So $u(s,r) = u_0(r)$.

Now we can solve for $x(s, r)$

$$\frac{dx}{ds} = u_0(r)$$

So

$$\begin{aligned} x(s, r) &= x(0, r) + u_0(r)s \\ &= r + u_0(r)s \end{aligned}$$

Characteristics are straight lines emanating from the x-axis with speed $u_0(x)$ ($u_0(r)$).

So our method doesn't seem to work for $t > 1$, at least around the t-axis.

2.2 Advection Equations

Where do equation of this form come from? Consider the equation

$$u_t + vu_x = 0\,.$$

The characteristics are lines of the form

$$x(t) = x_0 + v(x, t)t\,.$$

Solution of the PDE is given by

$$u(x, t) = f(x - vt)\,.$$

The solution profile is advected with speed v.

So the equation models advection of marker particles (but does not model conservation of chemical species, pollutant etc.)

2.3 Integral Conservation Laws

Consider a pipe or river with velocity of fluid given by $v(x, t)$. Consider some chemical which has a concentration $u(x, t)$ kg/m at position x and time t. See Figures 6 and 7.

The amount of chemical in the region from x_0 to x_1 at time t_0 is given by

$$\int_{x_0}^{x_1} u(x, t_0)\, dx\,.$$

At a later time t_1, the amount of chemical in the section of river x_0 to x_1 is

$$\int_{x_0}^{x_1} u(x, t_1)\, dx\,.$$

The amount of chemical flowing past a point per second will be

$$\{\text{concentration}\} \times \{\text{velocity}\} = uv \qquad (\text{kg/s as } u \text{ kg/m} \times v \text{ m/s})\,.$$

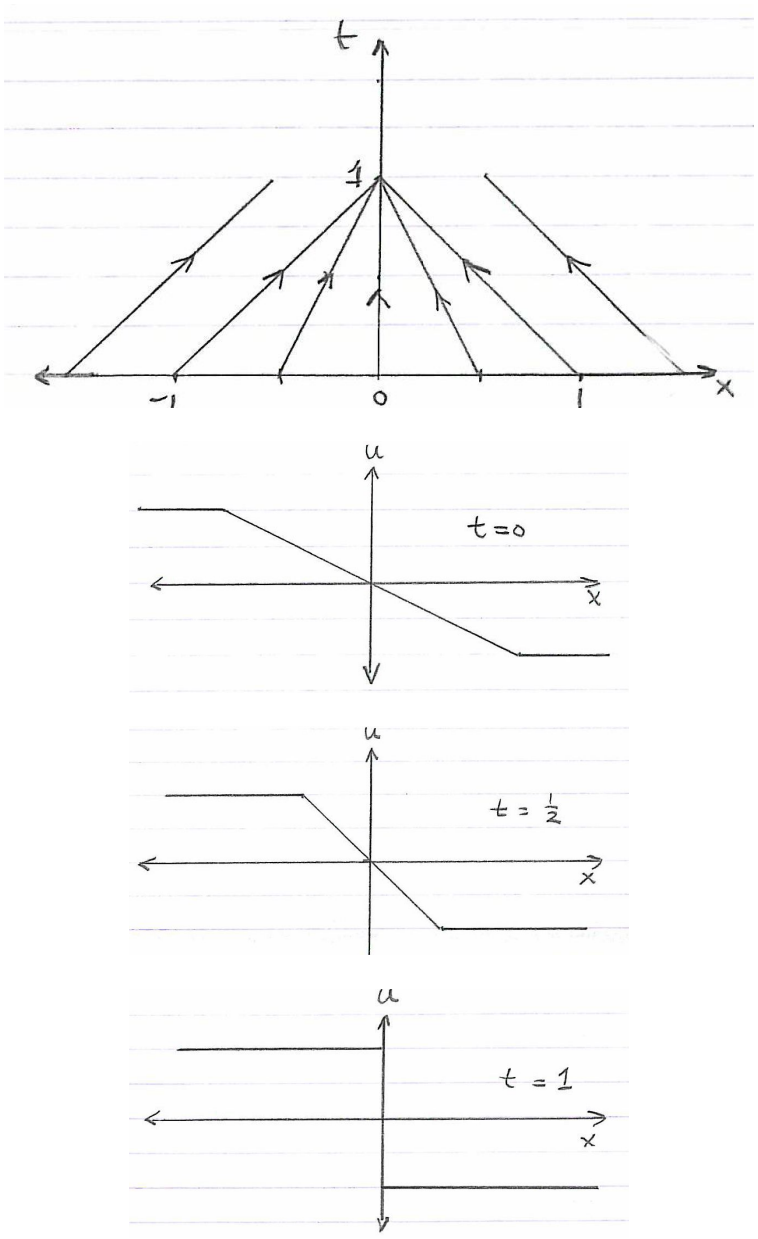


Figure 5: Solution obtained by characteristics (up to time $t = 1$).

So we have (in words) the conservation law

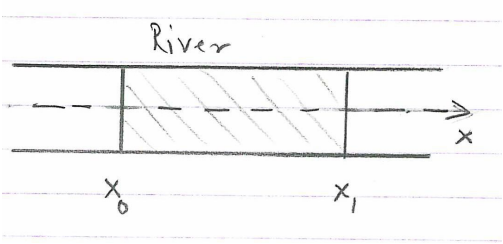


Figure 6: Flow in a pipe or river.

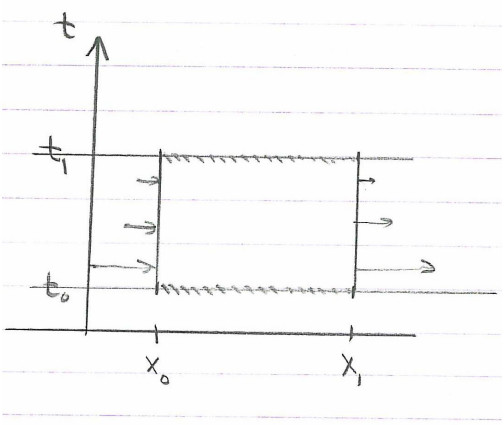


Figure 7: Flux diagram.

$$\left\{\begin{array}{c}\text{Amount of Chemical in}\\ \text{interval } [x_0, x_1] \text{ at time } t_1\end{array}\right\} = \left\{\begin{array}{c}\text{Amount of Chemical in}\\ \text{interval } [x_0, x_1] \text{ at time } t_0\end{array}\right\}$$

$$+ \left\{\begin{array}{c}\text{Amount of Chemical which}\\ \text{flows across } x_0 \text{ from time } t_0\\ \text{to } t_1\end{array}\right\}$$

$$- \left\{\begin{array}{c}\text{Amount of Chemical which}\\ \text{flows across } x_1 \text{ from time } t_0\\ \text{to } t_1\end{array}\right\}$$

In mathematical notation (which allows us to manipulate using all our skills from Calculus) is

$$\begin{aligned}\int_{x_0}^{x_1} u(x,t_1)\,dx = &\int_{x_0}^{x_1} u(x,t_0)\,dx \\ &+ \int_{t_0}^{t_1} v(x_0,t)u(x_0,t)\,dt \\ &- \int_{t_0}^{t_1} v(x_1,t)u(x_1,t)\,dt\,.\end{aligned}$$

This expression must hold for all choices of x_0, x_1, t_0 and t_1. We call this the integral form of the conservation law.

2.4 PDE Conservation Laws

Rearranging the conservation law gives

$$\int_{x_0}^{x_1} \left(u(x,t_1) - u(x,t_0)\right)\,dx + \int_{t_0}^{t_1} \left(v(x_1,t)u(x_1,t) - v(x_1,t)u(x_1,t)\right)\,dt = 0\,.$$

If u and v are smooth enough, these differences can be rewritten using the Fundamental theorem of Calculus to obtain

$$\int_{x_0}^{x_1}\int_{t_0}^{t_1} u_t\,dt\,dx + \int_{t_0}^{t_1}\int_{x_0}^{x_1} \left(vu\right)_x\,dx\,dt = 0\,.$$

Interchanging the order of integration we obtain

$$\int_{x_0}^{x_1}\int_{t_0}^{t_1} u_t + \left(vu\right)_x\,dt\,dx = 0\,,$$

for all choices of x_0, x_1, t_0 and t_1. So surely for smooth enough u and v the quantity u must satisfy the PDE

$$u_t + \left(vu\right)_x = 0\,.$$

For more general problems the flux uv may be more complicated. But if $f(u)$ is a flux function for a conserved quantity u, then u (if smooth enough) satisfies the PDE

$$u_t + f(u)_x = 0$$

f is the flux function and u is the conserved quantity.

A PDE of this form is called a Conservation Law.

Example: Burgers equation Our previous example

$$u_t + uu_x = 0$$

is not a PDE in conservation form. But the equation can be rewritten in the form

$$u_t + \left(\frac{1}{2}u^2\right)_x = 0$$

is a conservation equation (for the quantity u).

As long as the solution is smooth, either form produces the same solution. But the non-conservative form doesn't make sense for non-smooth functions, but the conservative form is essentially a short hand for the integral conservation equations which do make sense for non-smooth functions.

You do have to be careful, the equation

$$\left(\frac{1}{2}u^2\right)_t + \left(\frac{1}{3}u^3\right)_x = 0$$

is another conservation law, where $\frac{1}{2}u^2$ is the conserved quantity. It turns out that solutions to this conservation equation can be different to the conservation law for the quantity u.

Example: Burgers' Equation Burgers equation is a model for the non-linear conservation laws that arise in the equations for fluid flow, in particular the Navier-Stokes equation and the momentum equation given by Euler equations.

$$u_t + \left(\frac{1}{2}u^2\right)_x = 0.$$

For smooth solutions this is equivalent to the quasi-linear equation

$$u_t + uu_x = 0$$

We can solve the conservation equation using characteristics.

But what happens when the characteristics intersect. (i.e. when we get discontinuities). Go back to the derivation of the conservation law using the integral formulation

$$\int_{x_0}^{x_1} u(x,t_1)\,dx = \int_{x_0}^{x_1} u(x,t_0)\,dx + \int_{t_0}^{t_1} f(u(x_0,t))\,dt - \int_{t_0}^{t_1} f(u(x_1,t))\,dt$$

Suppose we start with an initial condition which is a step function, as shown in Figure 9.

$$u(x,0) = \begin{cases} u_L & x \le 0 \\ u_R & x > 0 \end{cases}$$

Can we find a solution to our integral equations with this initial condition. This type of problem, with a step function as initial conditions is called a Riemann problem.

Make a guess that the solution will remain a step function, but moves with speed s, as shown in Figure 10.

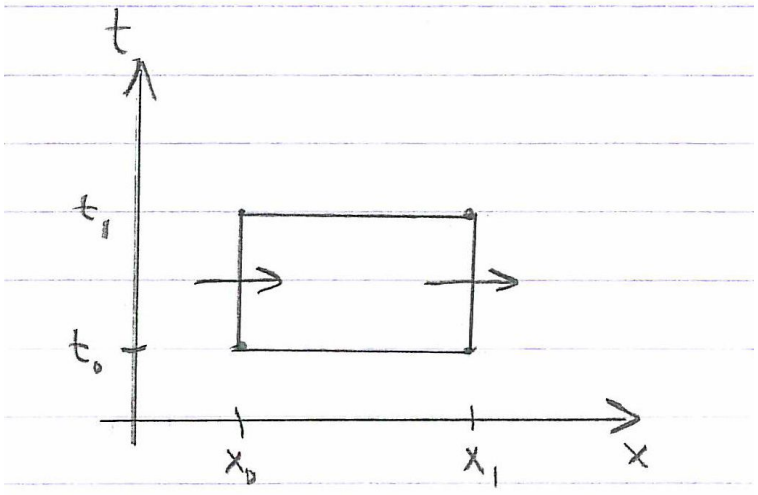


Figure 8: Domain associated with integral formulation of conservation law.

Choose a very simple region to apply integral equation. let $x_0 = 0$, $x_1 = s$, $t_0 = 0$ and $t_1 = 1$, see Figure 11

Then must have

$$\int_0^s u(x,1)\,dx = \int_0^s u(x,0)\,dx \tag{17}$$

$$+ \int_0^1 f(u(0,t))\,dt - \int_0^1 f(u(s,t))\,dt \tag{18}$$

Assumptions

- $u(x,1) = u_L$ for $0 \le x \le s$,
- $u(x,0) = u_R$ for $0 \le x \le s$,
- $u(0,y) = u_L$ for $0 \le y \le 1$,
- $u(s,y) = u_R$ for $0 \le y \le 1$.

So equation (18) implies

$$su_L = su_R + f(u_L) - f(u_R)$$

So

$$s = \frac{f(u_L) - f(u_R)}{u_L - u_R}$$

These are known as the Rankine Hugoniot conditions.

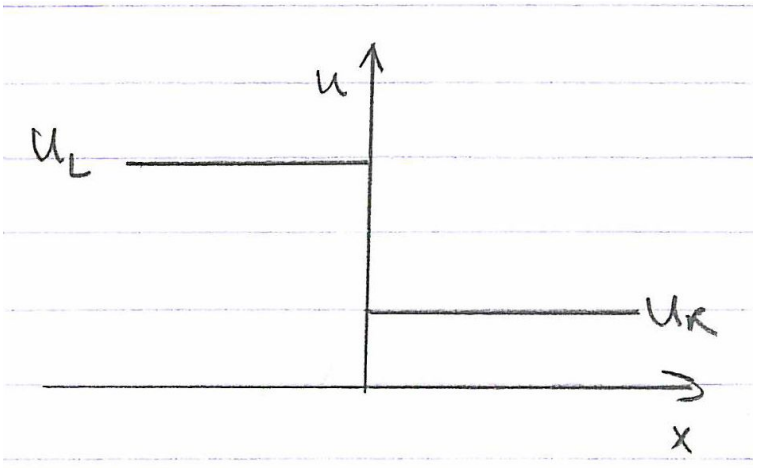


Figure 9: Step function initial condition as found in Riemann problems.

Example: Burgers Equation

$$u_t + \left(\frac{1}{2}u^2\right) = 0$$

So

$$s = \frac{\left(\frac{1}{2}u_L^2\right) - \left(\frac{1}{2}u_R^2\right)}{u_L - u_R} = \frac{1}{2}\frac{(u_L + u_R)(u_L - u_R)}{(u_L - u_R)}$$
$$= \frac{1}{2}(u_L + u_R)$$

Burgers equation with initial condition

$$u(x,0) = \begin{cases} 1 & x \leq 0 \\ 0 & x > 0 \end{cases}$$

$$s = \frac{1}{2}(u_L + u_R) = \frac{1}{2}$$

What about

$$u(x,0) = \begin{cases} 0 & x \leq 0 \\ 1 & x > 0 \end{cases}$$

Again the shock speed is

$$s = \frac{1}{2}(u_L + u_R) = \frac{1}{2}$$

This shock solution does solve the integral equations, but it is a non physical shock (information is generated from nowhere and it is unstable to slight perturbation in the initial conditions).

But there is another physically sensible solution, the rarefaction fan.

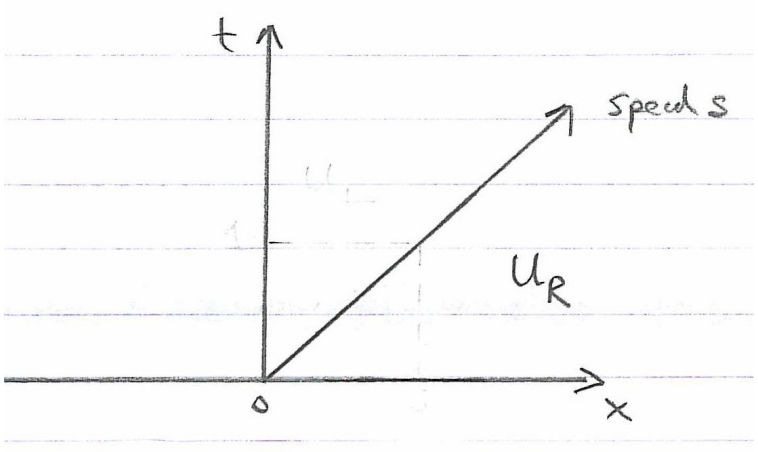


Figure 10: Left of shock $u = u_L$, right of shock $u = u_R$.

2.5 Rarefaction Fan

When the characteristics fan out (as opposed to collide as in the shock situation) it is reasonable to assume that the solution takes the form $u(x,t) = g(x/t)$ (where we have taken the origin to be a the location of the initial jump in the initial conditions.

Substituting the assumption $u(x,t) = g(x/t)$ into the quasi-linear form of the PDE, i.e.

$$u_t + f'(u)u_x = 0$$

leads to the requirement that g must satisfy

$$f'(g(\xi)) = \xi.$$

Example: Burgers equation In the case of Burgers equation, where $f'(u) = u$, this implies that $g(\xi) = \xi$. Consequently the rarefaction fan solution for Burgers equation is $u(x,t) = \frac{x}{t}$, in the region joining up the constant states.

$$u(x,0) = \begin{cases} 0 & x \leq 0 \\ 1 & x > 0 \end{cases}$$

In particular the solution is

$$u(x,t) = \begin{cases} 0 & x/t \leq 0 \\ x/t & 0 < x/t < 1 \\ 1 & x > 0. \end{cases}$$

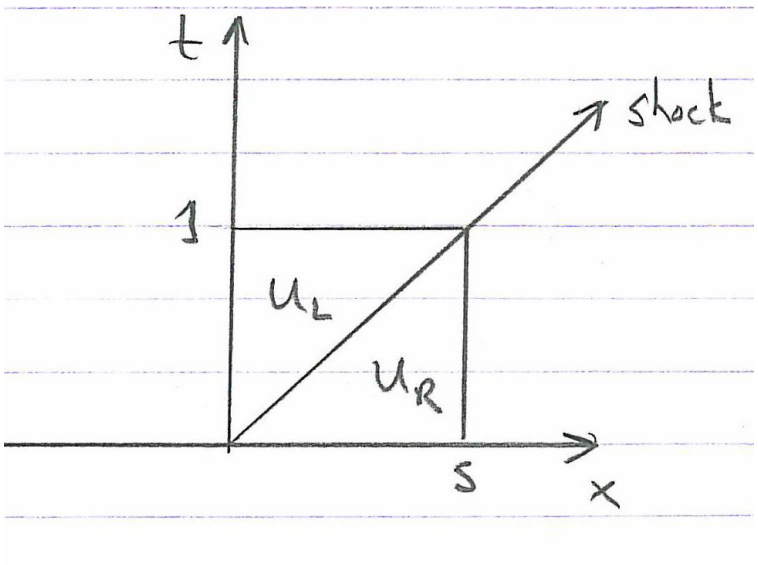


Figure 11: Simple domain used to derive Rankine-Hugoniot condition.

2.6 Problems

Problem 2.1 (Shock waves.)**.** Shock waves are often modeled by a first order (quasi-linear) PDE of the form

$$a(u)u_x + u_y = c(x, y)\,. \tag{19}$$

In this problem we will see that the method of characteristics can be applied to such equations as well.

(a) Consider the equation

$$a(u)u_x + u_y = 0\,. \tag{20}$$

Obviously, any constant function $u(x, y) = u_0$ is a solution of (20). Let $u(x, y) = u_0 + \epsilon v(x, y)$, where $\epsilon \ll 1$. Find the most general solution of this form to (20), to first order in ϵ, and argue that these represent waves moving with velocity $a(u_0)$ (where the variable y has the interpretation of time).

(b) Use the method of characteristics to show that the most general solution of (20) is implicitly given by

$$u = F(x - a(u)y)\,. \tag{21}$$

Find the (implicitly defined) solution to (20) for the initial condition $u(x, 0) = \phi(x)$.

(c) Solve the initial value problem (for small $y \geq 0$)

$$uu_x + u_y = 0\,, \qquad u(x, 0) = -x\,. \tag{22}$$

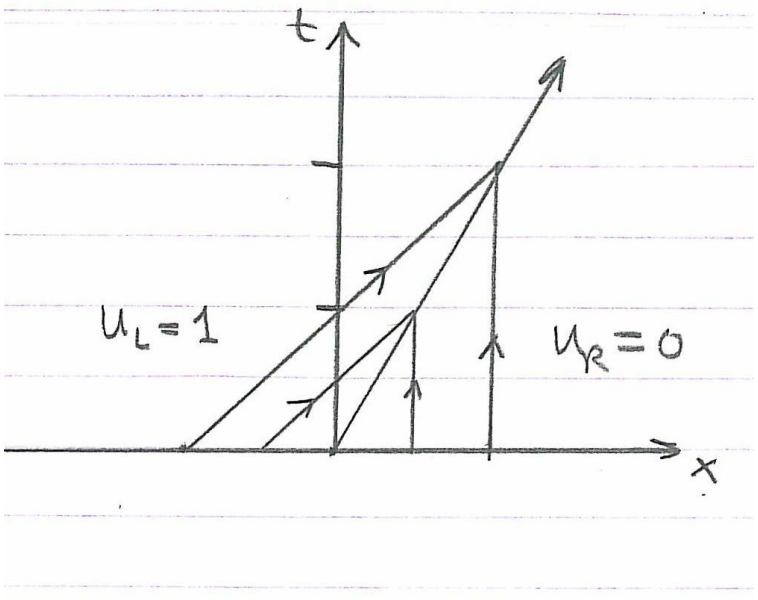


Figure 12: A physically sensible shock is formed when characteristics intersect.

Draw the lines in the (x, y)-plane along which the solution is constant. Do you see the development of a 'shock' for $y \geq 0$?

(d) Solve

$$uu_x + u_y = 1\,, \qquad u(x, x) = \frac{x}{2}\,. \tag{23}$$

[*Hint:* It is convenient to choose the parameter s along the characteristic curves such that for $s = 0$ we are at the intersection of the characteristic curve with the curve $(x(t), y(t)) = (t, t)$ for which $u(x(t), y(t)) = \frac{t}{2}$.]

Problem 2.2 (Riemann Problem for Fourth Power Flux.)**.** Consider the conservation law

$$u_t + \left(\frac{1}{32}u^4\right)_x = 0$$

with initial condition

$$u(x, 0) = \begin{cases} u_L, & x < 0, \\ u_R, & x \geq 0. \end{cases}$$

For $u_L > u_R$ we would expect a solution with a shock and for $u_L < u_R$ a rarefaction fan.

(a) Find the Rankine-Hugoniot shock speed condition for this problem.

(b) Hence find the solution to the problem with the initial condition $u_L = 2$, $u_R = -1$. Draw a picture of the x-t plane showing representative characteristic curves and the shock curve for this solution.

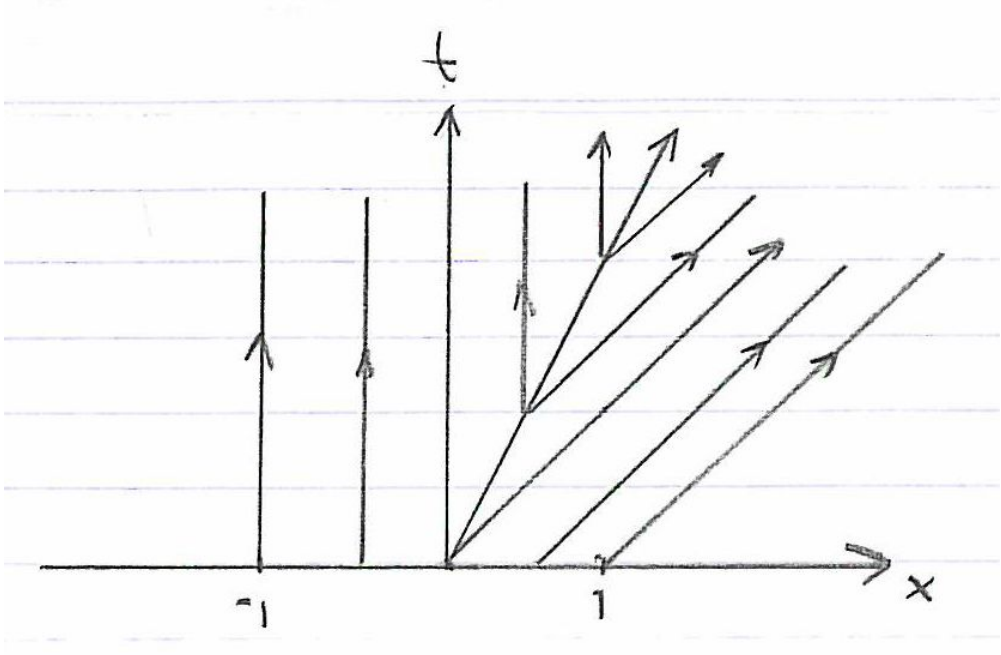


Figure 13: Non physical shock solution of Riemann Problem.

(c) Find the expression for the rarefaction fan for this problem. In particular find the function g such that $u(x,t) = g(x/t)$ is a solution of this PDE.

(d) Hence find the solution to the problem with the initial condition $u_L = -1$, $u_R = 2$. Draw a picture of the x-t plane showing representative characteristic curves and the rarefaction fan for this solution.

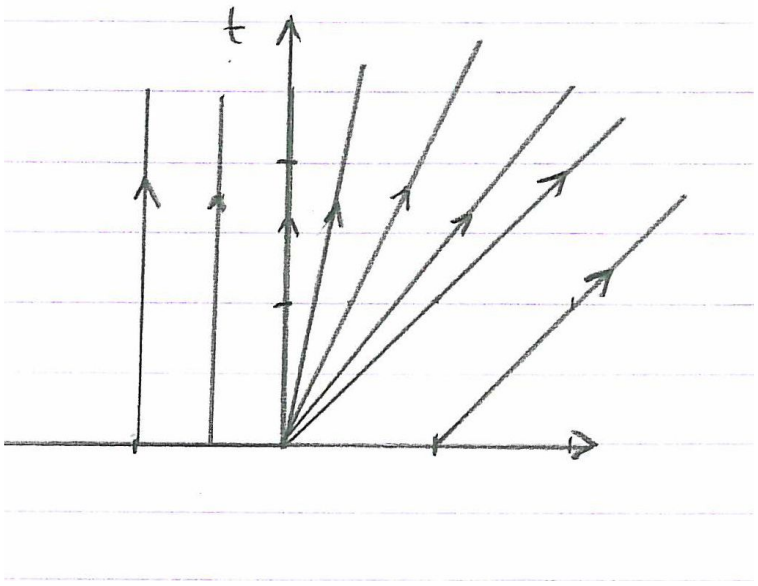


Figure 14: Rarefaction fan solution of Riemann Problem for Burgers Equation.

3 Dam Break Problem

The dam-break problem is an idealized instantaneous collapse of a dam in a wide horizontal frictionless rectangular channel. The resulting wave is considered to be one-dimensional.

3.1 Stoker's Solution

Stoker (1905-1992)[112], an American Engineer, extended Ritter's dam-break solution, which is described in this Section, to the propagation of a dam-break wave into a region of still tail-water. He derived analytical expressions for the free-surface profile and for the wave-front speed and depth in terms of the initial upstream and downstream water depths. The water on both sides of the dam is assumed to be at stagnant. At time $t = 0$s, the dam is suddenly removed.

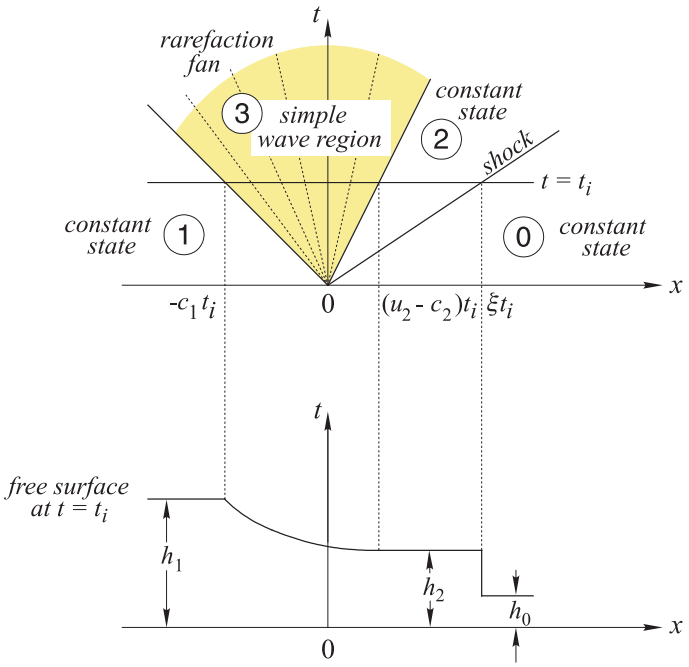


Figure 15: Stoker's solution to the behaviour of the dam break problem at time $t > 0$ with stagnant water downstream.

For $t > 0$s, the resulting water surface profile has four distinct regions, see Figure 15.

Each region is separated either by a (*i*) *shock wave* or (*ii*) *rarefaction fan*, also known as a *rarefaction fan*. The region of undisturbed water in region (1), where $u_1 = 0$m/s is separated from the *simple wave region* (3), which has a parabolic free-surface by a characteristic given by

$$x = -c_1 t. \tag{24}$$

The simple wave region is separated from the *constant wave region* (2), were the water depth

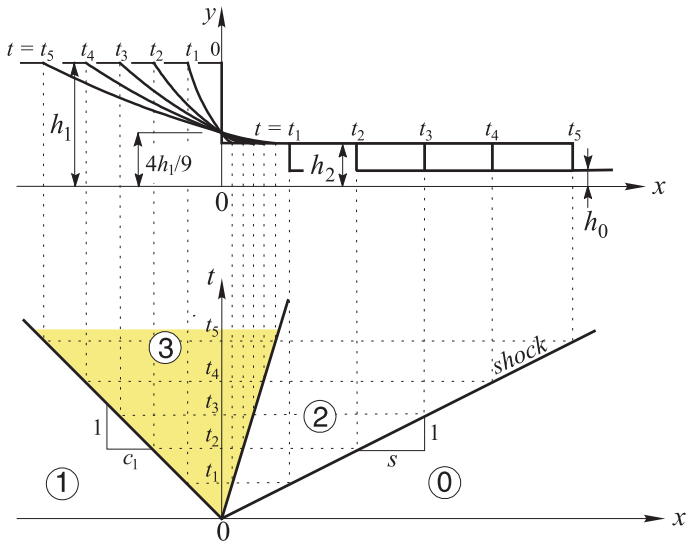


Figure 16: Behaviour of the dam break problem at time $t = t_i$ with stagnant water downstream.

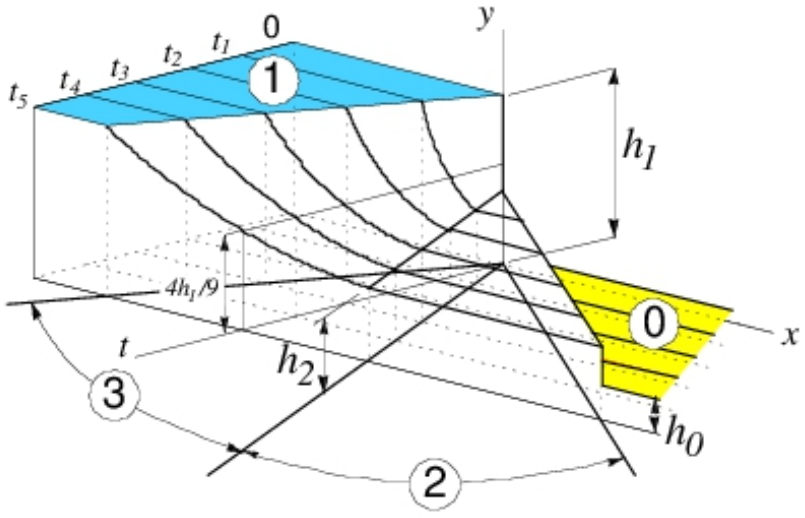


Figure 17: Behaviour of the dam break problem at time $t = t_i$ with stagnant water downstream.

remains constant. The characteristic separating these regions is given by

$$x = (u_2 - c_2)t. \tag{25}$$

Finally a sharp front connects the third region to the downstream undisturbed region. The discontinuity between states (0) and (2) is a surge moving at the speed ξ. The moving surge can be reduced to a stationary surge or hydraulic jump by having the co-ordinate system moving at the shock speed ξ. In this case, the conservation of mass through a hydraulic jump is given by

$$h_2(\xi - u_2) = h_0(\xi - u_0). \tag{26}$$

Since $u_0 = 0$m/s and $c = \sqrt{gh}$ then

$$\frac{c_0^2}{c_2^2} = \frac{h_0}{h_2} = \frac{\xi - u_2}{\xi}. \tag{27}$$

The conservation of momentum across a hydraulic jump is

$$\frac{1}{2}gh_2^2 - \frac{1}{2}gh_0^2 = h_2(\xi - u_2)(u_2 - u_0).$$

Making use of (26) then

$$2\xi(\xi - u_2) = (c_2^2 + c_0^2). \tag{28}$$

Eliminating c_2 between (27) and (28) and solving the resulting quadratic equation for the velocity of the fluid in region (2) u_2, then

$$u_2 = \xi - \frac{c_0^2}{4\xi}\left(1 + \sqrt{1 + \frac{8\xi^2}{c_0^2}}\right). \tag{29}$$

Eliminating u_2 between (27) and (28) and solving the resulting quadratic equation for c_2, then

$$\frac{c_2^2}{c_0^2} = \frac{h_2}{h_0} = \frac{1}{2}\sqrt{1 + \frac{8\xi^2}{c_0^2}} - \frac{1}{2}. \tag{30}$$

To solve these equations for c_2 (or h_2) and u_2, ξ or a relationship between c_2 (or h_2) and u_2 must be known. The additional relationship is obtained by considering the characteristic originating from the simple wave region (2). Along these characteristics

$$u_2 + 2c_2 = 2c_1. \tag{31}$$

Using this relationship with (27) and (28) provides the following non-linear expression for the speed of the surge, ξ in terms of the known quantities c_0 and c_1.

$$c_1 = \frac{\xi}{2} - \frac{c_0^2}{8\xi}\left[1 + \sqrt{1 + \frac{8\xi^2}{c_0^2}}\right] + \left[\frac{c_0^2}{2}\sqrt{1 + \frac{8\xi^2}{c_0^2}} - \frac{c_0^2}{2}\right]^{1/2}. \tag{32}$$

Solving this equation for ξ then (29) and (30) can be used to obtain u_2 and c_2 (and h_2) respectively.

In the centered simple wave region or rarefaction fan, region (3)

$$\frac{dx}{dt} = \frac{x}{t} = u_3 - c_3 \tag{33}$$

combined with

$$u_3 + 2c_3 = 2c_1$$

gives

$$u_3 = \frac{2}{3}\left(c_1 + \frac{x}{t}\right) \tag{34}$$

and

$$c_3 = \frac{2}{3}c_1 - \frac{x}{3t}. \tag{35}$$

Given the initial condition h_0, h_1, u_0 and u_1 at $t = 0$s, then the water surface profile and the velocity at any time $t > 0$s, can be determined using (24) to (35). When $h_0 = 0$m/s there is no hydraulic jump. At $x = 0$m, from (35) $c_3 = 2c_1/3$ or $h_3 = 4h_1/9$. This can only occur if the rarefaction fan straddles the $t-$axis, so that $u_2 - c_2 > 0$. At $x = 0$m, from (34) and (35) $u_3 = c_3$ and $Fr = 1$, which corresponds to critical flow conditions. The water depth at the location of the dam remains constant for all time and the depth only depends on the upstream quiescent flow depth, h_1. Equation (35) indicates that the water surface profile for $-c_1 t < x < (u_2 - c_2)t$ is also only dependent on the upstream water depth. This is the water profile within the simple wave region or rarefaction fan, see Figure 15.

More recently, Wu *et al.*[131] derived a single algebraic expression for the depth of the shock, h_2 which depends only on the ratio of the initial upstream and downstream water depths, h_0 and h_1.

Recalling that $u_0 = 0$m/s, then from (26)

$$\xi = \frac{h_2 u_2}{h_2 - h_0}. \tag{36}$$

Substituting (31), $c_1 = \sqrt{gh_1}$ and $c_2 = \sqrt{gh_2}$ into (36) so that

$$\xi = \frac{2h_2}{h_2 - h_0}\left(\sqrt{gh_1} - \sqrt{gh_2}\right)$$

which is used to eliminate ξ in (30) so that

$$\frac{h_2}{h_0} = \frac{1}{2}\sqrt{1 + 8\left(\frac{2h_2}{h_2 - h_0}\frac{\sqrt{gh_1} - \sqrt{gh_2}}{\sqrt{gh_0}}\right)^2} - \frac{1}{2}$$

where $c_0 = \sqrt{gh_0}$. Rearranging and dividing by h_1^3 to yield

$$\left(\frac{h_2}{h_1}\right)^3 - 9\left(\frac{h_2}{h_1}\right)^2\left(\frac{h_0}{h_1}\right) + 16\left(\frac{h_2}{h_1}\right)^{3/2}\left(\frac{h_0}{h_1}\right) - \left(\frac{h_2}{h_1}\right)\left(\frac{h_0}{h_1}\right)\left(\frac{h_0}{h_1} + 8\right) + \left(\frac{h_0}{h_1}\right)^3 = 0 \tag{37}$$

which shows that the height of the shock, h_2 is only dependent on the upstream and downstream flow depths, h_1 and h_2 respectively. This equation is much simpler to solve than the iterative solution of (32) followed by the solution of (30) for h_2. The monotone relationship between h_2/h_1 and h_0/h_1, given by (37) is illustrated in Figure 18.

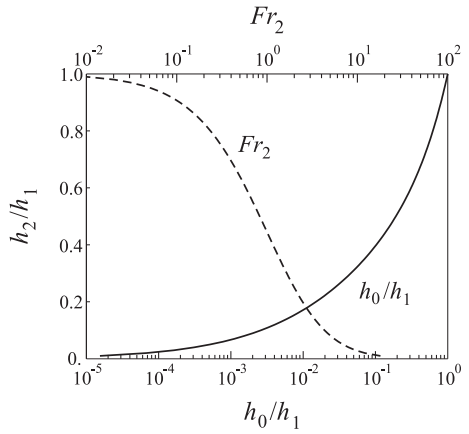


Figure 18: Relationship between the ratio of the water depth at the shock and the upstream water depth, h_2/h_1 to the ratio of the upstream water depth and the downstream water depth, h_1/h_0 and to the shock Froude number, Fr_2 for the dam break problem.

Equation (37) can be used to establish the maximum height of the shock wave and the ratio h_0/h_1 that this will occur. These quantities have been plotted in Figure 19. The maximum height of the shock wave $h_2 - h_0 \approx 0.133043 h_1$ and this occurs when the ratio of the upstream to downstream water depth $h_0/h_1 \approx 0.175692$. This is a useful results because it indicates the water level in the reservoir upstream of a dam that should be avoided in the event of a breach of a dam.

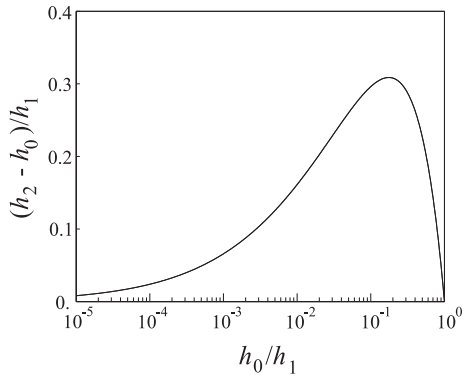


Figure 19: Relationship between the ratio of the shock height and the upstream water depth, $(h_2 - h_0)/h_1$ to the ratio of the upstream water depth and the downstream water depth, h_1/h_0 for the dam break problem.

Although the dam break will result in a shock traveling downstream when there is a finite water depth downstream, the shock does not necessarily result in supercritical flows downstream of the breach. From (31)

$$Fr_2 = \frac{u_2}{c_2} = 2\frac{c_1}{c_2} - 2 = 2\sqrt{\frac{h_1}{h_2}} - 2.$$

The Froude number for a range of water depth ratio, h_2/h_1 are also plotted in Figure 18. When $Fr_2 = 1$, $h_2/h_1 = 4/9$ or $h_0/h_1 \approx 0.1379$ and critical flow exist below the breach. When $h_0/h_1 < 0.1379$ the flow downstream of the dam is supercritical and when $h_0/h_1 > 0.1379$ the downstream flow remains subcritical.

Figure 18 can be used to establish the upstream to downstream water depth that is required to produce downstream flow with the desired Froude number. It is interesting to note that the maximum loss in energy across the jump, shown in Figure **??** does not coincide with the maximum shock height shown in Figure 19.

3.2 Ritter's Dry Bed Solution

Ritter, a German Engineer in 1892[99] considered a channel of infinite length, horizontal bottom, rectangular in cross-section and with a smooth bed and walls. The dam in the channel is assumed to be orthogonal to the longitudinal axis of the channel. Upstream of the dam there is a finite depth of stagnant water, whereas downstream the channel is dry.

As a limiting case, $c_0 = 0$m/s for a dry bed downstream, (30) gives $c_2/c_0 \to \infty$ and $c_2 = 0$m/s. Using (31) then $u_2 = 2c_1$ and (27) yields $\xi = u_2 = 2c_1$. In this case, the constant state with u_2 and c_2 and the discontinuity disappear, see Figure 20. A parabolic free-surface profile that connects the upstream undisturbed region and the downstream propagating wave front. For this special case, in which one of the initial states has no water results in a Riemann problem which consists of one wave only. This is always a rarefaction fan.

The parabolic water surface can be calculated using (35) and the corresponding velocity is given by (34). The wave profile is confined to the interval

$$-1 \leq \frac{x}{c_1 t} \leq 2.$$

From (34) the velocity, u_3 varies linearly between zero at the negative front, which travels at the speed c_1 and the avancing front, which travels at the speed $2c_1$. At the dam site, the flow is critical and the water depth and velocity remain constant in time at $4/9h_1$ and $2/3c_1$ respectively, yielding a discharge equal to $8/27h_1c_1$. From Figure 18 flow is subcritical upstream of the dam and supercritical downstream.

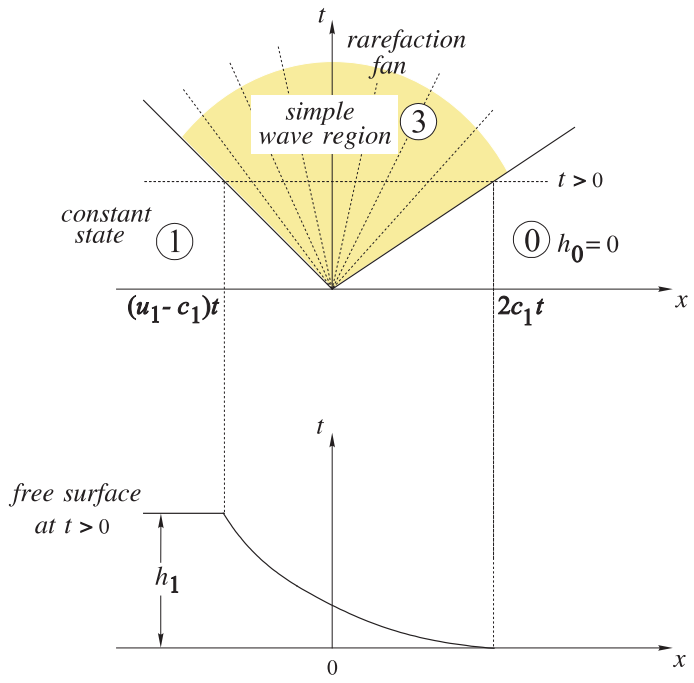


Figure 20: Behaviour of the dam break problem at time $t > 0$ on a dry bed.

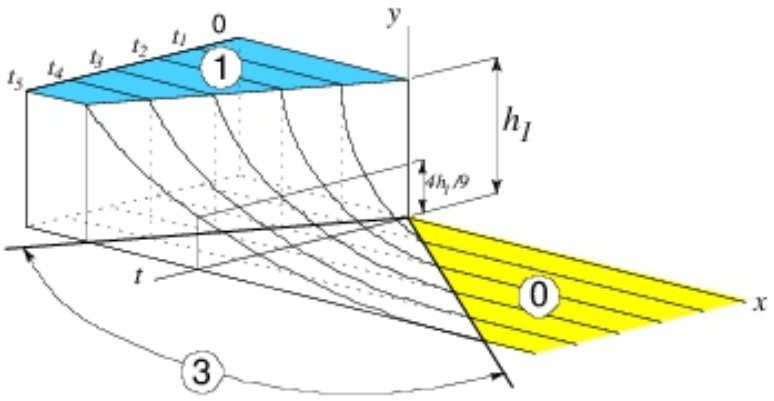


Figure 21: Ritter's solution to the behaviour of the dam break problem at time $t = t_i$ with dry downstream bed.

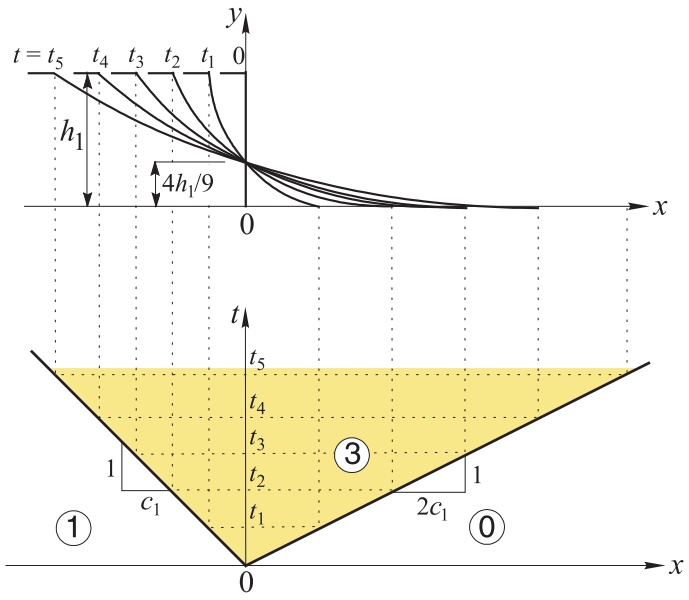


Figure 22: Behaviour of the dam break problem at time $t > 0$ with downstream dry bed.

4 Conservative Numerical Schemes

Solving the conservative form of the governing equations is important, so is ensuring that the numerical scheme is also conservative. For example, the inviscid Burgers equation can be discretized by the conservative upwind scheme to give

$$u_j^{n+1} = u_j^n - \frac{\Delta t}{2\Delta x}(u_j^n + u_{j-1}^n)(u_j^n - u_{j-1}^n) \tag{38}$$

where the flux $F(u) = u^2/2$. Using the non-conservative upwind method

$$u_j^{n+1} = u_j^n - \frac{\Delta t}{2\Delta x} u_j^n (u_j^n - u_{j-1}^n).$$

Both equations are identical for smooth solutions but not for problems with shock waves. This is illustrated in Figure 23. The shock is moving at the wrong speed when the Burgers equation is solved with the non-conservative upwind scheme. The true shock speed is $S = (u_l + u_r)/2 = 5\text{m/s}$ solution sought at $t = 10\text{s}$, $\Delta t = 0.01\text{s}$, $\Delta x = 0.1\text{m}$ and the computational domain is subdivided into $1,000$ segments. Since the governing equations are written in conservative form, it would

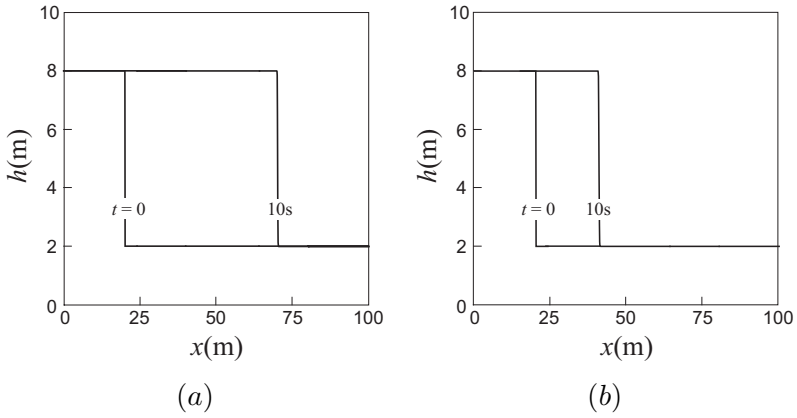


Figure 23: Solution of the inviscid Burgers equation using a first-order upwind scheme written in (a) conservative and (b) non-conservative form.

seem reasonable that the numerical scheme should also possess this property. Consider the one-dimensional problem. Partition the space $[0, T] \times [-\infty, \infty]$. Let Δt and Δx be the temporal and spatial step lengths respectively. Denote the numerical solution at the point (x_j, t^n) by $\mathbf{U}_j^{n+1}$, where $x_j = j\Delta x$, $t^n = n\Delta t$, $(j = 0, \pm 1, \pm 2, \ldots, n = 0, 1, 2, \ldots)$. Then a numerical scheme that can be written in the form of

$$\mathbf{U}_j^{n+1} = \mathbf{U}_j^n - \frac{\Delta x}{\Delta t}(\mathbf{f}_{j+1/2}^n - \mathbf{f}_{j-1/2}^n) \tag{39}$$

are considered to be in the desirable *conservative form* because it is a descritized form of the integral form of the conservation laws, which simply states that the rate of change in $\mathbf{U}$ in a cell is equal to the difference in the fluxes $\mathbf{f}$ entering the cell. It is the integral form of the governing equations,

which allows discontinuities in the solution. Methods differ in the way the intercell flux $\mathbf{f}^n_{j\pm1/2}$ is estimated, see Figure 24.

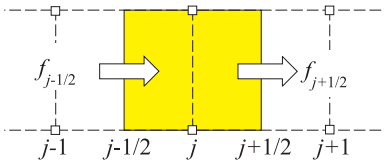


Figure 24: Inter-cell fluxes for cell j.

5 Modified Equivalent Partial Differential Equation

Finite difference approximations of partial differential equations generally exhibit properties that are not present in the differential equation. This behaviour may obscure the true response of the system being modelled. Therefore, it is imperative that numerical modellers have an appreciation and understanding of this behaviour. There are a number of useful techniques available to investigate the behaviour of finite difference approximations. However, many of these approaches assume that the solution is smooth. In many practical problems the initial or boundary conditions are not continuous. We examine the behaviour of finite difference analogues of the advection equation near a discontinuity. This behaviour is examined by investigating the properties of the modified partial differential equation of a finite difference scheme. Analytical solutions are given for the modified partial differential equations of first-, second- and third-order finite difference schemes used to solve the advection equation near a discontinuity. These analytical solutions provide a rational explanation of the behaviour of a number of numerical schemes in current use. These analytical solutions are also used to demonstrate the accuracy of these schemes for the simulation of problems containing discontinuities. The results indicate that Qth-order accurate finite difference schemes have only $O(\Delta x^{Q/(Q+1)})$ accuracy near discontinuities. The notation $O(\Delta x^Q)$ denotes a quantity which goes to zero with the same rate as Δx^Q, or faster, when Δx goes to zero. Therefore, a front can only be modelled to first-order accuracy at best. The transport of pollutants in aquatic systems involves the spread of the pollutant by advection and by diffusion. In many situations, the advection usually dominates the transport processes.

We are primarily interested in the solution of the constant coefficient advection equation

$$\frac{\partial h_0}{\partial t} + u\frac{\partial h_0}{\partial x} = 0 \quad 0 < x < \infty, \quad t > 0 \tag{40}$$

in which h_0 is a scalar quantity, u is the constant fluid velocity, x is the distance and t is the time, using finite difference schemes.

The analytical solution to this equation is simply; $h_0(x,t) = h_0(x-ut,0)$ the profile is advected at the speed u without deformation.

The numerical solution of equation (40) has received considerable attention in the literature and there are numerous finite difference schemes available for its solution. Although this is a very simple equation, many schemes fail dramatically. They have very demanding stability criteria, or exhibit undesirable properties such as spurious oscillations, artificial dispersion or numerical diffusion not present in the exact solution[100]. It is important for numerical modellers to understand the behaviour of the finite difference analogue of equation (40) otherwise this behaviour may obscure the true response of the system being modelled.

There are a number of techniques for investigating the accuracy and behaviour of numerical schemes. The most common is Fourier analysis where a Fourier component in space is substituted into the finite difference equations, and the evolution in time of the Fourier[114] modes are

observed or their group velocity[121][114]. This approach does not resolve the discontinuity very well. The process of using only a finite number of modes itself introduces oscillations known as *Gibb's phenomenon.* The oscillations around the shock are related to the well known *Gibb's phenomenon* in Fourier analysis. These oscillations may obscure the spurious oscillations introduced by the finite difference approximation. Therefore, these approaches may not provide any insight into the behaviour of finite difference schemes near discontinuities. In many practical problems the concentration profile is not continuous.

A Taylor series expansion of a finite difference scheme will generally reveal that the scheme is a more accurate approximation to a higher-order partial differential equation than the original equation. This equation is known as the *modified equivalent partial differential equation*[128][104][89]. By retaining only the leading term in the modified equivalent partial differential equation, it is possible to estimate the accuracy and establish the stability constraints for the numerical scheme[52]. We have called this equation the *modified partial differential equation.* The properties of this simpler partial differential equation can also be used to provide some insight into the behaviour of the finite difference approximation near a discontinuity. This has not been appreciated by a number of numerical modellers. For example, the linear advection equation was solved by Chen and Falconer[17], Tamamidis and Assanis[116] and Comini *et al.*[20] using a number of numerical schemes. All observe behaviour in the finite difference solution near discontinuities that is not predicted by the governing partial differential equation. They make the observation that some high-order numerical schemes produce oscillations trailing the discontinuity, whilst others produce oscillations that lead the discontinuity. They seem oblivious to the fact that the behaviour of a numerical scheme near a discontinuity can be deduced from the modified partial differential equation. This is even more apparent from the latter example[20], where the modified partial differential equation for a number of schemes were provided and were not used to explain the behaviour of the finite difference scheme. Motivated by this apparent oversight, analytical solutions are provided to the modified partial differential equation of several well known finite difference approximations to the linear advection equation. These analytical solutions are used to provide a rational explanation of the behaviour of these schemes near a discontinuity.

There are very few examples in the literature where the behaviour of finite difference approximations have been examined by investigating the properties of their modified partial differential equation near a discontinuity. Two notable exception are Hedstrom[48] and Chin and Hendstrom[18]. Based on their investigations, a number of well known numerical schemes for solving the advection equation are analyized. Analytical solutions to the modified partial differential equations are used to show that the *order of accuracy* of a finite difference scheme is significantly reduced for problems containing discontinuities.

In the next section we provide a general expression for the modified equivalent partial differential equation for any two-level finite difference scheme and discuss the implications of the dominant error term. We develop explicit solutions to the modified partial differential equation for first- second-

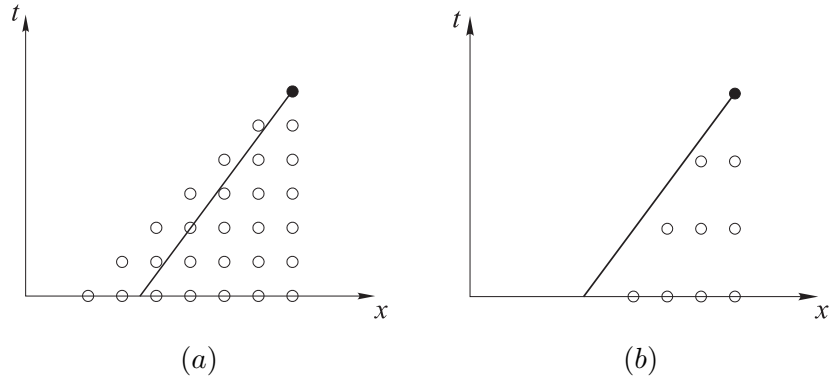


Figure 25: The computational points that are included in the domain of dependence of a computational point for (a) a stable explicit scheme and (b) an unstable explicit scheme, after Durran[27].

and third-order finite difference schemes for problems involving discontinuities. The properties of these solutions are investigated and their behaviour compared with that observed from a numerical scheme. The analytical solutions are then used to establish the expected order of accuracy of a numerical scheme near a discontinuity.

The *modified equivalent partial differential equation* for any finite difference approximations of equation (40) involving a maximum of six computational nodes over two time level is given by[90]

$$\frac{\partial h_0}{\partial t} + u\frac{\partial h_0}{\partial x} = -\sum_{q=2}^{\infty} \frac{u\Delta x^{q-1}}{q!}\eta_q(Cr)\frac{\partial^q h_0}{\partial x^q} \tag{41}$$

in which the *cell Courant number*[2] $Cr = u\Delta t/\Delta x$ and the functions η_q are dependent on u and the grid spacing Δt and Δx used in the numerical scheme.

If $\eta_2(Cr) > 0$ the finite difference scheme introduces numerical diffusion which has the affect of smearing the profile. The forward time upwind[100] scheme is an example of a first-order scheme where $\eta_2(Cr) = (Cr - 1)$. Another is the Lax-Friedrichs scheme where $\eta_2(Cr) = (Cr^2 - 1)/Cr$. Both these schemes introduce numerical diffusion and their modified partial differential equation is a diffusion equation.

The modified partial differential equation is a dispersion equation when $\eta_2(Cr) = 0$ and $\eta_3(Cr) \neq 0$. The finite difference scheme is second-order accurate. Second-order schemes eliminate the numerical diffusion and provide a better resolution to discontinuities in the solution than first-order schemes, however dispersion is introduced. Numerical dispersion is associated with different frequencies that make up a travelling wave, propagating at different speeds. This may result in oscillations in the numerical solution in the vicinity of steep gradients in the profile. Examples of second-order finite difference schemes are the Lax-Wendroff[73] scheme, where $\eta_3(Cr) = (1 - Cr^2)$ and the Beam-Warming[127] scheme, where $\eta_3(Cr) = -(2 - 3Cr + Cr^2)$.

[2]The Courant-Friedrichs-Lewy number is named after the German-American Mathematicians; Richard Courant (1888-1973), Kurt Otto Friedrichs (1901-1982) and Hans Lewy (1905-1988).

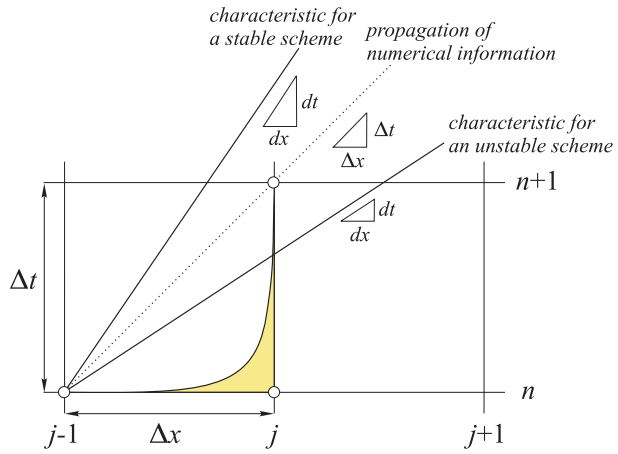


Figure 26: Relationship between the propagation of information and the computational stencile for an explicit scheme that will result in an unstable or stable numerical scheme.

The behaviour of several finite difference schemes can be established by finding analytical solutions to their modified partial differential equation.

5.1 Advective-diffusion equation

The linear advective-diffusion equation is given by

$$\frac{\partial h_2}{\partial t} + u\frac{\partial h_2}{\partial x} = D\frac{\partial^2 h_2}{\partial x^2} \quad 0 < x < \infty, \quad t > 0 \tag{42}$$

and the initial conditions are given by $h_2(x,0) = f(x)$ for $-\infty < x < \infty$ where D is the constant diffusion coefficient,

$$f(x) = \begin{cases} A & 0 < x \\ 0 & x \geq 0 \end{cases} \tag{43}$$

and A is the maximum value of the profile. This initial value problem can be solved using the Fourier transform in space [3] to provide the well known analytical solution

$$h_2(x,t) = \frac{A}{2}\text{erfc}\left(\frac{x-ut}{2\sqrt{Dt}}\right) \tag{44}$$

in which erfc is the *complementary error function* [4][1]. Consider a simple example, where $A = u = D = t = 1$. The diffusion of a front is shown in Figure 27. These results are similar to the results

[3] Fourier transform in space:

$$\mathcal{F}\{f(\alpha,t)\} = \frac{1}{2\pi}\int_{-\infty}^{\infty} f(x,t)\exp(-i\alpha x)\,dx$$

[4] Complementary error function:

$$\text{erfc}(z) = \frac{2}{\sqrt{\pi}}\int_{z}^{\infty}\exp(-t^2)\,dt$$

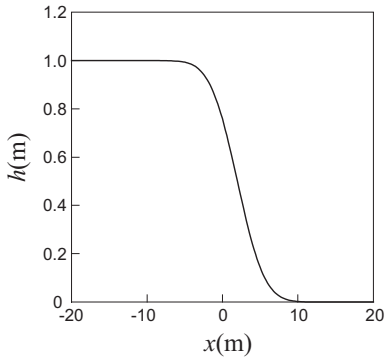


Figure 27: The diffusion of a step function.

produced for the evolution of an advancing front using first-order schemes. Consider the following test problem

$$h(x, t = 0) = A = \begin{cases} 100 & 0 \leq x \leq 50\Delta x \\ \\ 0 & \text{otherwise} \end{cases} \tag{45}$$

solved using the forward time upwind scheme. The step profile is advected by the fluid travelling at $u = 0.25$m/s. In the model $\Delta x = 1$m and $\Delta t = 1$s therefore, $Cr = 0.25$, and the simulated and analytical solutions are compared when $t = 200$s. The analytical solution to this problem and the results from the first-order forward time upwind scheme is shown in Figure 28. The simulated

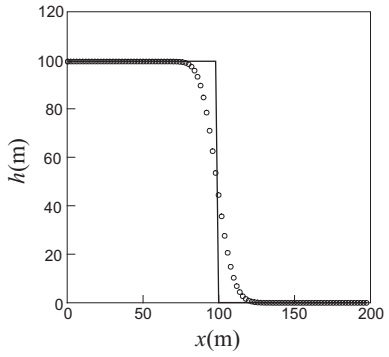


Figure 28: Analytical and forward time upwind solution of the advection equation.

profile closely resembles that predicted by equation (44), where the profile remains monotone but spreads with time. This is a well known result for first-order finite difference schemes.

5.2 Advective-dispersion equation

The linear advective-dispersion equation is given by

$$\frac{\partial h_3}{\partial t} + u\frac{\partial h_3}{\partial x} = -D\frac{\partial^3 h_3}{\partial x^3} \quad 0 < x < \infty, \quad t > 0 \tag{46}$$

and the initial condition is given by $h_3(x,0) = f(x)$ for $-\infty < x < \infty$. D is the dispersion coefficient in this equation. This equation is the linearized *Korteweg-de Vries* equation which plays an important role in the theory of water waves (see for example Whitham[130]).

Using the Fourier transform in space, equation (46) has the following solution (see, Appendix A)

$$h_3(x,t) = \frac{1}{(3Dt)^{1/3}} \int_{-\infty}^{\infty} f(\xi)\mathrm{Ai}\left(\frac{x-\xi-ut}{(3Dt)^{1/3}}\right) d\xi \tag{47}$$

where Ai is the *Airy function*[5][1], which is an integral representation of the solution of the initial boundary value problem. For the initial conditions given by (43) then equation (47) becomes

$$h_3(x,t) = \frac{A}{(3Dt)^{1/3}} \int_{-\infty}^{0} \mathrm{Ai}\left(\frac{x-\xi-ut}{(3Dt)^{1/3}}\right) d\xi. \tag{48}$$

This solution indicates that for a moving front at $x = ut$, oscillations should be observed for negative ξ and the profile decays exponentially for positive ξ. This is illustrated by considering a simple example, where $u = t = A = 1$ and $D = 1/3$. Introducing a new variable $\zeta = x - \xi - 1$, then equation (48) becomes

$$h_3(x,t) = \int_0^{\infty} \mathrm{Ai}(\zeta)\, d\zeta - \int_0^{x-1} \mathrm{Ai}(\zeta)\, d\zeta. \tag{49}$$

This can be simplified by making use of equation 10.4.82 in Abramowitz and Stegun[1] where

$$\int_0^{\infty} \mathrm{Ai}(z)\, dz = \frac{1}{3}$$

then equation (49) becomes

$$h_3(x,t) = \frac{1}{3} - \int_0^{x-1} \mathrm{Ai}(\zeta)\, d\zeta.$$

Evaluating the integral numerically, the resulting concentration profile is illustrated in Figure 29. These results are typical of those obtained for the solution of the advection equation using second-order schemes in which D is positive. Oscillations are observed behind the advancing front and the profile decays exponentially forward of the front. This is the case for the second-order Lax-Wendroff scheme, the results of which are shown in Figure 30 for the above test problem, equation (45). If

[5] Airy function:

$$\mathrm{Ai}(\pm(3a)^{-1/3}x) = \tfrac{(3a)^{1/3}}{\pi} \int_0^{\infty} \cos(at^3 \pm xt)\, dt.$$

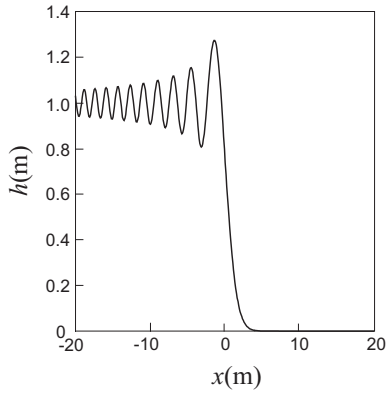


Figure 29: The dispersion of a step function for positive dispersion coefficient, D.

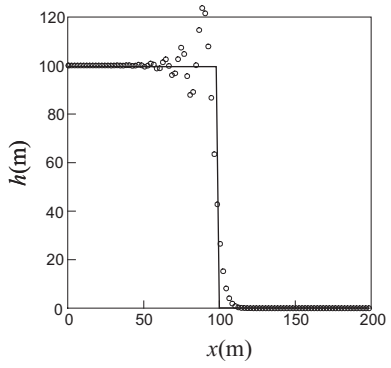


Figure 30: Analytical and Lax-Wendroff solution of the advection equation

the diffusion coefficient is negative, the linear advective-dispersion equation is written in the form

$$\frac{\partial h_3}{\partial t} + u\frac{\partial h_3}{\partial x} = D\frac{\partial^3 h_3}{\partial x^3} \quad 0 < x < \infty, \quad t > 0$$

with the initial conditions $h_3(x,0) = f(x)$ for $-\infty < x < \infty$. This equation has an analytical solution analogous to equation (48) given by

$$h_3(x,t) = \frac{C}{(3Dt)^{1/3}} \int_{-\infty}^{0} \mathrm{Ai}\left(\frac{ut + \xi - x}{(3Dt)^{1/3}}\right) d\xi \tag{50}$$

for the initial conditions given by (43). Using the test problem, where $u = t = A = 1$ and $D = 1/3$, then equation (50) becomes

$$h_3(x,t) = \int_0^{\infty} \mathrm{Ai}(-\zeta)\, d\zeta - \int_0^{x-1} \mathrm{Ai}(-\zeta)\, d\zeta \tag{51}$$

which can be simplified by making use of equation 10.4.83 in Abramowitz and Stegun[1]

$$\int_0^{\infty} \mathrm{Ai}(-z)\, dz = \frac{2}{3}.$$

Equation (51) becomes

$$h_3(x,t) = \frac{2}{3} - \int_0^{x-1} \mathrm{Ai}(-\zeta)\, d\zeta.$$

Evaluating the integral numerically, the resulting concentration profile is illustrated in Figure 31. In this case the analytical solution indicates that oscillations are observed prior to the arrival and

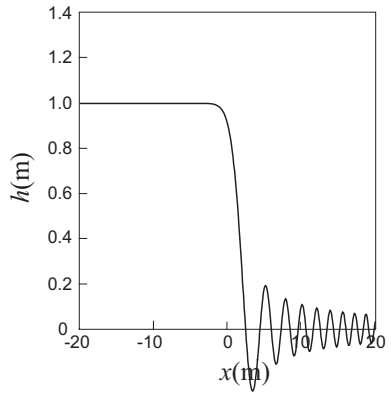


Figure 31: The dispersion of a step function for negative dispersion coefficient, D.

monotone after the arrival of the advancing front. Similar profiles are observed for the solution of the advection equation using second-order schemes such as the Beam and Warming schemes. The solution of the test problem using the Beam and Warming scheme is illustrated in Figure 32. As predicted oscillations occur downstream of the discontinuity and like the analytical solution the

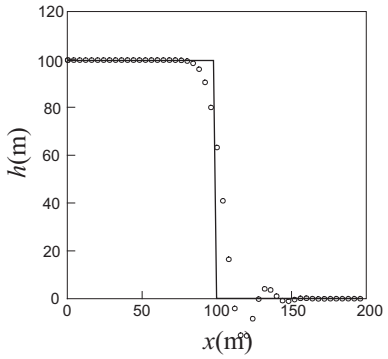


Figure 32: Analytical and Beam and Warming solution of the advection equation.

numerical scheme is also monotone upstream of the discontinuity. These results are not well known.

There is however, a major distinction between the results shown in Figures 29 and 31 and those obtained from a second-order numerical scheme. Since a numerical scheme has a finite domain, boundary conditions are imposed on the initial value problem. These boundary conditions must be satisfied. Therefore, the oscillations are dampened so that the boundary conditions are satisfied. These constraints do not exist in the analytical solution where the oscillations extend to infinity. Other higher-order even terms introduced by the finite difference approximation may also dampen these oscillations.

5.3 Higher-order equations

The above procedure can be generalised for any higher-order scheme. Only the third-order accurate scheme is considered here. In this case

$$h_4(x,t) = \int_{-\infty}^{\infty} f(\xi) \left[\frac{1}{2\pi} \int_{-\infty}^{\infty} \exp\left(-\alpha^4 Dt + i\alpha(x-\xi-ut)\right)\, d\alpha \right] d\xi.$$

For the initial conditions given by (43) then

$$h_4(x,t) = A \int_{-\infty}^{0} \left[\frac{1}{\pi} \int_{0}^{\infty} \exp\left(-\alpha^4 Dt\right) \cos\left(\alpha(x-\xi-ut)\right)\, d\alpha \right] d\xi \tag{52}$$

which makes use of *Euler's formula*[6]

$$\exp(z) = \exp(x)\exp(iy) = \exp(x)(\cos(y) + i\sin(y)) \quad \text{and} \quad \int_{\infty}^{-\infty} \sin(y)\, dy = 0.$$

The inner integral has been evaluated numerically for $D = 1$, $t = 1$ and $ut = \xi$. The results are shown in Figure 33. The function is symmetrical about $x = 0$ and oscillatory with its amplitude damping rapidly. Solving equation (52) numerically for $u = t = A = 1$ and $D = 1/3$, produced the results shown in Figure 34. The simulated profile is symmetrical with oscillations occurring both

[6] Euler's formula:

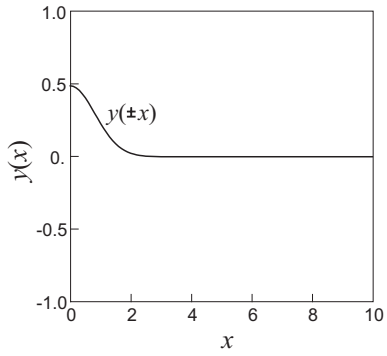


Figure 33: Evaluation of the inner integral in equation (52).

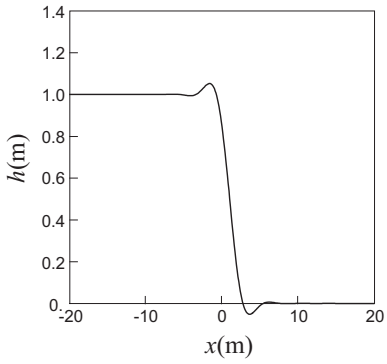


Figure 34: Analytical solution to the fourth-order equivalent modified partial differential equation.

upstream and downstream of the discontinuity. These results are typical of the results produced by third-order schemes which do not use shape preserving interpolation or flux limiters. Using the test problem, the results from the third-order accurate Holly and Priessmann[53] scheme are shown in Figure 35. Oscillations occur in the simulated profile both upstream and downstream of the

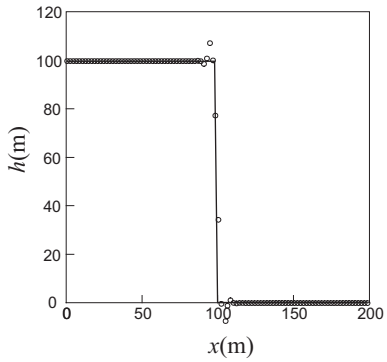


Figure 35: Analytical and Holly and Preissmann solution of the advection equation.

discontinuity. In addition, these oscillations are symmetrical about the discontinuity.

It is well known that numerical diffusion is the direct result of even-order derivatives of the dominant error term in the modified equivalent partial differential equation, and numerical dispersion is the direct result of the odd-order derivatives. However, for many difference schemes dispersion dominates diffusion[121]. This may be the case for many odd-order accurate schemes, such as the Holly and Preissmann scheme, which exhibit features of a dispersive scheme rather than a diffusive scheme.

Although this book concentrates on the numerical schemes for the solution of the shallow water wave equation, analytical solutions to their approximations to the shallow water wave equation provides considerable insight to the behaviour of finite difference approximations to the shallow water wave equation. Kevorkian[63] is an excellent resource for analytical solutions to partial differential equations, as is Myint-U and Debnath[86].

5.4 Accuracy in the presence of a discontinuity

Analytical solutions have been provided for the solution near a discontinuity of the modified partial differential equations of first-, second- and third-order finite difference approximation of the linear advection equation. These analytical solutions provide a rational explanation of the behaviour of finite difference approximations to the linear advection equation near a discontinuity. First-order schemes introduce numerical diffusion, which smears the profile. Second-order schemes introduce numerical dispersion which manifest as oscillations in the solution. Oscillations are only observed

$$\exp(i\alpha) = \cos(\alpha) + i\sin(\alpha).$$

after the arrival of a front for second-order schemes with positive dispersion coefficients. The profile remains monotone before the arrival of the front. When the dispersion coefficient is negative, oscillations are only observed before the arrival of the front and it remains monotone behind the front. For third-order schemes, oscillations are symmetrical about the discontinuity. The analytical solutions also show that the formal accuracy of finite difference schemes is dramatically reduced if there is a discontinuity in the initial or boundary conditions. A Qth-order accurate finite difference scheme will approximate a discontinuity with accuracy $O(\Delta x^{Q/(Q+1)})$ in the L_1 norm sense. Therefore, a front can only be modelled to first-order accuracy at best, which of course, is the best that can be expected.

The predicted accuracy of a numerical scheme using traditional techniques, such as the modified equivalent partial differential equations is dramatically reduced in the presence of a discontinuity in the initial or boundary conditions. The above analytical solutions can be used to estimate the accuracy of a finite difference scheme near discontinuities.

The L_1 norm[114];

$$L_1 = \int_{-\infty}^{\infty} |f(x) - h_n(x,t)| \, dx$$

is used as a measure of the error between the exact, $f(x)$ and numerical solutions, $h_n(x,t)$ of the advection equation.

5.5 First-order schemes

Consider a first-order finite difference scheme. The standardised L_1 norm between the exact and numerical solution is given by

$$\| \, h_0(x,t) - h_2(x,t) \, \|_{L_1} = \int_{-\infty}^{\infty} |f(x-ut) - h_2(x,t)| \, dx. \tag{53}$$

Recalling the initial conditions are

$$f(x) = \begin{cases} 1 & x < 0 \\ \\ 0 & x \geq 0 \end{cases} \tag{54}$$

then (53) becomes (see Appendix B)

$$\| \, h_0(x,t) - h_2(x,t) \, \|_{L_1} = A_1 \sqrt{Dt}$$

where A_1 is a constant. The L_1norm behaves like $t^{1/2}$. The dispersion coefficient, D is a function of u, Δx and Δt in first-order schemes. When Cr and u are given, then $D \propto \Delta x$ and the L_1 norm becomes

$$\| \, h_0(x,t) - h_2(x,t) \, \|_{L_1} = A_2 \sqrt{\Delta x t}$$

in which A_2 is a constant. The error decays only like $\Delta x^{1/2}$ even though the modified equivalent partial differential equation indicates that the method is formally *first-order accurate*.

5.6 Second-order schemes

Standard second-order finite difference schemes introduce a dispersion term. The analytical solution of the advective-dispersion equation for a unit step function is given by equation (47) in which $f(\xi)$ is the initial conditions.

Using the L_1 norm as a measure of the error between the exact and numerical solution of the advection equation, then

$$\| h_0(x,t) - h_3(x,t) \|_{L_1} = \int_{-\infty}^{\infty} |f(x-ut) - h_3(x,t)| \, dx.$$

Using (48) and for (54) this equation has the solution (see Appendix B)

$$\| h_0(x,t) - h_3(x,t) \|_{L_1} \leq A_1 (3Dt)^{1/3}. \tag{55}$$

This suggests that at best, a second-order scheme behaves like $t^{1/3}$ if a discontinuity exists in the initial conditions.

The behaviour of second-order numerical schemes for which the dispersion coefficient D is known, can now be established by substituting for D in equation (55). Generally, for a given Cr and u, $D \propto \Delta x^2$ and equation (55) becomes

$$\| h_0(x,t) - h_3(x,t) \|_{L_1} \leq A_2 \Delta x^{2/3} t^{1/3}$$

in which A_2 is a constant. Therefore the L_1norm decays like $\Delta x^{2/3}$ for second-order accurate schemes. These results have been confirmed using numerical experiments. For example the rate of convergence of the forward time upwind, Lax-Wendroff and the third-order Holly and Priessmann schemes, shown in Figure 36 were obtained for the following problem; $h_0(x,0) = A$ for $0 \leq x \leq 45$m, $h_0(0,t) = A$, $h_0(L,t) = 0$, $A = 100$, $u = 0.25$m/s and $L = 200$m. Fifteen data sets were used, containing between 201 to $20,001$ computational nodes ($\Delta x = 1$ to 0.001m). The L_1 norm between the analytical and numerical solution when, $t = 200$s was calculated using all the computational points in each data set. The Courant-Friedrichs-Lewy number, $Cr = 0.25$ remains constant for all data sets. These numerical results confirm the above predictions. The L_1norm for the first-order scheme behaves like $\Delta x^{1/2}$ and for the second-order scheme like $\Delta x^{2/3}$. Using similar analysis, it is possible to show that for Q^{th}-order schemes the L_1 norm decays like $\Delta x^{Q/(Q+1)}$.

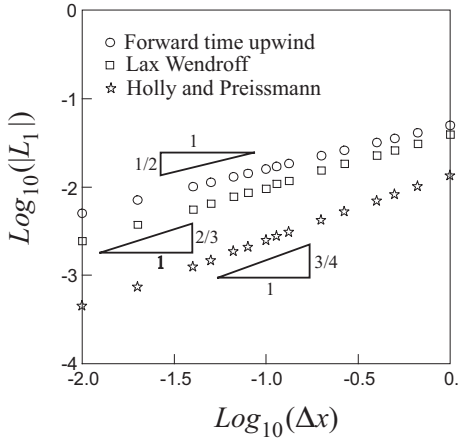


Figure 36: Rate of convergence of several finite difference schemes for the solution of the advection of a discontinuous profile.

6 Naive Finite Difference Schemes

Numerical solution of differential equations is very commonplace and in fact necessary in order to solve a variety of problems of practical interest which defy analytical treatment. Several techniques to solve the differential equations numerically have become popular such as finite difference methods, finite element methods, spectral methods and boundary element methods. A wealth of literature is available in the form of books and journals on these methods. We confine ourselves to the essential basis of finite difference methods in this part, we start with the basic concepts of finite difference schemes.

With the exception of *analogue computers*, modern computers cannot deal with continuous functions (see, for example Karplus[62]). They can only deal with discrete values which approximate a continuous function. The approximation of a continuous function by a finite number of discrete values is the essence of *finite difference schemes*. The basic idea of the finite difference method is to replace a function $h(x,t)$ in two-dimensions with h at discrete points in the $x-t$ plane. The approximation also extends to the derivatives of $h(x,t)$. For example, there are many ways to approximate the first derivative $\partial h/\partial x$ at node j shown in Figure 37 in terms of its neighbouring nodal values. Some approximations are better than others. The *forward difference*, involving function values at x_j and $x_j + \Delta x$ and the *backward differencing*, involving x_j and $x_j - \Delta x$ are first-order approximations. Using information at $x_j - \Delta x$ and $x_j + \Delta x$ results in the second-order *centered* or *central differencing.* The significance of the order of the approximations is that if $h(x)$ were a first or second-order polynomial, then the central differencing

$$\left.\frac{\partial h}{\partial x}\right|_{x_j} = \frac{h(x_{j+1}) - h(x_{j-1})}{2\Delta x}$$

would correspond to the exact solution.

There is no best method for obtaining approximate finite difference formulae. The only requirement is that the formula, having been obtained, must pass certain tests of *accuracy*, *consistency*, *stability* and *convergence.*

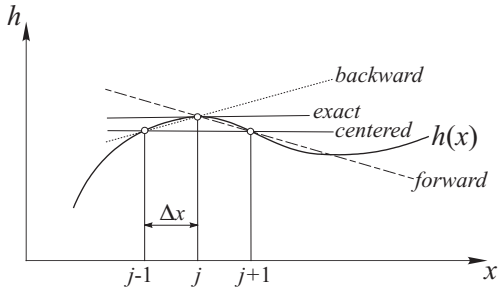


Figure 37: The exact, forward, backward and central finite difference approximation of $h(x_j)$.

Convergence

The problem of convergence of a finite difference method consists of finding the conditions under which

$$E = h(x,t) - \widetilde{h}(x,t)$$

is the difference between the exact solution, $h(x,t)$ of the partial differential equation and the solution to the finite difference approximation, $\widetilde{h}(x,t)$ at a fixed point (x,t), tend to zero uniformly as Δx and $\Delta t \to 0$. Unfortunately, conventional methods for establishing the convergence properties of a finite difference approximation is very difficult to apply. An indication of convergence may be the accuracy of the finite difference solutions of simplified problems for which exact solutions are known, or by the application of *Lax's equivalence theorem* if consistency and stability can be determined.

If the *Courant-Friedrichs-Lewy* condition is not satisfied, convergence cannot take place when $\Delta x \to 0$ with $\Delta t = p\Delta x$, where p is a constant.

The domain of dependence includes all grid points which contribute to the calculated value at the point (x_j, t^n).

In the present case of $Cr < 1$, in which the computational domain of dependence contains the physical domain of dependence, all the information required to determine the condition at P is included in the computational stencils so that the numerical scheme is stable. The results is inaccurate because of the inclusion of some unnecessary information originating from the region between the numerical and physical domains of dependence. If $Cr > 1$, the characteristics the domain of dependence of the computational scheme does not enclose the domain of dependence of the physical system. In this case only a part of the needed information is used to determine the solution at P and the computational scheme is unstable. When $Cr = 1$ the computational and physical domains of dependence coincide and the numerical solution produces the exact solution.

There are numerous finite difference schemes which have been used to solve partial differential equations. Many of these schemes do not explicitly consider the properties of the governing equations. These are considered as *naive finite difference schemes*. Since the shallow water wave equations are conservative laws, only conservative numerical schemes have been examined. These are the first-order Lax-Friedrichs, the second-order Lax-Wendroff and MacCormack schemes.

Let us introduce in the one-dimensional $x - t$ plane a mesh size Δx and Δt and denote

$$x_j = j\Delta x, \quad t^n = n\Delta t, \quad h(x_j, t^n) = h_j^n.$$

Any continuous differentiable function $h(x)$ can, in the vicinity x_j be expressed as a Taylor series.

A finite difference scheme reduces to an algorithm for the computation of $h(x)$, assuming that h_j^n is known. The value of $h(x)$ can be expressed in terms of h_j^n with the Taylor series expansion

in space

$$h(x) = h_j^n + (x - x_j)\left.\frac{\partial h}{\partial x}\right|_j^n + \frac{(x - x_j)^2}{2}\left.\frac{\partial^2 h}{\partial x^2}\right|_j^n + \frac{(x - x_j)^3}{6}\left.\frac{\partial^3 h}{\partial x^3}\right|_j^n + O(\Delta x^4) \tag{56}$$

to terms of order Δx^3 inclusive and where $O(\Delta x^4)$ denotes terms of the order of Δx^4 have been neglected and represents the truncation error.

Obviously from the definition of a derivative

$$\left.\frac{\partial h}{\partial x}\right|_{x_j} = \lim_{\Delta x \to 0} \frac{h(x_j) - h(x_{j-1})}{\Delta x}$$

the closer the computational points the greater the accuracy of the approximation. However, there is a limit on the number of computational points that can be used to approximate derivatives in the partial differential equation. Therefore, each finite difference approximation introduces *discretisation errors*. Using these expansions, one can obtain approximate expressions for the first and higher derivatives at point x_j in terms of the function values at neighbouring points. For example, writing (56) for $h(x)$ at x_{j+1} then

$$\left.\frac{\partial h}{\partial x}\right|_j^n = \frac{h_{j+1}^n - h_j^n}{\Delta x} - \Delta x \left.\frac{\partial^2 h}{\partial x^2}\right|_j^n + O(\Delta x^2) \tag{57}$$

which is *downwind differencing*.

Alternatively, writing $h(x)$ at x_{j-1} then

$$\left.\frac{\partial h}{\partial x}\right|_j^n = \frac{h_j^n - h_{j-1}^n}{\Delta x} + \Delta x \left.\frac{\partial^2 h}{\partial x^2}\right|_j^n + O(\Delta x^2) \tag{58}$$

which is *upwind differencing*.

Higher-order approximations for the first derivative are possible by using more points to eliminate more of the higher order terms in the Taylor series. If $h(x)$ at x_{j+1} and $h(x)$ at x_{j-1} were used in (56), then

$$\left.\frac{\partial h}{\partial x}\right|_j^n = \frac{h_{j+1}^n - h_{j-1}^n}{2\Delta x} - \frac{\Delta x^2}{6}\left.\frac{\partial^3 h}{\partial x^3}\right|_j^n + O(\Delta x^3) \tag{59}$$

which is *centered difference* approximation for $\partial h/\partial x|_j^n$. Equation (59) could also be obtained by adding (57) with (58).

Second and higher derivatives can also be approximated in a similar manner. For example, to estimate the second derivative at a point, one may use the approximation to the first derivative twice. For example, the second derivative can be expressed using

$$\left.\frac{\partial^2 h}{\partial x^2}\right|_j^n = \frac{\left.\frac{\partial h}{\partial x}\right|_{j+1}^n - \left.\frac{\partial h}{\partial x}\right|_j^n}{\Delta x}$$

with the first derivatives replaced by their finite difference approximations. Using the upwind scheme to approximate $\partial h/\partial x|_j^n$ and $\partial h/\partial x|_{j+1}^n$ then

$$\left.\frac{\partial^2 h}{\partial x^2}\right|_j^n = \frac{h_{j+1}^n - 2h_j^n - h_{j-1}^n}{\Delta x^2} - \frac{\Delta x^2}{12}\left.\frac{\partial^4 h}{\partial x^4}\right|_j^n + O(\Delta x^3).$$

By repeatedly applying two first-order schemes, that are of opposite bias, the leading truncation error terms are cancelled and the resulting becomes second-order accurate.

This is only possible with constant grid spacing, Δx.

The order of accuracy referred to here is called the *local order of accuracy* defined as $E_j = h(x_j) - h_j$ for all j does not include the propagation of errors outside the computational stencil nor does it include the influence of boundary conditions. *Global order of accuracy*, $\sum_{\forall j} E_j$ is usually lower than the local order of accuracy.

We have already demonstrated that in the presence of a discontinuity, the order of accuracy is reduced even further.

If Δx is small, the higher order terms will also be small except in the situation where the higher derivatives are locally very large. Ignoring the latter possibility and retaining only the first term, these equations represent possible finite difference expressions for the first derivative. The first derivative approximation can be made higher-order accurate at the cost of introducing more computational points in the finite difference stencil.

Similar expressions can be obtained for the temporal derivatives. Examples of approximations for $\partial h/\partial t|_j^n$ over two time levels are

$$\left.\frac{\partial h}{\partial x}\right|_j^n = \frac{h_j^{n+1} - h_j^n}{\Delta t} + O(\Delta t) \qquad \text{Forward or forward Euler} \tag{60}$$

$$\left.\frac{\partial h}{\partial x}\right|_j^n = \frac{h_j^n - h_j^{n-1}}{\Delta t} + O(\Delta t) \qquad \text{Backward or backward Euler} \tag{61}$$

$$\left.\frac{\partial h}{\partial x}\right|_j^n = \frac{h_j^{n+1} - h_j^n}{\Delta t/2} + O(\Delta t^2) \qquad \text{Centered or trapezoidal.} \tag{62}$$

To illustrate one way that these approximations to the spatial and temporal derivatives can be used to produce a finite difference approximation, consider the advection equation of the form (**??**) to which an artificial dissipation term, $\nu\partial^2 h/\partial x^2$ has been added. Namely

$$\frac{\partial h}{\partial t} + u\frac{\partial h}{\partial x} = \nu\frac{\partial^2 h}{\partial x^2}.$$

Using centered differencing to approximate both x derivatives to second-order in Δx and forward difference for the t derivative, the result is

$$\frac{h_j^{n+1} - h_j^n}{\Delta t} + u\frac{h_{j+1}^n - h_{j-1}^n}{\Delta x} = \nu\frac{\delta_x^2 h_j^n}{\Delta x^2}$$

or

$$h_j^{n+1} = h_j^n - \frac{Cr}{2}(h_{j+1}^n - h_{j-1}^n) + \frac{\nu\Delta t}{\Delta x^2}\delta_x^2 h_j^n$$

where $\delta_x^2 = h_{j+1}^n - 2h_j^n - h_{j-1}^n$ and $Cr = u\Delta t/\Delta x$ is known as the *cell Courant number.*

In the case $\nu = 0$ this is known as the *forward time centered space scheme.* When; $\nu\Delta t/\Delta x^2 = 1/2$ results in the *Lax-Friedrichs scheme*, $\nu\Delta t/\Delta x^2 = Cr^2/2$ results in the *Lax-Wendroff scheme* and when $\nu\Delta t/\Delta x^2 = Cr/2$ results in the *forward time backward space scheme* or *upwind scheme.*

Even-order derivatives in the truncation error account for dissipation and odd-order derivative account for dispersion.

Accuracy: The terms that are ignored in (57) to (59) are called truncation errors. They represent a measure of the accuracy of the finite difference approximation and determine the rate at which the *truncation error* decreases as the grid resolution is refined. This is known as the *convergence rate* of the finite difference approximation. It does not indicate the magnitude of the error.

The accuracy of any numerical approximation is impaired by *truncation* and*round-off errors*, the former being the most important. Truncation errors are a result of using a set of discrete points to approximate a continuous variable. The magnitude and the behaviour of this error in the approximating scheme, which is dependent on the computational mesh size, can be established qualitatively only. With modern computers, accumulation of round-off errors is usually not considered as a serious problem. This is a second-order approximation. The term $O(\Delta x^2)$ implies that the truncation error will go to zero faster as $\Delta x \to 0$ than an $O(\Delta x)$ scheme. This does not imply that a $O(\Delta x^2)$ is more accurate than a $O(\Delta x)$ scheme.

How closely h_j^n approximates $h(x_j, t^n)$ as Δt and Δx vary.

Convergence: A fundamental requirement of an approximating scheme is that its solution approximates the solution of the differential equation that it approximates. If this is satisfied the scheme is converges to the partial differential equation that they approximate. Convergence refers to the relationship between the solutions as Δx and Δt vanish.

Whether a finite difference solution approaches the exact differential solution as Δx and Δt vanish, i.e. $h_j^n \to h(x_j, t^n)$ as $\Delta t, \Delta x \to 0$.

Consistency: Another requirement of an approximating scheme is that the approximation is identical to the differential equation in the limit as the mesh size Δx and Δt vanish. Then the numerical scheme is consistent with the original differential equation and consistency refers to the relationship between equations. Consistency is necessary if the finite difference solution is to converge to the solution of the partial differential equation being approximated. However, this is not a sufficient condition, for even though the finite difference equation might become equivalent to a certain partial differential equation as the grid spacing tends to zero, it does not follow that the solution of the finite difference equation approaches the solution of the partial differential equation.

Whether a difference operator, $\mathcal{L}(h_j^n)$ can truly represent the differential operator, $\mathcal{L}(h(x,t))$ in the limiting solution as Δx and Δt vanish.

Stability: The application of a numerical scheme is determined by whether it is stable, that is, whether the numerical solution grows and becomes unbounded after repeatedly applying the scheme. i.e. whether h_j^n is bounded as $n \to \infty$.

A numerical scheme is stable if the accumulation error in the numerical scheme remains bounded as the solution evolves with time. There are a number of methods of stability analysis. These techniques include;

- *discrete perturbation analysis*

  An error is introduced into the finite difference scheme at an arbitrary grid point, (x_j, t^n) and its effect on the computation of subsequent values of the solution to the finite difference equation is examined. Stability is indicated if the perturbation decays as the solution evolves.

- *the matrix method*

  This is similar but more rigorous than the discrete perturbation analysis. Expressing the governing equations in matrix form, the error vector can be uniquely expressed in terms of the eigenvalues of the coefficient matrix. The matrix method can only be used if a set of linearly independent eigenvalues for the coefficient matric can be found. It is one of the very few stability analysis techniques where the effect of the boundary conditions are considered explicitly.

- *von Neumann analysis*

  In this method, the errors distributed along grid-points at one time level are expanded as a finite Fourier series. The stability or otherwise of the finite difference approximation is determined by considering whether separate Fourier components of the error distribution decay or amplify with time. This is the most common method of stability analysis and the easiest to apply. Another advantage of this method is that it not only provides information of the attenuation or amplification produced by the finite difference approximation, but also on any phase effects produced. However, like the matrix method, it can only be used to establish necessary and sufficient conditions for stability of linear initial value problems with constant coefficients. For nonlinear and variable coefficient problems, these methods can only be applied locally.

  We assume that time and space variables are separable and that the solution of the finite difference equation can be expanded in the form of a Fourier series.

- *heuristic stability analysis*[52]

The terms of the finite difference approximation are expanded in a Taylor series in order to develop a corresponding continuum partial differential equation. Stability of the difference approximation is then determined from the known behaviour of the resulting difference equation. Hirt[52] argued that the effect of

The present procedure for determining the stability, based on a differential equation constructed from the finite difference equation through Taylor series expansion, is called *Hirt's stability analysis*. It is based on the observation of the behaviour of equations with higher-order terms. For example, a diffusion term will smear out a disturbances in the solution. A negative diffusion coefficient is physically impossible, or otherwise the opposite would happen.

Hirt's stability analysis is based on the fact that a diffusion coefficient cannot be negative.

Unfortunately, none of these stability analysis techniques can be considered as robust, so when actually computing solutions to a finite difference equation they should be supplemented by numerical experimentation. Despite the most common use of the von Neumann's stability analysis, each method has its own advantages as well as limitations. A comparison and evaluation of these three methods can be found in Roach[100].

Lax's Equivalence Theorem: There exist important connections between the stability of linear finite difference schemes and the convergence of their solutions to that of the linear partial differential equations to which they are consistent. An important example is *Lax's equivalence theorem*.

This theorem, provided by Lax and described in Lax and Richtmyer[71] is; "*given a properly posed linear initial value problem and a finite difference approximation to it that satisfies the consistency condition, stability is the necessary and sufficient condition for convergence*."

This theorem is important because, whilst it is relatively easy to the stability and consistency of a finite difference approximation, it is generally very difficult to show convergence of its solution to that of the partial differential equation. Although Lax's equivalence theorem was restricted to linear equations, Ansorge[4] has since extended the theorem to a wider range of applications.

Increasing the accuracy of the solution of a numerical scheme is usually accomplished by an increase in complexity of a numerical scheme. A complex numerical scheme may consume significant amount of computer resources for a small gain in accuracy. A compromise must be reached in order to obtain a numerical model which is both accurate and efficient. Therefore, *efficiency* of a numerical scheme is as important as the above parameters in the selection of a suitable numerical scheme.

To demonstrate the above properties consider a number of finite difference approximations to the linear advection equation given by (40). Equation (40) may be replaced by a three-point, two-time level finite difference scheme method with appropriate boundary conditions in the region $0 \leq x \leq L$, $0 \leq t \leq T$. The general form of many of these finite difference schemes can be written

for $u > 0$ in the form[88]

$$\gamma\frac{h_{j-1}^{n+1}-h_{j-1}^{n}}{\Delta t}+(1-2\gamma)\frac{h_{j}^{n+1}-h_{j}^{n}}{\Delta t}+\gamma\frac{h_{j+1}^{n+1}-h_{j+1}^{n}}{\Delta t}$$
$$+u\left\{\theta\left[\varphi\frac{h_{j}^{n+1}-h_{j-1}^{n+1}}{\Delta x}+(1-\varphi)\frac{h_{j+1}^{n+1}-h_{j-1}^{n}}{2\Delta x}\right]\right.$$
$$\left.+(1-\theta)\left[\varphi\frac{h_{j}^{n}-h_{j-1}^{n}}{\Delta x}+(1-\varphi)\frac{h_{j+1}^{n}-h_{j-1}^{n}}{2\Delta x}\right]\right\}=0 \tag{63}$$

for $u > 0$. Different finite difference approximations are obtained depending on the choice of the weighting coefficients γ, φ and θ. Equation (63) can be simplified into

$$\alpha_{-1}h_{j-1}^{n+1}+\alpha_0 h_j^{n+1}+\alpha_1 h_{j+1}^{n+1}=\beta_{-1}h_{j-1}^{n}+\beta_0 h_j^{n}+\beta_1 h_{j+1}^{n} \tag{64}$$

in which

$$\begin{aligned}\alpha_{-1}&=2\gamma-\theta Cr(1+\varphi),\\ \alpha_0&=2(1-2\gamma+\theta\varphi Cr),\\ \alpha_1&=2\gamma+\theta Cr(1-\varphi),\\ \beta_{-1}&=2\gamma+Cr(1-\theta)(1+\varphi),\\ \beta_0&=2(1-2\gamma-\varphi Cr(1-\theta))\end{aligned}$$

and

$$\beta_1=2\gamma-Cr(1-\theta)(1-\varphi).$$

When $\alpha_{-1}=\alpha_1=0$ results in an *explicit scheme*, which can be expressed as

$$h_j^{n+1}=\sum_{q=-1}^{1}\beta_q h_{j+q}^{n} \tag{66}$$

otherwise it is an *implicit scheme*.

Finite difference approximations by which values at any time level are computed in terms of known values at previous time levels are known as implicit schemes.

The modified equivalent partial differential equation for (63) is given by (41) with the following coefficients for the first three higher-order derivatives[90]

$$\eta_2(Cr)=Cr(1-2\theta)-\varphi, \tag{67a}$$
$$\eta_3(Cr)=1-6\gamma-3\varphi Cr(1-2\theta)+2Cr^2(1-3\theta+3\theta^2) \tag{67b}$$

and

$$\eta_4(Cr)=\varphi(12\gamma-1)+Cr(1-2\theta)(4+3\varphi^2-24\gamma)$$
$$-12\varphi Cr^2(1-3\theta+2\theta^2)+6Cr^2(1-4\theta+6\theta^2-4\theta^3). \tag{67c}$$

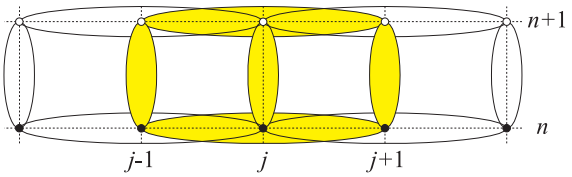


Figure 38: Computational stencil for the implicit Crank-Nicholson finite difference scheme; • known data ◦ unknown data.

If $\eta_q(Cr) = 0$ for $q = 2, 3, \ldots, Q$ the finite difference approximation is said to be of order Q, since the *local truncation error* is

$$E = \sum_{q=Q+1}^{\infty} \frac{u\Delta x^{q-1}}{q!}\eta_q(Cr)\frac{\partial^q h}{\partial x^q}$$

is $O(\Delta x^Q)$. This is the order of accuracy of a finite difference approximation if there are no discontinuities in the initial or boundary conditions. It indicates how closely the solution of the differential equation satisfies the approximating finite difference equation, that is, it shows consistency if it exists, it also indicates how rapidly the solution of the finite difference equation approaches the solution of the given differential equation as $\Delta x \to 0$, $\Delta t \to 0$ if the solution is convergent

Consider the forward time centered space scheme, which can be written for the advection equation as

$$h_j^{n+1} = \frac{Cr}{2}h_{j-1}^n + h_j^n - \frac{Cr}{2}h_{j+1}^n \tag{68}$$

with the modified equivalent partial differential equation

$$\frac{\partial h}{\partial t} + u\frac{\partial h}{\partial x} + \frac{u\Delta x}{2}Cr\frac{\partial^2 h}{\partial x^2} + O(\Delta x^2).$$

The scheme is consistent because its modified equivalent partial differential equation is identical to the differential equation as $\Delta x, \Delta t \to 0$. It is first-order accurate because the truncation error goes to zero like the first power in Δx. However it is unconditionally unstable because $\eta_2(Cr) = Cr$ and $Cr > 0$. The scheme introduces anti-diffusion, the solution grows with time.

The Lax-Friedrichs scheme, given by

$$h_j^{n+1} = \frac{(1+Cr)}{2}h_{j-1}^n + \frac{(1-Cr)}{2}h_{j+1}^n \tag{69}$$

can be obtained from (64) with $\alpha_{-1} = 0$, $\alpha_0 = 0$, $\alpha_1 = 1$, $\beta_{-1} = (1+Cr)/2$, $\beta_0 = 0$, and $\beta_1 = (1-Cr)/2$ or substituting $\varphi = 1/Cr$ and $\theta = \gamma = 0$ into (63). Substituting these values in (67), then

$$\eta_2(Cr) = \frac{Cr^2 - 1}{Cr},$$
$$\eta_3(Cr) = -2 + 2Cr^2$$

and

$$\eta_4(Cr) = \frac{2}{Cr} - 8Cr + 6Cr^2.$$

Since the non-vanishing term occurs for $Q = 1$, then this is also a first-order approximation and it is consistent with the advection equation. The modified equivalent partial differential equation for the Lax-Friedrichs scheme is

$$\frac{\partial h}{\partial t} + u\frac{\partial h}{\partial x} + \frac{u\Delta x}{2}\left(\frac{Cr^2 - 1}{Cr}\right)\frac{\partial^2 h}{\partial x^2} + O(\Delta x^2) = 0$$

which indicates that either artificial diffusion or anti-diffusion is introduced depending on the magnitude of Cr. The numerical scheme is unstable when $|Cr| > 1$ because it introduces anti-diffusion which will accentuate any irregularities that occur including round off errors. The scheme introduces artificial diffusion which will tend to smooth the solution and is therefore stable for values of $Cr \leq 1$.

Consider the forward time upwind scheme

$$h_j^{n+1} = Crh_{j-1}^n + (1 - Cr)h_j^n \tag{71}$$

obtained using $\varphi = 1$, $\theta = \gamma = 0$, then $\eta_2(Cr) = Cr - 1$ and the modified partial differential equation is

$$\frac{\partial h}{\partial t} + u\frac{\partial h}{\partial x} + \frac{u\Delta x}{2}(Cr - 1)\frac{\partial^2 h}{\partial x^2} + O(\Delta x^2) = 0.$$

The first-order upwind scheme introduces artificial diffusion if $0 \leq Cr\ leq1$, otherwise artificial anti-diffusion is introduced and the scheme becomes unstable. Comparing the truncation error between the Lax-Friedrichs and forward time upwind scheme reveals that more diffusion is introduced by the Lax-Friedrichs scheme compared with the forward time upwind scheme.

The formal order of accuracy of a numerical approximation generally increases with the number of grid points involved. Higher-order schemes can be developed by introducing more points in the finite difference stencil. Consider, the second-order Lax-Wendroff scheme, where $\varphi = Cr$, $\theta = \gamma = 0$ in (63), given by

$$h_j^{n+1} = \frac{Cr(1 + Cr)}{2}h_{j-1}^n + (1 - Cr^2)h_j^n - \frac{Cr(1 - Cr)}{2}h_{j+1}^n \tag{72}$$

with $\eta_2(Cr) = 0$, $\eta_3(Cr) = 1 - Cr^2$ and the modified equivalent partial differential equation is

$$\frac{\partial h}{\partial t} + u\frac{\partial h}{\partial x} + \frac{u\Delta x^2}{6}\left(1 - Cr^2\right)\frac{\partial^3 h}{\partial x^3} + O(\Delta x^3) = 0.$$

The Lax-Wendroff scheme is indeed a second-order scheme, consistent with the advection equation and does not introduce artificial diffusion. Instead, this scheme introduces artificial dispersion. For

a stable solution, $-1 \leq Cr \leq 1$, the dispersion coefficient is always positive. This is in contrast to the explicit Beam-Warming scheme given by

$$h_j^{n+1} = h_j^n - \frac{Cr}{2}(3h_j^n - 4h_{j-1}^n + h_{j-2}^n) + \frac{Cr^2}{2}(h_j^n - 2h_{j-1}^n + h_{j-2}^n)$$

which has the modified equivalent partial differential equation given by

$$\frac{\partial h}{\partial t} + u\frac{\partial h}{\partial x} - \frac{u\Delta x^2}{6}\left(2 - 3Cr + Cr^2\right)\frac{\partial^3 h}{\partial x^3} + O(\Delta x^3) = 0.$$

In this case the dispersion coefficient is negative in the stable region $0 \leq Cr \leq 2$.

Equations (69), (71) and (79) can be written in the form

$$h_j^{n+1} = h_j^n - \frac{Cr}{2}(h_{j+1}^n - h_{j-1}^n) + \frac{q}{2}(h_{j+1}^n - 2h_j^n + h_{j-1}^n) \tag{73}$$

where, with $q = 1$ produces the Lax-Friedrichs scheme, when $q = Cr$ and $Cr > 0$ produces the forward time upwind scheme and when $q = Cr^2$ results in the Lax-Wendroff scheme. From (73) the above schemes can be interpreted as solving

$$\frac{\partial h}{\partial t} + u\frac{\partial h}{\partial x} = \frac{q\Delta x^2}{2\Delta x}\frac{\partial^2 h}{\partial x^2}$$

which is the parabolic advective-diffusion equation. The original equation being modelled, the advection equation is hyperbolic. The extra term involving q is effectively artificial numerical viscosity introduced by the finite difference approximation. This artificial viscosity stabilizes the numerical scheme by damping any oscillations in the numerical scheme, even in the presence of rapid transitions in the solution.

It is relatively easy to show that the domain of dependence for the finite difference equation must not lie inside that for the differential equation. For if it does, it follows that the solution of the difference equation at a grid point is independent of the initial data of the problem which lies outside the domain of dependence of the difference equation but inside the domain of dependence of the difference equation. Alteration of this initial data will modify the solution of the differential equation, but leave the solution of the difference equation unaltered. The Courant-Friedrichs-Lewy condition states that "*the domain of dependence of the difference equation must include the domain of dependence of the differential equation.*" [22]

Characteristic curves play an important role in the numerical solution of hyperbolic equations. The reason for this is that the solution (x_j, t^{n+1}) depends only on the initial data on that segment of the x-axis cut out by the characteristics which meet at (x_j, t^{n+1}). The region of influence of the computational points $j-1$ to j *must* enclose the domain of dependence of the the point (x_j, t^{n+1}) otherwise the numerical scheme will calculate the solution at (x_j, t^{n+1}) using information that does not arrive and therefore not relevant to the information at (x_j, t^{n+1}) during the time interval Δt. In this case the solution is obtained using *extrapolation*, with perhaps unexpected results. If the

domain of dependence of the numerical method includes the domain of dependence of the differential equation, then the numerical scheme *interpolates* the solution from available data. Obviously, if the domain of dependence of the true and numerical solution coincide , the exact solution is obtained. Otherwise, the accuracy of the numerical solution will depend on how close the numerical domain of dependence approximates the true domain of dependence.

This condition is generally known as the *Courant-Friedrichs-Lewy condition* or simply the *CFL condition* [21]. The numerical domain of dependence contains the continuous domain of dependence.

In other words, the ratio of the maximum speed, $\max(u \pm \sqrt{gh})$ to the grid speed $\Delta x/\Delta t$ should be bounded by unity. Mathematically this is expressed as

$$\max(u \pm \sqrt{gh})\frac{\Delta x}{\Delta t} \leq 1$$

and is illustrated in Figure 40.

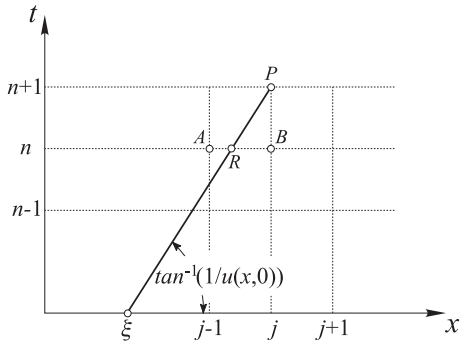


Figure 39: Domain of dependence for a single hyperbolic system; • known data ◦ unknown data.

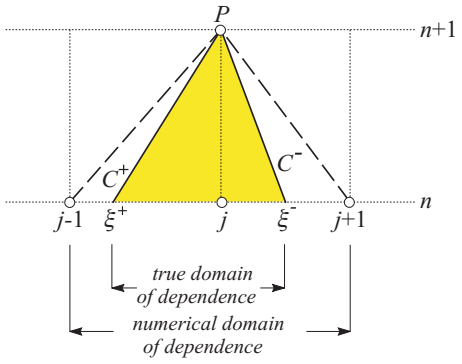


Figure 40: Characteristics and the domain of dependence for the true solution and the propagation of information in explicit numerical schemes approximating a 2×2 hyperbolic system of equations.

The Courant-Friedrichs-Lewy condition applies to explicit finite difference approximations to hyperbolic equations. Consider the characteristic for a linear first-order hyperbolic equation, that

passes through a computational grid shown in Figure 39. At the computational points located at (x_j, t^n), (x_{j-1}, t^n) information is known. The solution is required at (x_j, t^{n+1}). The Courant-Friedrichs-Lewy condition requires that the characteristic which passes through the point P, in Figure 39 intersects the line $t = n\Delta t$ between the points A and B. If the Courant-Friedrichs-Lewy condition is not satisfied, convergence cannot take place when $\Delta x \to 0$ with $\Delta t = k\Delta x$, where k is a constant. Since $\tan(\theta) = 1/u(h) = \Delta t/RB$, then $RB = \Delta t u(h)$. The Courant-Friedrichs-Lewy condition requires that $0 \leq RB \leq \Delta x$, or $0 \leq \Delta t u(h)$, or $0 \leq ku(h) \leq 1$ ($\Delta x \neq 0$). The area enclosed by these curves have been called the domain of dependence of the numerical scheme. The area enclosed for a 2×2 system by the characteristic curves is called the domain of dependence of the true solution.

A great deal is known about the behaviour of conservative laws and their solution. This knowledge can be exploited to develop special methods that overcome some of the numerical difficulties encountered with naive finite differences.

A finite difference scheme is defined to be *monotone* if the solution at all time levels is monotonically increasing or decreasing whenever the initial data is monotonically increasing or decreasing in space. Consider the difference between explicit finite difference schemes, given by (66) at two neighbouring grid points

$$\begin{aligned} h_{j+1}^{n+1} - h_j^{n+1} &= \sum_{q=-1}^{1} \beta_q h_{j+q+1}^n - \sum_{q=-1}^{1} \beta_q h_{j+q}^n \\ &= \sum_{q=-1}^{1} \beta_q (h_{j+q+1}^n - h_{j+q}^n) \end{aligned}$$

If the solution at t^n is monotonic then for all j's, $h_{j+q+1}^n - h_{j+q}^n$ will have the same sign. In addition, if $\beta_q \geq 0$, then $h_{j+1}^{n+1} - h_j^{n+1}$ will also have the same sign as $h_{j+q+1}^n - h_{j+q}^n$ and the solution remains monotonic and the scheme is monotone. If however, one of the coefficients in (66) $\beta_{q=k} < 0$, then for a monotone solution at t^n shown in Figure 41

$$h_{k+1}^{n+1} - h_k^{n+1} = \beta_k (h_{k+1}^n - h_k^n) = \beta_k < 0$$

since $h_{j+1}^n - h_j^n = 0 \ \forall j$ except when $j = k$ where $h_{j+1}^n - h_j^n = 1$. The solution oscillates between a monotone increasing and decreasing solution. Therefore, the requirement for an explicit finite difference schemes given by (73) is that all β_q's are non-negative. Within their respective regions of stability, examples of monotone finite difference schemes are the Lax-Friedrichs and forward time centered space schemes. This is not the case for the Lax-Wendroff scheme.

Godunov schemes are named after the Russian mathematician Sergei Konstantinovich Godunov (1929-). Godunov's scheme piecewise constant over a mesh at each time step and the time evolution is determined by the exact solution of the Riemann problem at the cell interface.

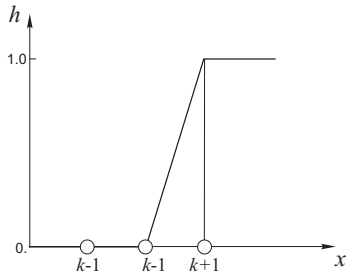


Figure 41:

Consider (66), using the Taylor series in time, then

$$h_j^{n+1} = h_j^n + \Delta t \left.\frac{\partial h}{\partial t}\right|_j^n + \frac{\Delta t^2}{2} \left.\frac{\partial^2 h}{\partial t^2}\right|_j^n + O(\Delta t^3).$$

From the original partial differential equation $\partial h/\partial t = -u\partial h/\partial x$ and $\partial^2 h/\partial t^2 = u^2 \partial^2 h/\partial x^2$ then

$$h_j^{n+1} = h_j^n - u\Delta t \left.\frac{\partial h}{\partial x}\right|_j^n + \frac{u^2 \Delta t^2}{2} \left.\frac{\partial^2 h}{\partial x^2}\right|_j^n + O(\Delta x^3). \tag{74}$$

The Taylor series expansion of the right hand side terms in (66) yields

$$\sum_{q=-1}^{1} \beta_q h_{j+q}^n = (\beta_{-1} + \beta_0 + \beta_1) h_j^n + (-\beta_{-1} + \beta_1)\Delta t \left.\frac{\partial h}{\partial x}\right|_j^n + (\beta_{-1} + \beta_1)\frac{\Delta x^2}{2} \left.\frac{\partial^2 h}{\partial x^2}\right|_j^n + O(\Delta x^3). \tag{75}$$

To obtain second-order accuracy in both space and time, the coefficients in (74) and (75) must cancel. This will occur when

$$\sum_{q=-1}^{1} \beta_q = 1, \tag{76a}$$

$$Cr = -\sum_{q=-1}^{1} q\beta_q \tag{76b}$$

and

$$Cr^2 = \sum_{q=-1}^{1} q^2 \beta_q. \tag{76c}$$

For a monotone scheme, $\beta_q \geq 0$, so the following variables can be defined $a_q = \sqrt{\beta_q}$, $b_q = q\sqrt{\beta_q}$

and (76) becomes

$$\sum_{q=-1}^{1} (a_q)^2 = 1,$$

$$Cr^2 = \left(\sum_{q=-1}^{1} a_q b_q \right)^2$$

and

$$Cr^2 = \sum_{q=-1}^{1} (b_q)^2.$$

Manipulating these expressions, then

$$\sum_{q=-1}^{1} (a_q)^2 \sum_{q=-1}^{1} (b_q)^2 = \left(\sum_{q=-1}^{1} a_q b_q \right)^2 .$$

Using *Schwartz inequality*[7], then

$$\sum_{q=-1}^{1} (a_q)^2 \sum_{q=-1}^{1} (b_q)^2 \geq \left(\sum_{q=-1}^{1} a_q b_q \right)^2$$

and the equality holds only if

$$a_q \propto b_q.$$

According to the definitions for a_q and b_q the coefficients of proportionality must be

There are three equations with three unknowns

$$\begin{bmatrix} 1 & 1 & 1 \\ 1 & 0 & -1 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_{-1} \\ \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} 1 \\ Cr \\ Cr^2 \end{bmatrix}$$

Solving, yields $\beta_{-1} = 1$, $\beta_0 = Cr - 1$ and $\beta_1 = Cr^2 - Cr$.

For stability $Cr \leq 1$, therefore $\beta_{-1} = 1$, $\beta_0 \leq 0$ and $\beta_1 \leq 0$. The scheme is monotone only when $Cr = 1$ were $h_j^{n+1} = h_{j-1}^n$ which is the exact solution. For all other values of Cr, β_0 and β_1 are negative violating the condition for a monotonic scheme. Therefore, second and higher-order schemes will not be monotone schemes.

Consider conditionally stable first-order explicit schemes. These will involve only two spatial computational points and therefore a combination of; β_{-1}, β_0 or β_{-1}, β_1. These schemes, which

[7] Schwartz inequality:

$$L_2(a)\ L_2(b) \geq L_2(ab).$$

correspond to the forward time upwind scheme and Lax-Friedrichs produce $\beta_{-1} = Cr$, $\beta_0 = 1 - Cr$ and $\beta_{-1} = 1 - Cr$, $\beta_0 = Cr$ and are monotone preserving schemes. Therefore, monotone schemes can not be higher than first-order accurate.

Monotonicity also plays an important role for conservative finite difference schemes. The most important is the theorem attributed to Law and Wendroff[72], which states that "*the solution of a conservative* monotone *finite difference scheme of a conservation law will converge to the weak solution to that law as* Δt *and* $\Delta x \to 0$*, and the entropy condition is satisfied across all discontinuities of the dependent variables*".

Some numerical approximations produce oscillations in the vicinity of steep gradients. If a requirement of the numerical scheme is that it is monotonicity preserving, then the oscillations are eliminated.

The numerical scheme for a conservative law in the form

$$\begin{aligned} h_j^{n+1} &= h_j^n - \frac{\Delta t}{\Delta x}\left[F(h_{j-p}^n, \ldots, h_{j+q+1}^n) - F(h_{j-p-1}^n, \ldots, h_{j+q}^n)\right] \\ &= R(h_{j-p-1}, \ldots, h_{j+p+1}) \end{aligned}$$

in which the numerical flux F is an approximation of the intercell flux $F_{j\pm 1/2}$ and depends on a finite number of values $h_{j-p-1}, \ldots, h_{j+p+1}$, is said to be monotone if the right hand side is monotone function for each argument $h_{j-p-1}, \ldots, h_{j+p+1}$. For example, consider a finite difference scheme where $p = 0$ and $q = 0$ then

$$h_j^{n+1} = R(h_{j-1}^n, h_j^n, h_{j+1}^n) \quad \text{and} \quad h_{j+1}^{n+1} = R(h_j^n, h_{j+1}^n, h_{j+2}^n).$$

Subtracting these values, then

$$h_{j+1}^{n+1} - h_j^{n+1} = R(h_j^n, h_{j+1}^n, h_{j+2}^n) - R(h_{j-1}^n, h_j^n, h_{j+1}^n).$$

Making use of the mean value theorem, then there exists a constant γ, $0 < \gamma < 1$ such that

$$\begin{aligned} h_{j+1}^{n+1} - h_j^{n+1} &= R(h_j^n, h_{j+1}^n, h_{j+2}^n) - R(h_{j-1}^n, h_j^n, h_{j+1}^n) \\ &= \sum_{i=n-1}^{n+1} \frac{\partial R(\overline{h}_{j-1}^n, \overline{h}_j^n, \overline{h}_{j+1}^n)}{\partial h_j^n}(h_{j+1}^n - h_j^n) \end{aligned}$$

where

$$(\overline{h}_{j-1}^n, \overline{h}_j^n, \overline{h}_{j+1}^n) = (\gamma \overline{h}_j^n + (1-\gamma), \gamma \overline{h}_{j+1}^n + (1-\gamma)\overline{h}_j^n, \gamma \overline{h}_{j+2}^n + (1-\gamma)\overline{h}_{j+1}^n).$$

Since h_j^n is monotone increasing or decreasing, then $(h_j^n - h_{j-1}^n)$, $(h_{j+1}^n - h_j^n)$ and $(h_{j+2}^n - h_{j+1}^n)$ are all either positive or negative. For a monotone preserving scheme, then $(h_{j+1}^{n+1} - h_j^{n+1})$ must have the same sign as $(h_j^n - h_{j-1}^n)$. This can only occur if

$$\frac{\partial R}{\partial h_i^n} \geq 0 \quad \text{for} \quad i = j - p - 1, \ldots, j + q + 1.$$

This is a special condition that ensures that a monotone solution remains monotone and represents a severe constraint on a numerical scheme.

These results would be identical to the results obtained using other linear stability analysis, such as the von Neuman stability analysis. Unfortunately, linear stability analysis does not imply convergence if discontinuities are present[107].

6.1 Lax-Friedrichs

A well known first-order scheme is the Lax-Friedrichs scheme where the flux (Lax [69]) consists of a diffusive term approximating $\Delta x^2/2/\Delta t \mathbf{U}_x$ added to the unstable centered flux term.

$$\mathbf{U}_j^{n+1} = \frac{1}{2}\left(\mathbf{U}_{j+1}^n + \mathbf{U}_{j-1}^n\right) - \frac{\Delta t}{2\Delta x}\left(\mathbf{F}_{j+1}^n - \mathbf{F}_{j-1}^n\right).$$

In terms of fluxes

$$\mathbf{f}_{j+1/2}^n = \frac{1}{2}(\mathbf{F}_{j+1}^n + \mathbf{F}_j^n) - \frac{\Delta x}{2\Delta t}(\mathbf{U}_{j+1}^n - \mathbf{U}_j^n) \tag{78}$$

6.2 Richtmyer Scheme

The Richtmyer scheme is a two-step scheme (see, for example Richtmyer and Morton[97], p. 303). For the first step

$$\bar{\mathbf{U}}_{j+1/2}^{n+1} = \frac{1}{2}\left(\mathbf{U}_{j+1}^n + \mathbf{U}_j^n\right) - \frac{\Delta t}{2\Delta x}\left(\mathbf{F}_{j+1}^n - \mathbf{F}_j^n\right)$$

and for the second step;

$$\mathbf{U}_j^{n+1} = \mathbf{U}_j^n - \frac{\Delta t}{\Delta x}\left(\bar{\mathbf{F}}_{j+1/2}^{n+1} - \bar{\mathbf{F}}_{j-1/2}^{n+1}\right)$$

These results are similar to those produced by the Lax-Wendroff scheme. The exception is that this scheme was able to provide some results for the second example. There is excellent resolution of the shock and rarefaction wave. However, the remainder of the solution is corrupted by oscillations introduced by the dispersive behaviour of this second-order scheme. However, unlike first-order schemes, it seems that the solution satisfies the entropy condition.

6.3 Lax-Wendroff Scheme

If downwind differences are used to approximate $\mathbf{U}_x$ this will result in the second-order Lax-Wendroff scheme[72]) written as

$$\begin{aligned}\mathbf{U}_j^{n+1} = \mathbf{U}_j^n &- \frac{\Delta t}{2\Delta x}\left(\mathbf{F}_{j+1}^n - \mathbf{F}_{j-1}^n\right) \\ &+ \frac{\Delta t^2}{2\Delta x^2}\left(\mathbf{A}_{j+1/2}^n\left(\mathbf{F}_{j+1}^n - \mathbf{F}_j^n\right) - \mathbf{A}_{j-1/2}^n\left(\mathbf{F}_j^n - \mathbf{F}_{j-1}^n\right)\right)\end{aligned}$$

in which $\mathbf{A}_{j+1/2}^n = \mathbf{A}\left(\mathbf{U}_{j+1/2}^n\right) = \mathbf{A}\left(\left(\mathbf{U}_j^n + \mathbf{U}_{j+1}^n\right)/2\right)$ or $\mathbf{A}_{j+1/2}^n = \left(\mathbf{A}_j^n + \mathbf{A}_{j+1}^n\right)/2$. In terms of fluxes given in (103), then

$$\mathbf{f}_{j+1/2}^{n+1} = \frac{1}{2}\left(\mathbf{F}_{j+1}^n + \mathbf{F}_{j-1}^n\right) - \frac{\Delta t}{2\Delta x}\mathbf{A}_{j+1/2}^n\left(\mathbf{F}_{j+1}^n - \mathbf{F}_j^n\right) \tag{79}$$

in which $\mathbf{A}^n_{j+1/2} = (\mathbf{A}^n_j + \mathbf{A}^n_{j+1})/2$.

Table 1: Fluxes for various numerical scheme.

Scheme	**Numerical flux**	**Equation reference**
Unstable	$\mathbf{f}^n_{j+1/2} = \frac{1}{2}(\mathbf{F}^n_{j+1} + \mathbf{F}^n_j)$	
Lax-Friedrichs	$\mathbf{f}^n_{j+1/2} = \frac{1}{2}(\mathbf{F}^n_{j+1} + \mathbf{F}^n_j) - \frac{\Delta x}{2\Delta t}(\mathbf{U}^n_{j+1} - \mathbf{U}^n_j)$	(78)
Lax-Wendroff	$\mathbf{f}^n_{j+1/2} = \frac{1}{2}\left(\mathbf{F}^n_{j+1} + \mathbf{F}^n_j\right) - \frac{\Delta t}{2\Delta x}\mathbf{A}^n_{j+1/2}\left(\mathbf{F}^n_{j+1} - \mathbf{F}^n_j\right)$	(79)
Beam-Warming		
MacCormack	$\mathbf{f}^n_{j+1/2} = \frac{1}{2}(\mathbf{F}^n_{j+1} + \overline{\mathbf{F}}_j)$	(80)
Bermudez-Vazquez	$\mathbf{f}^n_{j+1/2} = \frac{1}{2}\left(\mathbf{F}^n_{j+1} + \mathbf{F}^n_j\right) + \frac{1}{2}\left(\lvert\mathbf{G}\rvert^n_j\mathbf{U}^n_j - \lvert\mathbf{G}\rvert^n_{j+1}\mathbf{U}^n_{j+1}\right)$	(104)
Steger-Warming	$\mathbf{f}^n_{j+1/2} = \frac{1}{2}\left(\mathbf{F}^n_{j+1} + \mathbf{F}^n_j\right) + \frac{1}{2}\left(\lvert\mathbf{A}\rvert^n_j\mathbf{U}^n_j - \lvert\mathbf{A}\rvert^n_{j+1}\mathbf{U}^n_{j+1}\right)$	(108)
Yang	$\mathbf{f}^n_{j+1/2} = (\mathbf{A})^-_{j+1/2}(\mathbf{F}^n_{j+1} - \mathbf{F}^n_j)$	(109)
TVD Lax-Wendroff	$\mathbf{f}^n_{j+1/2} = \mathbf{f}^{(UP)n}_{j+1/2} + \phi(\beta^+, \beta^-)\left(\mathbf{f}^{(LW)n}_{j+1/2} - \mathbf{f}^{(UP)n}_{j+1/2}\right)$	
Harten	$\mathbf{f}^n_{j-1/2} = \mathbf{F}^n_{j-1} + \left[1 - \delta^n_{j-1/2}\right]\left(\mathbf{F}^n_j - \mathbf{F}^n_{j-1}\right)$	(110)
Biased average procedure	$\mathbf{f}^n_{j+1/2} = \mathbf{F}^+_j(x)\lvert^n_{x_{j+1/2}} + \mathbf{F}^-_{j+1}(x)\lvert^n_{x_{j+1/2}}$	(**??**)

Table 2: Fluxes for various numerical scheme, cont.

Scheme	**Numerical flux**	**Equation reference**
Roe's	$\mathbf{f}^n_{j+1/2} = \mathbf{F}(\mathbf{U}^n_l) + \sum_{\overline{\lambda}_i \leq 0} \overline{\lambda}_i \overline{\alpha}_i \overline{\mathbf{e}}_i$ $\mathbf{f}^n_{j+1/2} = \mathbf{F}(\mathbf{U}^n_r) - \sum_{\overline{\lambda}_i > 0} \overline{\lambda}_i \overline{\alpha}_i \overline{\mathbf{e}}_i$ $\mathbf{f}^n_{j+1/2} = \frac{1}{2}\left(\mathbf{F}(\mathbf{U}^n_r) + \mathbf{F}(\mathbf{U}^n_l)\right) - \frac{1}{2}\sum_{i=1}^{2} \lvert\overline{\lambda}_i\rvert \overline{\alpha}_i \overline{\mathbf{e}}_i$	(144)
Harten, Lax and van Leer	$\mathbf{f}^n_{j+1/2} = \begin{cases} \mathbf{F}(\mathbf{U}^n_l) & \text{if} \quad S_1 \geq 0 \\ \mathbf{F}(\mathbf{U}^n_m) & \text{if} \quad S_1 < 0 < S_2 \\ \mathbf{F}(\mathbf{U}^n_r) & \text{if} \quad S_2 \leq 0. \end{cases}$	(157)
Weighted average flux	$\mathbf{f}^n_{j+1/2} = \dfrac{\mathbf{F}^n_j + \mathbf{F}^n_{j+1}}{2} - \dfrac{1}{2}\sum_{k=1}^{3} \operatorname{sgn}(S^{(k)}_{j+1/2}) \psi^{(k)}_{j+1/2} \Delta\mathbf{f}^{(k)}_{j+1/2}$	(164)
Essentially non-oscillatory $O(\Delta x^2)$[91]	$\mathbf{f}_{j+1/2} = \frac{1}{2}(\mathbf{F}_l + \mathbf{F}_r) - \frac{\alpha}{2}(\mathbf{U}_r - \mathbf{U}_l)$	(??)
Essentially non-oscillatory $O(\Delta x^3)$third-order[122]	$\mathbf{f}^n_{j+1/2} = \dfrac{1}{2}(\mathbf{F}^n_j + \mathbf{F}^n_{j+1} + \mathbf{R}^n_{j+1/2}\psi^n_{j+1/2})$	(??)
Central [78]	$\mathbf{f}^n_{j+1/2} = \mathbf{f}(\mathbf{U}^{n+1/2}_j) + \dfrac{\Delta x}{8\Delta t}\mathbf{U}'_j$	
Central upwind[66]		
Kinetic		

6.4 Beam-Warming Scheme

Upwind differences result in the Beam-Warming scheme (Warming and Beam [127]).

6.5 MacCormack Scheme

The MacCormack scheme [79] is also a two step predictor-corrector scheme given by

$$\bar{\mathbf{U}}_j^{n+1} = \mathbf{U}_j^n - \frac{\Delta t}{\Delta x}\left(\mathbf{F}_{j+1}^n - \mathbf{F}_j^n\right)$$

and for the second step;

$$\mathbf{U}_j^{n+1} = \frac{1}{2}\left(\mathbf{U}_j^n + \bar{\mathbf{U}}_j^{n+1}\right) - \frac{\Delta t}{2\Delta x}\left(\bar{\mathbf{F}}_j^{n+1} - \bar{\mathbf{F}}_{j-1}^{n+1}\right).$$

In terms of the intercell flux

$$\mathbf{f}_{j+1/2}^n = \frac{1}{2}(\mathbf{F}_{j+1}^n + \overline{\mathbf{F}}_j) \tag{80}$$

For linear problems it is identical to the Lax-Wendroff scheme.

This schemes behaviour is similar to the Lax-Wendroff scheme. Oscillations occur in the vicinity of abrupt changes in the dependent variables. In addition, this scheme also failed to produce a solution for the second-example. These oscillation can be reduced by introducing artificial viscosity or by using a flux limiter, such as those described in the following section.

These results are similar to those produced by the Lax-Wendroff scheme. Oscillations occur in the vicinity of abrupt changes in the dependent variables. In addition, this scheme also failed to produce a solution to the second example. To obtain a higher resolution of shocks and rarefaction fans, higher-order finite difference schemes can be employed.

In general, methods based on naive finite difference approximations may behave well for smooth solutions but can give disastrous results when discontinuities are present. Naive discretizations of the governing equations will generally result in either very smeared solutions or oscillations adjacent to discontinuities. This is because finite differences produce poor approximations to discontinuous functions.

7 Artificial Viscosity and Limited Schemes

The suppression of spurious oscillations near discontinuities in the solution profile by employing a monotone numerical scheme will only result at best in a first-order accurate solution. Therefore, there is no advantage in using higher-order schemes, which will not posses the montonicity property, and will generally result in oscillations in the solution in comparison with first-order schemes. This predicament can be resolved by recognizing that the conditions for a monotone scheme is very restrictive than is required to guarantee that these oscillations are supressed. This is achieved by incorporating some form of *slope limiter* or suitable *bounding*. The purpose of limiting or bounding a numerical scheme is to produce a high resolution scheme without the spurious oscillations associated with the more classical second-order schemes. This is achieved by developing a numerical scheme where the *total variation (TV)* defined by

$$TV(h^{n+1}) = \sum_{j=-\infty} |h_{j+1}^{n+1} - h_j^{n+1}|$$

is not increasing with time. The class of finite difference schemes for conservative laws, which possess the property

$$TV(h^{n+1}) < TV(h^n)$$

are known as *total variation diminishing (TVD)*. These properties are sufficient to obtain some convergence results and also guarantee that monotone profiles remain monotone, thereby preventing the formation of spurious oscillations in the solution. If an additional entropy condition is satisfied, the scheme is convergent and stable[47]. For example, if a numerical scheme has the following form

$$h_j^{n+1} = h_j^n - \alpha_{j-1}(h_j^n - h_{j-1}^n) + \beta_j(h_{j+1}^n - h_j^n)$$

then the scheme is TVD if

$$\alpha_j \geq 0, \quad \beta_j \geq 0 \quad \text{and} \quad \alpha_{j-1} + \beta_j \leq 1 \quad \forall\, j.$$

Schemes differ on the expressions used to define the coefficients α_j and β_j.

Clearly the Lax-Friedrichs, (69) and the forward time upwind schemes, (71) schemes satisfy these requirements and are therefore first-order TVD schemes.

There are also *total variation bounded (TVB)* schemes that satisfy $TV(h^{n+1}) \leq B$, for some fixed $B > 0$ depending only on $TV(h(x, 0))$. Clearly TVD implies TVB. Several bounding TVB or flux limiter TVD schemes have been proposed. Amongst these are Roe[101], Chakravarthy and Osher[14] and Shu[105][106]. A desirable property of all TVD and TVB schemes is that as $\Delta x \to 0$, if an entropy condition is satisfied, which implies uniqueness of the weak solution, then the scheme is also convergent[42].

TVD methods are restricted to second-order accuracy on smooth solutions and produce sharp resolution of discontinuities without excessive smearing using a first-order accurate scheme. TVB schemes can be globally high-order accurate in space.

TVD schemes reduce locally to first-order accuracy near local extrema in the solution profile, in order to maintain the *TVD* property which is required to suppress spurious oscillations near sharp gradients. This is achieved in general by constructing the *TVD* scheme as a nonlinear combination of a higher-order scheme and a first-order monotone scheme. The numerical dispersion of the higher-order component of the scheme, which produces oscillations in the solution is suppressed by the numerical diffusion introduced by the first-order monotone scheme in regions where necessary.

A difference scheme is called *total variation decreasing* or *diminishing* (*TVD*) if it produces a solution satisfying

$$\sum_{j=-\infty}^{\infty} |u_{j+1}^{n+1} - u_j^{n+1}| \leq \sum_{j=-\infty}^{\infty} |u_{j+1}^{n} - u_j^{n}|$$

for all $n \geq 0$.

It is natural to require *TVD* since the solution of the continuous problem $u(x,t)$ satisfies

$$TV(u(t)) \equiv \lim_{\Delta x \to 0} \frac{1}{\Delta x} \int_R |u(x+\Delta x, t) - u(x,t)|\, dx \leq TV(u(0)).$$

However, the *TVD* is sometimes too restrictive. This severe restriction can be relaxed producing an *essentially non-oscillatory* (*ENO*) scheme A difference scheme is called essentially non oscillatory if it produces a solution satisfying

$$\sum_{j=-\infty}^{\infty} |u_{j+1}^{n+1} - u_j^{n+1}| \leq \sum_{j=-\infty}^{\infty} |u_{j+1}^{n} - u_j^{n}| + O(\Delta x^p)$$

for all $n \geq 0$ and some $p \geq 1$.

A simple high resolution scheme for the advection equation is given by[74]

$$h_j^{n+1} = h_j^n - Cr(h_j^n - h_{j-1}^n) - \frac{Cr(\mathrm{sgn}(Cr) - Cr)\Delta x(\sigma_j^n - \sigma_{j-1}^n)}{2} \tag{81}$$

in which $\mathrm{sgn}(x) = 1$ if $x > 0$, -1 otherwise and σ_j^n is a measure of the slope of the profile. If $\sigma_j^n = 0$ the scheme reduces to the upwind scheme. If $\sigma_j^n = (h_{j+1}^n - h_j^n)/\Delta x$ it becomes the Lax-Wendroff scheme. Since the upwind scheme is known to be *TVD*, the oscillation which arise with the Lax-Wendroff scheme can be interpreted as a poor choice of slopes σ_j^n near steep gradients in the solution. Alternatively, this scheme could be considered as an upwind scheme with an additional term approximating a second-order diffusion term.

Slope limiters can be used to eliminate these oscillations and produce a *TVD* scheme.

In smooth regions the second-order Law-Wendroff scheme is used.

The principles of limiting is illustrated using the minmod limiter given by (92) and the solution of the linear advection equation using the basic Lax-Wendroff scheme, the Lax-Wendroff scheme using the minmod limiter and the upwind scheme using (81). In Figure 42(a) the profile at time t is evolved for one time step using the Lax-Wendroff scheme. The variable σ_j at each computational node has also been plotted Figure 42.

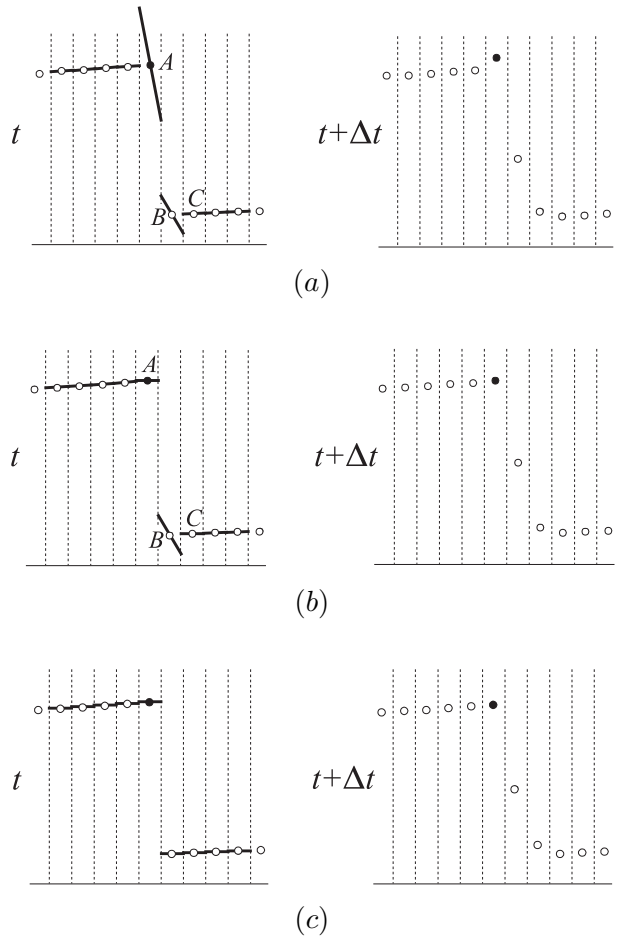


Figure 42: Piecewise linear reconstruction using (a) Lax-Wendroff slopes, (b) minmod limiter and (c) piecewise constant using first-order upwind scheme.

An overshoot is introduced at the upstream of the steep gradient in the profile, defined by the solid point in Figure 42(a). The growth of oscillations will continue as the profile evolves polluting the solution upstream of the steep gradient. At point A in Figure 42(a) $\Delta_+ h_j < 0 < \Delta_- h_j$, therefore $\sigma_j^n < 0 < \sigma_{j-1}^n$ and the last term in (81) is negative because for stability, $Cr > 0$. Therefore, negative diffusion is added to the upwind scheme. However, when the minmod limiter is used the slopes $\Delta_\pm h_j$ have opposite sign at the A and $\sigma_j = 0$ is set at this point, shown in Figure 42(b) which is equivalent as using the first-order upwind scheme. At point B, $\Delta_\pm h_j$ have the same sign. However, $|\Delta_+ h_j| > |\Delta_- h_j|$ and there is no change in σ_j at B as shown in Figure 42(b). At point C, $\Delta_\pm h_j$ have opposite sign and $\sigma_j = 0$ is used in the limited scheme, Figure 42(b). For this problem, a combination of upwind and Lax-Wendroff scheme is used to evolve and produce a solution free from oscillations at time $t + \Delta t$. Adopting $\sigma_j = 0$ at all computational nodes, which corresponds to using the upwind scheme, has not produced vastly different results shown in Figure 42(c) from those produced by the limited scheme. Both solutions eliminate the occurrence of numerical oscillations in the solution by introducing sufficient diffusion, in the limited scheme strategically to suppress these oscillations. Recalling that these results are for the evolution of the profile over one time step. Over many time steps, the diffusion introduced by the upwind scheme is greater than that produced by the limited Lax-Wendroff scheme.

There are numerous finite difference schemes which have been used to solve partial differential equations. Many of these schemes do not explicitly consider the properties of the governing equations. These are considered as *naive finite difference schemes*. A naive approximation of the intercell flux $\mathbf{f}_{j+1/2}^n$ from the known cell values at time t, is to use simple averaging between the nodal flux values, $\mathbf{F}_{j+1}$ and $\mathbf{F}_j$ so that $\mathbf{f}_{j+1/2} = \frac{1}{2}(\mathbf{F}_{j+1} + \mathbf{F}_j)$. Unfortunately, this averaging results in an unstable scheme. Other schemes attempt to remedy this by approximating higher derivatives.

We now review a number of naive methods. Two well known first-order schemes are the first-order upwind and Lax-Friedrichs schemes. To obtain higher resolution of shocks and rarefaction fan, higher-order schemes can be employed. Naive second-order finite difference schemes include; (i) Lax-Wendroff, (ii) Richtmyer and (iii) MacCormack schemes. To obtain higher resolution of shocks and rarefaction fan, higher-order schemes can be employed.

Table 3: Examples of limiters applicable to second and higher-order numerical schemes for solving conservation laws.

Limiter	$\phi(\beta^+, \beta^-)$
minmod	$\text{minmod}(1, \beta^+, \beta^-) = \begin{cases} \min(1, \lvert\beta^+\rvert, \lvert\beta^-\rvert) & \text{if } \beta^+\beta^- > 0 \\ 0 & \text{if } \beta^+\beta^- \leq 0 \end{cases}$
superbee	$\max(0, \min(1, 2\beta^-), \min(2, \beta^-)) + \max(0, \min(1, 2\beta^+), \min(2, \beta^+)) - 1$
MUSCL	$\max(0, \min(2, 2\beta^+, 2\beta^-, (\beta^+ + \beta^-)/2))$
van Leer[125]	$\frac{\beta^+ + \lvert\beta^+\rvert}{1 + \lvert\beta^+\rvert} + \frac{\beta^- + \lvert\beta^-\rvert}{1 + \lvert\beta^-\rvert} - 1$
van Albada[123]	$\frac{\beta^+ + (\beta^+)^2}{1 + (\beta^+)^2} + \frac{\beta^- + (\beta^-)^2}{1 + (\beta^-)^2} - 1$
monotonic	$\frac{\beta^+ + \lvert\beta^+\rvert}{1 + \beta^+} + \frac{\beta^- + \lvert\beta^-\rvert}{1 + \beta^-} - 2$

Equation (92) can be written as a flux limiter as [74]

$$\begin{aligned} \phi(\theta) &= \begin{cases} 0 & \text{if} \quad \theta \leq 0 \\ \theta & \text{if} \quad 0 \leq \theta \leq 1 \\ 1 & \text{if} \quad \theta \geq 1 \end{cases} \\ &= \max(0, \min(1, \theta)) \end{aligned}$$

where $\theta_j = \Delta_- h_j / \Delta_+ h_j$.

The effect of θ in (91) on the accuracy of a numerical scheme is demonstrated using the analytical solution, (**??**) and (**??**) the central upwind scheme, (**??**) with the flux given by (**??**) with the Heun time integration, (**??**) resulting in a second-order space and time. The standardized L_1 norm, given by (**??**) is calculated for the simulated and analytical water depth and velocity for several values of $\theta \in [0, 2]$ using all the computational points. All the other variables remain unchanged. In the analytical solution of flow in the one-dimensional canal with a parabolic shaped bed. The truncated canal is subdivided into 500 segments, each $\Delta x = 4.0$m, The simulated solution at $t = 1000$s is

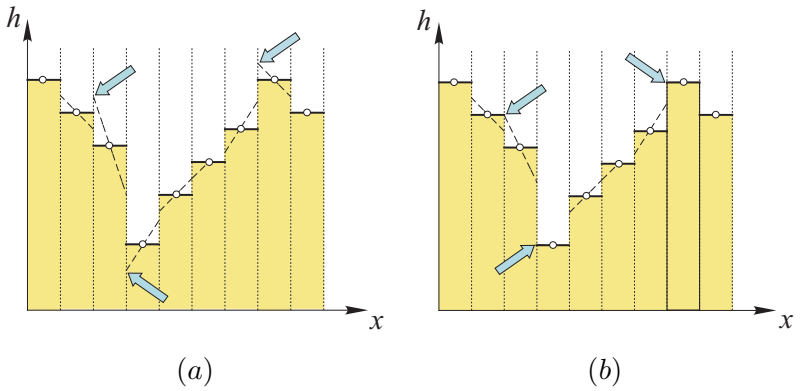


Figure 43: The reconstruction of the piecewise constant water depth using the (a) Lax-Wendroff scheme introducing extrema in the solution and (b) the TVD Lax-Wendroff scheme with the SUPERBEE limiter avoiding the introduction of extrema.

compared with the analytical solution. $B_0 = 5\text{m/s}$, $L = 3000\text{m}$, $z_\infty = 10\text{m}$, the linear friction coefficient set at $\tau = 0.001/\text{s}$, $\Delta t = \Delta x/\sqrt{(gh_0)}/2\text{s}$, the canal length is 2000m and the analytical solution, (**??**) and (**??**) is used as the boundary conditions in the numerical scheme. The effect of θ on the accuracy of the central upwind scheme is shown in Figure 44. The analytical solution chosen should be considered as a smooth solution and latter in this book is used to reveal that this numerical scheme is second-order accurate in both space and time. When $\theta = 0$ the scheme is only first-order because the limited values are simply averages of the nodal values, piecewise constant. This produces the least accurate simulations only slightly improving for $\theta < 1$. When $\theta = 1$ the limiter is second-order and the accuracy of the scheme increases significantly. Since this problem is smooth, increasing θ does not improve the accuracy of the scheme.

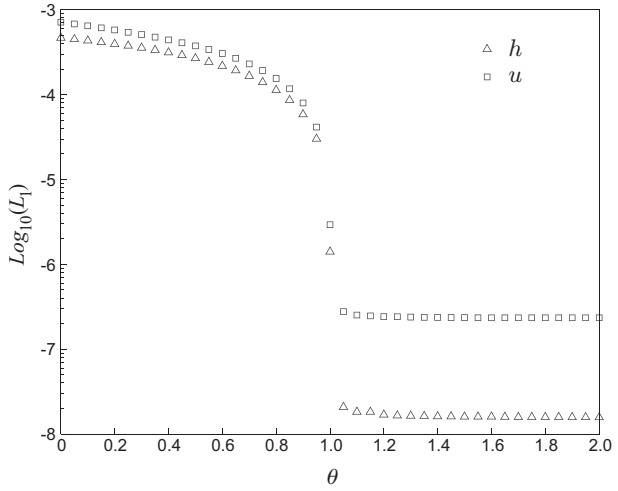


Figure 44: The effect of θ in the limiter, (91) on the accuracy of the second-order central upwind scheme.

7.1 Limiters

Examples of *total variation diminishing* (TVD) limiters include *minmod* (82), *superbee* (83), *van Leer* and *MUSCL* (85) limiters given by Yang and Prezekwas[132] and Chang[15] as

$$\phi(\beta^+, \beta^-) = \text{minmod}(1, \beta^+, \beta^-) \tag{82}$$

$$\begin{aligned}\phi(\beta^+, \beta^-) = &\max(0, \min(1, 2\beta^-), \min(2, \beta^-)) \\ &+ \max(0, \min(1, 2\beta^+), \min(2, \beta^+)) - 1\end{aligned} \tag{83}$$

$$\phi(\beta^+, \beta^-) = \frac{\beta^+ + |\beta^+|^\omega}{1 + |\beta^+|^\omega} + \frac{\beta^- + |\beta^-|^\omega}{1 + |\beta^-|^\omega} - 1 \tag{84}$$

$$\phi(\beta^+, \beta^-) = \max(0, \min(2, 2\beta^+, 2\beta^-, (\beta^+ + \beta^-)/2)) \tag{85}$$

in which $\beta^+ = \Delta_{j-1/2}/\Delta_{j+1/2}$, $\beta^- = \Delta_{j+3/2}/\Delta_{j+1/2}$, $\text{minmod}(1, \beta^\pm) = \min(1, \beta^\pm)$ if $\beta^\pm > 0$ otherwise it is equal to 0 and $\Delta_{j+1/2} = \mathbf{U}_{j+1} - \mathbf{U}_j$. When $\omega = 1$, (84) is the well known van Leer limiter[125], and when $\omega = 2$ (84) is the van Albada limiter[123] which has been used by Jha *et al.*[57]. The ratio $\beta^\pm$ is a measure of the smoothness of the data near x_j. For smooth data $\beta^\pm \approx 1$ and far from 1 near a discontinuity.

A suitable slope limiter is the *minimum-modulus (minmod) limiter*

$$\sigma_j = \frac{\text{minmod}(\Delta_+ h_j^n, \Delta_- h_j^n)}{\Delta x}$$

where

$$\begin{aligned}\text{minmod}(a, b) &= \begin{cases} a & \text{if} \quad |a| < |\Delta_- h_j^n| \quad \text{and} \quad ab > 0 \\ b & \text{if} \quad |b| < |\Delta_+ h_j^n| \quad \text{and} \quad ab > 0 \\ 0 & \text{if} \quad ab \le 0 \end{cases} \\ &= (\text{sgn}(a) + \text{sgn}(b))\text{sgn}(|a|, |b|)/2\end{aligned} \tag{86}$$

and $\Delta_\pm h_j = \pm(h_{j\pm1}^n - h_j^n)$. At extreme points of h_j^n, the slopes $\Delta_\pm h_j$ have opposite sign and $\sigma_j^n = 0$ and the first-order upwind scheme is used.

Jamison has developed a symmetric limiter positive (SLIP) limiter

$$\phi(\alpha, \beta) = \frac{\alpha + \beta}{2}\left(1 - \left|\frac{\alpha - \beta}{|\alpha| + |\beta|}\right|^q\right) \tag{87}$$

When $q = 1$ the minmod limiter is obtained, $q = 2$ the van Leer limiter and for $q > 3$ the higher-order SLIP method is obtained.

Monotonic limiter:

$$\phi(\beta^+, \beta^-) = \frac{\beta^+ + |\beta^+|}{1 + \beta^+} + \frac{\beta^- + |\beta^-|}{1 + \beta^-} - 2 \tag{88}$$

$$\text{minmod}(a_1, a_2, \cdots, a_n) = \begin{cases} 0 & \text{if} \quad \exists(i,j)\text{sign}(a_i) \neq \text{sign}(a_j) \\ \\ \text{sign}(a_i)\min(|a_i|) & \text{otherwise} \end{cases}$$

Limiters are generally nonlinear.

Flux limiters are also referred to as *slope limiters* because they both have the same mathematical form and both have the effect of limiting the solution gradient near shocks or discontinuities. In general, the term flux limiter is used when the limiter acts on system *fluxes* and slope limiter is used when the limiter acts on system *states*.

Consider the one-dimensional semi-discrete scheme

$$\frac{du_i}{dt} = \frac{1}{\Delta x_i}\left[F(u_{i+1/2}) - f(u_{j-1/2})\right] = 0$$

where $F(u_{j+1/2}$ and $F(u_{j-1/2}$ represent edge fluxes for the jth cell. If these edge fluxes can be represented by *low*-order diffusive and *high*-order anti-diffusive schemes, then a limited flux can be expressed as

$$F(u_{j+1/2}) = f^{\text{low}}_{j+1/2} - \phi(r_j)\left(f^{\text{low}}_{j+1/2} - f^{\text{high}}_{j+1/2}\right)$$

and

$$F(u_{j-1/2}) = f^{\text{low}}_{j-1/2} - \phi(r_{j-1})\left(f^{\text{low}}_{j-1/2} - f^{\text{high}}_{j-1/2}\right)$$

where f^{low} represents the flux estimated from a low-order scheme, f^{high} represents the flux estimated from a high-order scheme and $\phi(r)$ is the flux limiter with r representing the ratio of successive gradients on the solution mesh, given by

$$r_j = \frac{u_j - u_{j-1}}{u_{j+1} - u_j}.$$

Negative values of r are associated with extrema.

The limiter can be considered as a switch between the low and high-order schemes. For example Generally the limiter function is constrained to be greater than or equal to zero. Therefore, when the limiter is equal to zero, generally at sharp gradients, opposite slopes or zero gradients, the flux is represented by a *low resolution schemes*. Similarly, when the limiter is equal to 1, in smooth regions, it is represented by a *high resolution scheme*. The various limiters have different switching characteristics and are selected according to the particular application. No particular limiter has been found to work well for all problems and a particular choice is usually made on a trial and error basis.

The flux limiter must satisfy certain constraints in order to produce a TVD satisfying schemes. They should satisfy at least the following criteria,

$$r \leq \phi(r) \leq 2r \quad 0 \leq r \leq 1,$$

$$r \leq \phi(r) \leq r \quad 1 \leq r \leq 2,$$

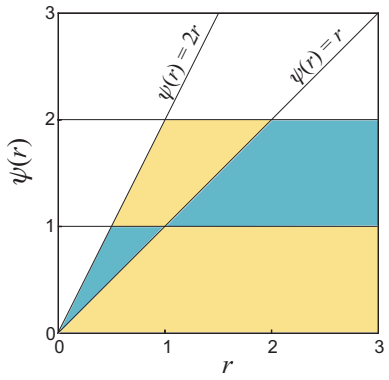


Figure 45: The admissible region for second-order TVD limiters that guarantee stability of the numerical scheme.

$$r \leq \phi(r) \leq 2 \quad r > 2,$$

and

$$\phi(1) = 1.$$

Most set $\phi(r \leq) = 0$ at extrema where $r \leq 0$. The last condition is necessary and sufficient for the scheme to be second-order accurate away from extrema. It is generally accepted that at TVD schemes revert to first-order accuracy at extrema regardless of the accuracy of the scheme in smooth regions. Zijlema and Wesseling show that appropriately designing the limiter higher accuracy is achievable at extrema and that $\phi(r) \neq 0$ for $r \leq 0$ and $\phi(r)$ may also be negative.

The admissible limiter region for second-order TVD schemes is shown in the *Sweby diagram*[115] shown in Figure 45.

Minmod is the most diffusive, and the superbee is less diffusive than the Barth-Jespersen, van Leer, van Albada and minmod limiter. The superbee follows the boundary of the TVD region. Whilst the solution is known to remain monotone with the superbee limiter, it is known to steepen gradients. Along with the Barth-Jespersen limiter they are the most compressive. van Leer is more compressive than the van Albada which is more compressive than the minmod limiter.

second-order van Leer's limiter

$$\psi(r) = \frac{r + |r|}{1 + |r|}$$

second-order monotonized central limiter[125]

$$\psi(r) = \max[0, \min(2r, (1 + r)/2, 2)]; \quad \lim_{r \to \infty} \psi(r) = 2$$

second-order minmod limiter

$$\psi(r) = \max[0, \min(1, r)]; \quad \lim_{r \to \infty} \psi(r) = 1$$

second-order superbee limiter

$$\psi(r) = \max[0, \min(2r, 1), \min(r, 2)]; \quad \lim_{r\to\infty} \psi(r) = 2$$

second-order Chakravarthy-Osher limiter

$$\psi(r) = \max[0, \min(r, \beta)]; \quad \lim_{r\to\infty} \psi(r) = \beta$$

where $1 \leq \beta \leq 2$ and it reverts to the minmod limiter when $\beta = 1$.

second-order MUSCL limiter[126]

$$\psi(r) = \max[0, \min(2r, \frac{1}{2} + \frac{1}{2}r, 2)]; \quad \lim_{r\to\infty} \psi(r) = 2$$

second-order Sweby limiter

$$\psi(r) = \max[0, \min(\beta r, 1), \min(r, \beta)]; \quad \lim_{r\to\infty} \psi(r) = \beta$$

where $1 \leq \beta \leq 2$ and it reverts to the minmod limiter when $\beta = 1$.

second-order Ospre limiter

$$\psi(r) = \frac{3}{2}\frac{r^2 + r}{r^2 + r + 1}; \quad \lim_{r\to\infty} \psi(r) = 1.5$$

second-order van Albada's limiter

$$\psi(r) = \frac{r^2 + r}{1 + r^2}; \quad \lim_{r\to\infty} \psi(r) = 1$$

second-order van Leer limiter

$$\psi(r) = \frac{r + |r|}{1 + |r|}; \quad \lim_{r\to\infty} \psi(r) = 2$$

All the above limiters, other than Osher's limiter exhibit the following symmetry property, $\phi(r)/r = \phi(1/r)$. This ensures that the limits for forward and backward gradients are treated the same way. All pass through $\phi(r = 1) = 1.0$ and it is expected that $\phi(r) \geq 0 \; \forall \; r$, but this is not the case.

Third-order Spekreijse's limiter

$$\psi(r) = \left(\frac{1}{3} + \frac{2}{3}r\right)\frac{3r^3 - 2r^2 + 3r}{2 + 2r^4}; \quad \lim_{r\to\infty} \psi(r) = 1$$

Third-order Koren's limiter

$$\psi(r) = \max(0, \min(2, 2r, \frac{1}{3} + \frac{2}{6}r)); \quad \lim_{r\to\infty} \psi(r) = 2$$

Third-order Piperno Depeyre[95]

$$\psi(r) = \left(\frac{1}{3} + \frac{2}{3}r\right)\varphi(\frac{1}{r})$$

where

$$\varphi(r) = \begin{cases} 0 & \text{if} \quad r \leq 0 \\ 1 + (\frac{3}{2}r + 1)(r-1)^3 & \text{if} \quad 0 \leq r \leq 1 \\ \frac{3r^2 - 6r + 19}{r^3 - 3r + 18} & \text{if} \quad 1 \leq r \end{cases}$$

minmod limiter

$$\text{minmod}(a, b) = sign(a) \max[0, \min(|a|, sign(a)b)].$$

The limiters are shown in Figure 47

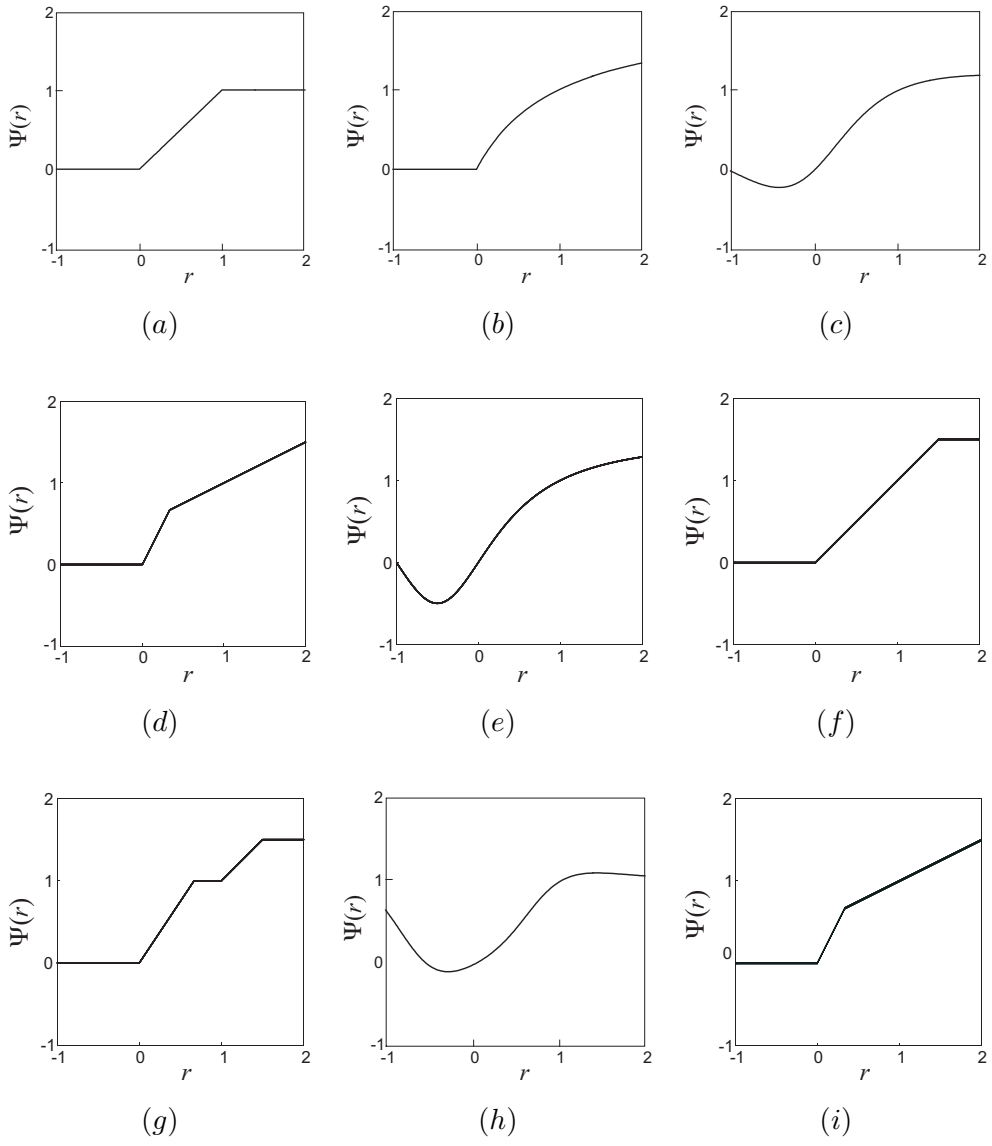


Figure 46: Second-order limiter functions $\psi(r)$ for (a) mimmod, (b) van Leer, (c) van Albada, (d) monotonized central, (e) Ospre, (f) Osher with $\beta = 1.5$ when $\beta = 1$ this is equivalent to the minmod limiter????, (g) Sweby with $\beta = 1.5$ this is equivalent to the minmod limiter????, (h) superbee and (i) MUSCL limiters.

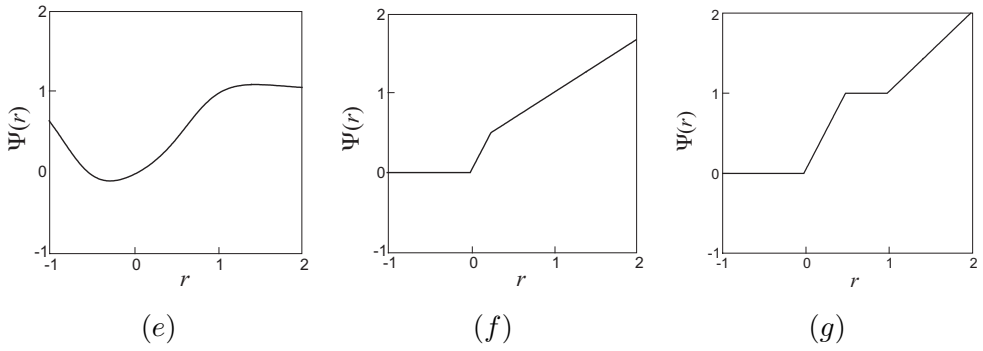


Figure 47: Third-order limiter functions $\psi(r)$ for (a) Spekreijse, (f) Koren and (g) Piperno and Depeyre limiters.

Observing from Figure 43(a) that the Lax-Wendroff slope introduces extrema near minima and maxima of the average cell solution. The total variation is increasing. To obtain a total variation diminishing solution the slope σ_j^n is limited so that the condition $h_{j\pm1/2}^{\mp} \in [h_{j-1}^n, h_j^n]$ where $h_{j\pm1/2}^{-}$ is the value of $h_{j-1/2}^n$ in the interval $[x_{j-3/2}, x_{j-1/2}]$ and $h_{j\pm1/2}^{+}$ is the value of $h_{j-1/2}^n$ in the interval $[x_{j-1/2}, x_{j+1/2}]$. The limited slope is expressed as

$$\sigma_j^n = \max(0, \min(1, 2\theta_j^n), \min(2, \theta_j^n)) \frac{h_{j+1}^n - h_j^n}{\Delta x} = \phi(\theta_j^n)\sigma_j^n$$

which is equal to 1 except near extrema where this flux limits the Lax-Wendroff slope. The application of σ_j^n is shown in Figure 43(b). The scheme is TVD and the resulting scheme is called the TVD Lax-Wendroff SUPERBEE scheme.

Jaing and Tadmor[58] modified minmod limiter

$$w_j' = \phi(\theta(w_{j+1}^n - w_j^n), (w_{j+1}^n - w_{j-1}^n)/2, \theta(w_j^n - w_{j-1}^n)) \tag{91}$$

where

$$\phi(a_1, a_2, \cdots, a_n) = \begin{cases} \min_p(a_i) & \text{if} \quad a_i > 0 \quad \forall \quad i \\ \max_p(a_i) & \text{if} \quad a_i < 0 \quad \forall \quad i \\ 0 & \text{otherwise.} \end{cases}$$

When $\theta = 1$ results in the classical minmod limiter. When $1 < \theta < 2$, it is a variant of the minmod limiter and when $\theta > 2$ it is the Harten and Osher UNO limiter[46]. When $\theta < 1$ the limiter is not second-order, with $\theta = 0$ it is a first-order piecewise linear limiter. A non-oscillatory approximation of the derivative is possible using

$$u_j' = \phi(\Delta u_{j+1/2}, \Delta u_{j-1/2})$$

or

$$u_j' = \phi(\theta\Delta u_{j+1/2}, (\Delta u_{j+1} - \Delta u_{j-1})/2, \theta\Delta u_{j-1/2}), \qquad 1 \le \theta \le 4$$

where $\phi(\omega_1, \omega_2, \ldots)$ is the minmod non-linear limiter given by

$$\phi(\omega_1, \omega_2, \ldots) = \begin{cases} \min(\omega_i) \ \text{ if } \ \omega_i > 0 & \forall\, i \\ \max(\omega_i) \ \text{ if } \ \omega_i < 0 & \forall\, i \\ 0 & \text{otherwise,} \end{cases} \tag{92}$$

with $\Delta u_j = u_{j+1} - u_j$. An alternative to the above total variation diminishing (TVD) limiters is the following universally non-oscillatory (UNO) limiter

$$u_j' = \phi(\Delta u_{j-1/2} + \frac{1}{2}\phi(\Delta^2 u_{j-1}, \Delta^2 u_j), \Delta u_{j+1/2} - \frac{1}{2}\phi(\Delta^2 u_j, \Delta^2 u_{j+1}))$$

with $\Delta^2 u_j = u_{j+1} - 2u_j + u_{j-1}$.

An example of a non-linear limiter is the *minmod* limiter

$$s_j = \text{minmod}\left(\theta\frac{\overline{u}_j - \overline{u}_{j-1}}{\Delta x}, \frac{\overline{u}_{j+1} - \overline{u}_{j-1}}{2\Delta x}, \frac{\overline{u}_{j+1} - \overline{u}}{\Delta x}\right),\ 1 \leq \theta \leq 2,$$

where

$$\text{minmod}(x_1, x_2, x_3, ...) = \begin{cases} \min\limits_j(x_j) & \text{if } x_j > 0,\ \forall j, \\ \max\limits_j(x_j) & \text{if } x_j < 0,\ \forall j, \\ 0 & \text{otherwise.} \end{cases}$$

To eliminate the oscillations introduced by naive second-order schemes, some form of dissipation or limiter is used. A number of limiting and dissipation strategies have been developed for the Lax-Wendroff and McCormack schemes.

Dissipation involves the addition of an artificial viscosity which approximates $\mathbf{U}_x$ to the basic flux. For example, the Lax-Wendroff flux given by (79) is replaced with $\mathbf{f}_{j+1/2} = \mathbf{f}^{LW}_{j+1/2} - \varepsilon_{j+1/2}(\mathbf{U}_{j+1} - \mathbf{U}_j)$ where $\varepsilon_{j+1/2}$ is at least proportional to Δx to maintain second-order accuracy (Hirsch[51]). There are numerous approaches for estimating the artificial viscosity that is added to a numerical scheme. A number of these are described and demonstrated using the hypothetical examples.

Oscillations in the second-order MacCormack scheme can be reduced by using artificial viscosity or dissipation applied to the conserved variables at the predictor-corrector steps, $\mathbf{U}^{n+1}_j = \mathbf{U}^{\text{Mac}}_j + \varepsilon_{j+1/2}(\mathbf{U}^{\text{Mac}}_{j+1} - \mathbf{U}^{\text{Mac}}_j) - \varepsilon_{j-1/2}(\mathbf{U}^{\text{Mac}}_j - \mathbf{U}^{\text{Mac}}_{j-1})$. This approach has been used by Fennema and Chaudhry[31], Fennema and Chaudhry[32], Gharangik and Chaudhry[36] and Rahman and Chaudhry[96]. The diffusion coefficient $\varepsilon_{j+1/2}$ is computed from a normalized form of the gradients of one of the dependent variables and it also depends on an empirical parameter $K, (0.5 \leq K \leq 3)$ which is used to regulate the amount of dissipation (Jameson*et al.*[55]). Third-order artificial viscosity can also be used (see, for example Hirsch[50], p. 279). The unusual feature of this scheme is that the viscosity is applied after the solution has been calculated.

It has been demonstrated from the above examples that second-order naive finite difference schemes introduce dispersion. To eliminate the oscillations introduced by these schemes, some form of limiter or dissipation is used. There are two types of limiters; *slope* limiters and *flux* limiters. Slope limiters impose constraints on the gradients of the dependent variables. Flux limiters impose constraints on the gradients of the flux functions. The two types of limiters are equivalent for constant coefficient scalar equations. A number of these schemes are also considered here. These include; (*i*) MacCormack scheme with artificial viscosity and dissipation, (*ii*) Lax-Wendroff scheme with flux limiters and (*iii*) the Flux Corrected Transport scheme.

Second-order naive finite difference schemes introduce dispersion. To eliminate the oscillations introduced by these schemes, some form of limiter or dissipation is used in the scheme.

Lax-Wendroff Scheme

Sweby[115] used the underlying philosophy of the flux corrected transport scheme of Boris and Book [9] to develop a generalization of the total variation diminishing Lax-Wendroff schemes. It simply adds a conservative dissipation term to the Lax-Wendroff scheme which ensures that the final scheme is total variation diminishing. Yee[134] extended the work of Sweby and developed symmetric *total variation diminishing* schemes.

The solution is advanced in time using

$$\mathbf{U}_j^{n+1} = \mathbf{U}_j^n - \frac{\Delta t}{\Delta x}\left(\mathbf{f}_{j+1/2}^n - \mathbf{f}_{j-1/2}^n\right) \tag{93}$$

in which

$$\mathbf{f}_{j+1/2} = \mathbf{f}_{j+1/2}^{UP} + \phi(\beta^+, \beta^-)\left(\mathbf{f}_{j+1/2}^{LW} - \mathbf{f}_{j+1/2}^{UP}\right)$$

where, $\mathbf{f}_{j+1/2}^{UP}$ is the first-order conservative upwind flux, $\mathbf{f}_{j+1/2}^{LW}$ is the second-order conservative Lax-Wendroff flux, and $\phi(\beta^+, \beta^-)$ is a limiter function which ensures that the scheme is total variation diminishing in which $\beta^{\pm}$ are ratios of conservative variations. There are several limiters available[132]. These include; (i) *minmod*, (ii) *MUSCL* and (iii) *superbee* limiters, see Table 3.

Results for the test problem using ; Steger Warming, Upwind Steger and Warming first-order upwind scheme[109] and the above limiters are illustrated in Figures **??**(c), **??**(d), and **??**(e) respectively.

For scalar equations, all these limiters add sufficient diffusion in the scheme to produce a monotone preserving scheme. This is not necessarily the case for a non-linear system of equations. For the shallow water wave equations, the minmod and MUSCL limiters have not introduced sufficient diffusion to completely eliminate the oscillations in the solution. They have however produced better resolution of the shocks and rarefaction fan than the schemes which use the monotonic and superbee limiters. In these schemes, the spurious oscillations in the solution have been suppressed by the additional diffusion introduced. The price paid for suppressing oscillations in the solution, is that these schemes are also diffusive. A problem encountered by many first-order upwind schemes is the simulation of a non-physical solution at the sonic point, that is, at the transition from subcritical to supercritical flows. This problem will be discussed in Section 5. They do not resolve shocks or rarefaction fan to the same accuracy as the schemes which use the minmod or MUSCL limiters.

MacCormack Scheme

Oscillations in the second-order MacCormack scheme can be reduced by using artificial viscosity or dissipation. The MacCormack scheme with artificial viscosity can be written as

$$\frac{\partial \mathbf{U}}{\partial t} + \frac{\partial \mathbf{F}}{\partial x} = \mathbf{D}\mathbf{U} \tag{94}$$

where $\mathbf{D}$ is a diffusion operator. Using second-order differences, the operator becomes

$$\mathbf{D}\mathbf{U} = \varepsilon_{j+1/2}(\mathbf{U}_{j+1}^{n+1} - \mathbf{U}_j^{n+1}) - \varepsilon_{j-1/2}(\mathbf{U}_j^{n+1} - \mathbf{U}_{j-1}^{n+1})$$

where $\varepsilon_{j+1/2}$ is computed from a normalized form of the gradients of one dependent variable. The non-linear coefficient $\varepsilon_{j+1/2}$ is defined as[55]

$$\varepsilon_{j+1/2} = K \max(\varepsilon_j, \varepsilon_{j+1})$$

where, for the dependent variable, h

$$\varepsilon_j = \frac{|h_{j+1} - 2h_j - h_{j-1}|}{|h_{j+1}| + |2h_j| + |h_{j-1}|}$$

and an empirical parameter $K, (0.5 \leq K \leq 3)$ which is used to regulate the amount of dissipation[55]. Other dependent variables could be used to define ε_j. Fennema and Chaudhry[31] used $K = 1.3$ in their simulation of the propagation of a bore in a rectangular frictionless channel with a horizontal bed. Fennema and Chaudhry[32] obtained satisfactory results with $K = 0.25$ in the simulation of a partial breach of a two-dimensional dam break in a frictionless horizontal channel. The unusual feature of this scheme is that the viscosity is applied after the solution has been obtained.

Davis [24] describes a strategy for introducing artificial dissipation in a numerical scheme so that the scheme is *total variation diminishing*. The result is a dissipation term which is based on the theory of total variation diminishing schemes and which does not contain any free parameters. The dissipative terms are added to the solution of the MacCormack scheme so that the solution is

$$\mathbf{U}_j^{n+1} = \mathbf{U}_j^{\text{mac}} + \left(\mathbf{D}_{j+1/2} - \mathbf{D}_{j-1/2}\right)$$

where $\mathbf{U}_j^{\text{mac}}$ is the MacCormack scheme solution and $\mathbf{D}_{j\pm1/2}$ are the artificial dissipation introduced to produce a total variation diminishing scheme. The dissipation terms are estimated using

$$\mathbf{D}_{j+1/2} = \bar{\mathbf{K}}^+_{j+1/2}(\beta_j^+) + \bar{\mathbf{K}}^-_{j+1/2}(\beta_{j+1}^-)\left(\mathbf{U}_{j+1}^n - \mathbf{U}_j^n\right)$$

in which

$$\bar{\mathbf{K}}^{\pm}(\beta^{\pm}) = 0.5C(\nu)(1 - \phi(\beta^{\pm}))$$

where

$$\nu = \max_j |\lambda_j| \frac{\Delta t}{\Delta x},$$

$$C(\nu) = \begin{cases} \nu(1-\nu) & \text{if} \quad \nu \leq 0.5 \\ 0.25 & \text{if} \quad \nu > 0.5 \end{cases},$$

$$\beta_j^{\pm} = \frac{\left(\Delta\mathbf{U}_{j-1/2}^n, \Delta\mathbf{U}_{j+1/2}^n\right)}{\left(\Delta\mathbf{U}_{j\pm1/2}^n, \Delta\mathbf{U}_{j\pm1/2}^n\right)}$$

where $(\cdot,\cdot)$ denotes the usual inner product and λ_j are the eigenvalues. The limiter $\phi(\beta)$ is defined as

$$\phi(\beta) = \begin{cases} \min(2\beta, 1) & \text{if} \quad \beta > 0 \\ 0 & \text{if} \quad \beta \leq 0 \end{cases}$$

which is the *minmod* limiter.

The theory which is used to establish a total variation diminishing scheme is related to scalar equations. Their application to non-linear problems does not ensure that the solution will be total variation diminishing. This is the case here, where the solution contains some small oscillations. However, it has produced excellent resolution of the shocks and rarefaction fan and most of the oscillations associated with second-order schemes scheme have been suppressed. In addition, this scheme was able to produce a solution for the second example, which was not the case for the second-order Lax-Wendroff scheme and the MacCormack scheme, with and without artificial viscosity.

Most of the numerical schemes examined in this article produced similar results to those shown in Figure **??**(e). This was not necessarily the case for the more severe example shown in Figure **??**(a)

7.2 Flux Corrected Transport

The flux corrected transport scheme was originally developed by Boris and Book [9] and was extended later by Zalasek [136]. The two-step flux corrected transport scheme is a hybrid scheme consisting of a combination of a low-order total variation diminishing scheme and a higher-order scheme. Here, the low-order solutions are calculated using

$$\mathbf{U}_j^{n+1} = \mathbf{U}_j^n - \frac{\Delta t}{\Delta x}\left(\mathbf{f}_{j+1/2}^{\mathrm{UP}} - \mathbf{f}_{j-1/2}^{\mathrm{UP}}\right)$$

which is an upwinding scheme. These values are corrected by adding some anti-diffusion flux to the upwind solution. The amount of anti-diffusion is given by

$$\mathbf{A}_{j+1/2} = \mathbf{f}_{j+1/2}^{\mathrm{LW}} - \mathbf{f}_{j+1/2}^{\mathrm{UP}}$$

where, $\mathbf{f}^{\mathrm{LW}}$ is the flux calculated using a higher-order scheme such as the second-order Lax-Wendroff scheme and $\mathbf{f}^{\mathrm{UP}}$ is the flux calculated using the Steger and Warming first-order upwind scheme. This flux is limited so that neither new extrema are created or existing extrema are accentuated. The limiter is;

$$\mathbf{A}_{j+1/2}^c = \mathbf{S}\max\left(0, \min\left[\left|\mathbf{A}_{j+1/2}\right|, \frac{\mathbf{S}\Delta x\left(\mathbf{U}_{j+2}^{\mathrm{UP}} - \mathbf{U}_{j+1}^{\mathrm{UP}}\right)}{\Delta t}, \frac{\mathbf{S}\Delta x\left(\mathbf{U}_{j}^{\mathrm{UP}} - \mathbf{U}_{j-1}^{\mathrm{UP}}\right)}{\Delta t}\right]\right)$$

where $\mathbf{S} = \mathrm{sgn}(\mathbf{A}_{j+1/2})$. Therefore, the corrected values are given by;

$$\mathbf{U}_j^{n+1} = \mathbf{U}_j^{\mathrm{UP}} - \frac{\Delta t}{\Delta x}\left(\mathbf{A}_{j+1/2}^c - \mathbf{A}_{j-1/2}^c\right).$$

In regions where the variables vary smoothly, the high-order scheme is employed. In regions where the variables vary rapidly, the low-order scheme is employed.

For the test example, the flux corrected transport produced the results shown in Figure **??**(f) There is excellent resolution of the shocks and rarefaction fan by this scheme. Unfortunately, the

limiter employed in this scheme does not suppress all the oscillations associated with the second-order scheme. In addition, the scheme has also produced an entropy violating solution.

Alternative strategies have been developed so that only sufficient diffusion is applied to naive second-order schemes in an attempt to eliminate the numerical oscillations. Central to this strategy and to the application of *flux* and *slope* limiters is how much limiting is necessary to avoid numerical oscillations in the solution. Slope limiters impose constraints on the gradients of the dependent variables. Flux limiters impose constraints on the gradients of the flux functions. A measure of the oscillations in a solution is provided by the *total variation* $TV(\mathbf{U}^n) = \sum_{j=-\infty}^{\infty} |\mathbf{U}_j^n - \mathbf{U}_{j-1}^n|$ of the grid values $\mathbf{U}_j$ at time level n. Oscillations are avoided if the total variation does not increase with time. Many higher-order schemes are constructed so that $TV(\mathbf{U}^n) \leq TV(\mathbf{U}^{n+1})$. It is possible to use limiters to produce TVD MacCormack and Lax-Wendroff schemes (see for example Garcia-Navarro *et al.*[34], Jha *et al.*[57], Nakatani and Komura[87], Rider[98] and Yee[135]).

Davis[24] describes a strategy for introducing artificial dissipation in a numerical scheme so that the scheme is total variation diminishing. The result is a dissipation term which does not contain free parameters. The dissipative terms are added to the solution of the MacCormack scheme so that the solution is $\mathbf{U}_j^{n+1} = \mathbf{U}_j^{\text{Mac}} + (\mathbf{D}_{j+1/2} - \mathbf{D}_{j-1/2})$ where $\mathbf{U}_j^{\text{Mac}}$ is the MacCormack scheme solution and $\mathbf{D}_{j\pm1/2}$ is artificial dissipation introduced to produce a total variation diminishing scheme. The nonlinear dissipation term depends on the Courant-Friedrichs-Lewy number and is limited so that oscillations in the solution are suppressed.

The theory which is used to establish a total variation diminishing scheme is based on scalar equations. Their application to non-linear problems does not ensure that the solution will be total variation diminishing. This is the case here, where the solution contains some small oscillations. However, it has produced excellent resolution of the shocks and rarefaction fan and most of the oscillations associated with second-order schemes have been suppressed. In addition, this scheme was able to produce a solution for the second example, which was not the case for the second-order Lax-Wendroff and the MacCormack naive finite difference scheme, with and without artificial viscosity.

Most of the numerical schemes examined in this paper produce similar results to those shown in Figure **??**(e). This was not necessarily the case for the more severe example shown in Figure **??**.

In addition, the scheme has produced an entropy violating solution at the *sonic point* (Hirsch[51]) which is located at the dam site $x = 1,000$m. There is a discontinuity in the simulated water depth, h and velocity, u which occurs when *critical flow*, $\lambda = 0$m/s occurs and there is a transition from subcritical to supercritical flow or *visa versa*. In the example where $h_0 = 5$m, flow is always subcritical, sonic points are not encountered and the solution is continuous along the rarefaction fan. In the second example, where $h_0 = 0.1$m a transition between subcritical and supercritical flow occurs. In this case the sonic point occurs at the original location of the dam.

Sweby[115] used the underlying philosophy of the flux-corrected transport scheme of Boris and Book[9] to develop a generalization of the total variation diminishing Lax-Wendroff schemes. It

simply adds a dissipation term to the Lax-Wendroff scheme which ensures that the final scheme is total variation diminishing. Yee[134] extended the work of Sweby and developed symmetric total variation diminishing schemes. The flux is given by

$$\mathbf{f}_{j+1/2} = \mathbf{f}^{\mathrm{UP}}_{j+1/2} + \phi(\beta^+, \beta^-) \left(\mathbf{f}^{\mathrm{LW}}_{j+1/2} - \mathbf{f}^{\mathrm{UP}}_{j+1/2} \right)$$

where, $\mathbf{f}^{\mathrm{UP}}_{j+1/2}$ is the first-order conservative upwind flux, $\mathbf{f}^{\mathrm{LW}}_{j+1/2}$ is the second-order conservative Lax-Wendroff flux, and $\phi(\beta^+, \beta^-)$ is a limiter function which ensures that the scheme is total variation diminishing. Results for the test problem using Steger and Warming first-order upwind scheme (Steger and Warming[109]) are illustrated in Figures **??**(b), **??**(c), **??**(d) and **??**(e) using the minmod, monotonic, *Monotonic Upstream Schemes for Conservation Laws* (MUSCL) (van Leer [124]) and superbee limiters respectively. For scalar equations, all these limiters add sufficient diffusion in the scheme to produce a monotone preserving scheme. This is not necessarily the case for a non-linear system of equations. For the shallow water wave equations, the minmod and MUSCL limiters have not introduced sufficient diffusion to completely eliminate the oscillations in the solution. They have however, produced better resolution of the shocks and rarefaction fan than the schemes which use the *monotonic* (88) and superbee limiters. In these schemes, the spurious oscillations in the solution have been suppressed by the additional diffusion they have introduced. The price paid for suppressing oscillations in the solution, is that these schemes are also diffusive. In addition, the Steger and Warming scheme, like many other first-order upwind schemes, produce entropy violation solutions at the original location of the dam. This is demonstrated in the next section. In the symmetric total variation diminishing scheme, the overall performance of the scheme is very dependent on the performance of the monotone first-order scheme.

8 Upwind Schemes

Schemes can be developed which incorporate the physical propagation properties of the solution. These schemes are called *upwind* schemes, in which space discretizations are consistent with the propagation properties of the problem. A number of first-order *flux vector splitting* upwind schemes have been considered. These are by Steger and Warming[109], Bermudez and Vazquez[7], Yang *et al.*[133] and Harten.

The simplest first-order conservative scheme is the Upwind scheme (see, for example Hirsch[50], p. 414). For subcritical flow, from left to right, both characteristic speeds are positive. Information only travels downstream and the first-order upwind scheme is;

$$\mathbf{U}_j^{n+1} = \mathbf{U}_j^n - \frac{\Delta t}{\Delta x}\left(\mathbf{F}_j^n - \mathbf{F}_{j-1}^n\right) - \Delta t \mathbf{S}_j^n. \tag{95}$$

For supercritical flow, from right to left, both characteristic speeds are also positive. Information only travels upstream and the first-order upwinding scheme becomes

$$\mathbf{U}_j^{n+1} = \mathbf{U}_j^n - \frac{\Delta t}{\Delta x}\left(\mathbf{F}_{j+1}^n - \mathbf{F}_j^n\right) - \Delta t \mathbf{S}_j^n. \tag{96}$$

For subcritical flows the two characteristics have different signs. Information travels both upstream and downstream and a combination of the above upwind schemes is required. To construct an upwind scheme valid for all flow conditions, the flux related to the positive characteristic speed must be separated from the flux related to the negative characteristic. Therefore, (**??**) becomes

$$\frac{\partial \mathbf{U}}{\partial t} + \frac{\partial \mathbf{F}^+}{\partial x} + \frac{\partial \mathbf{F}^-}{\partial x} + \mathbf{S} = 0 \tag{97}$$

where $\partial \mathbf{F}^+/\partial x$ is discretized using backward differencing and $\partial \mathbf{F}^-/\partial x$ is discretized using forward differencing. This equation can be written in the quasi-linear form by splitting the Jacobian, $\mathbf{A}(\mathbf{U})$ so that

$$\frac{\partial \mathbf{U}}{\partial t} + \mathbf{A}\frac{\partial \mathbf{U}}{\partial x} = \mathbf{A}^+\frac{\partial \mathbf{U}}{\partial x} + \mathbf{A}^-\frac{\partial \mathbf{U}}{\partial x} + \mathbf{S} = 0.$$

This is a form of the Steger-Warming[109] flux splitting. It however, implies that $\mathbf{F} = \mathbf{A}\mathbf{U}$. This is valid if $\mathbf{F}(\mathbf{U})$ is a homogeneous function of *degree one*[8]. The Euler equations have this special

[8] A function $f(\mathbf{x})$ of $m \geq 2$ variables $x_1, \ldots, x_m$ is homogeneous of degree n in these variables if and only if, for $\alpha > 0$; $f(\alpha x_1, \ldots, \alpha x_m) = \alpha^n f(x_1, \ldots, x_m)$. For a function which is homogeneous of degree 1 $f(x) = x\, df(x)/dx$ for any state x. An equation $\phi(x, y, z) = 0$ is a homogeneous equation of degree n if

$$\phi(\lambda x, \lambda y, \lambda z) = \lambda^n \phi(x, y, z).$$

Differentiating $\phi(\lambda x, \lambda y, \lambda z)$ with respect to λ, keeping x, y and z unchanged, then

$$\frac{\partial \phi(\lambda x, \lambda y, \lambda z)}{\partial \lambda} = x\frac{\partial \phi}{\partial(\lambda x)} + y\frac{\partial \phi}{\partial(\lambda y)} + z\frac{\partial \phi}{\partial(\lambda z)}. \tag{98}$$

Using the definition of the homogeneous equation, this becomes

$$\frac{\partial \phi(\lambda x, \lambda y, \lambda z)}{\partial \lambda} = \frac{\partial(\lambda^n \phi)}{\partial \lambda} = n\lambda^{n-1}\phi(x, y, z) \tag{99}$$

property, however both inviscid Burgers and the shallow water wave equations are homogeneous functions of *degree two*, so that $\mathbf{F} \neq \mathbf{AU}$ and this form of flux splitting is not appropriate here (see, for example Hirsch [50], p. 417).

8.1 Bermudez and Vazquez Scheme

Equation (102) assumes that $\mathbf{F} = \mathbf{AU}$. This is valid if $\mathbf{F}(\mathbf{U})$ is a homogeneous function of degree one. The shallow water wave equation is a homogeneous function of order two, so that $\mathbf{F} \neq \mathbf{AU}$. The form of flux splitting used by Steger and Warming is not appropriate here (see, for example Hirsch[50], p. 417). It is possible to construct a matrix, $\mathbf{G}$ that does satisfy $\mathbf{F} = \mathbf{GU}$[7]. For (**??**)

$$\mathbf{G} = \mathbf{P}\mathbf{\Lambda}\mathbf{P}^{-1} = \begin{bmatrix} 0 & 1 \\ -u^2 + \frac{1}{2}gh & 2u \end{bmatrix}$$

where

$$\mathbf{P} = \begin{bmatrix} 1 & -1 \\ u + c\sqrt{2} & u - c\sqrt{2} \end{bmatrix}, \ \mathbf{P}^{-1} = \frac{1}{2c} \begin{bmatrix} -u\sqrt{2} + c & \sqrt{2} \\ u\sqrt{2} + c & \sqrt{2} \end{bmatrix}$$

and the eigenvalues of $\mathbf{G}$ are

$$\mathbf{\Lambda} = \begin{bmatrix} u + c/\sqrt{2} & 0 \\ 0 & u - c/\sqrt{2} \end{bmatrix}.$$

which are used in (102) to estimate $|\mathbf{F}| = |\mathbf{G}|\mathbf{U}$.

The flux can now be decomposed as follows. Let $\mathbf{\Lambda} = \mathbf{\Lambda}^+ + \mathbf{\Lambda}^-$, where $\mathbf{\Lambda}^+$ and $\mathbf{\Lambda}^-$ have diagonal elements, λ_1^+, λ_2^+ and λ_1^-, λ_2^- respectively. The eigenvalues, λ_i associated with $\mathbf{\Lambda}$ can also be decomposed into $\lambda_i = \lambda_i^+ + \lambda_i^-$ where

$$\lambda_i^+ = \frac{\lambda_i + |\lambda_i|}{2}, \lambda_i^- = \frac{\lambda_i - |\lambda_i|}{2}. \tag{100}$$

Since $\mathbf{F} = \mathbf{GU} = \mathbf{P}\mathbf{\Lambda}\mathbf{P}^{-1}\mathbf{U}$, then

$$\begin{aligned} \mathbf{F} &= \mathbf{P}\left(\mathbf{\Lambda}^+ + \mathbf{\Lambda}^-\right)\mathbf{P}^-1\mathbf{U} = \left(\mathbf{P}\mathbf{\Lambda}^+\mathbf{P}^{-1} + \mathbf{P}\mathbf{\Lambda}^-\mathbf{P}^{-1}\right)\mathbf{U} \\ &= \left(\mathbf{G}^+ + \mathbf{G}^-\right)\mathbf{U} = \mathbf{F}^+ + \mathbf{F}^- \end{aligned}$$

where, $\mathbf{G}^+ = \mathbf{P}\mathbf{\Lambda}^+\mathbf{P}^{-1}$, $\mathbf{G}^- = \mathbf{P}\mathbf{\Lambda}^-\mathbf{P}^{-1}$, $\mathbf{F}^+ = \mathbf{G}^+\mathbf{U}$, $\mathbf{F}^- = \mathbf{G}^-\mathbf{U}$ and $\mathbf{G} = \mathbf{G}^+ + \mathbf{G}^-$. Therefore, it is possible to decompose the fluxes.

---

and for the special case when $\lambda = 1$, combining (98) and (99) then

$$n\phi = x \left.\frac{\partial\phi}{\partial(\lambda x)}\right|_{y,z} + y \left.\frac{\partial\phi}{\partial(\lambda y)}\right|_{x,z} + z \left.\frac{\partial\phi}{\partial(\lambda z)}\right|_{x,y}$$

which is known as the *Euler's identity*

Discretizing, (97) with $\mathbf{S}^T = [0, 0]$ (see, for example Hirsch[50], p. 427);

$$\mathbf{U}_j^{n+1} = \mathbf{U}_j^n - \frac{\Delta t}{\Delta x}\left[\left(\mathbf{F}_j^+ - \mathbf{F}_{j-1}^+\right)^n + \left(\mathbf{F}_{j+1}^- - \mathbf{F}_j^-\right)^n\right]. \tag{101}$$

Need to evaluate $\mathbf{F}^+$ and $\mathbf{F}^-$ in terms of the nodal fluxes. Using (100)

$$\lambda_i^+ = \lambda_i^- + |\lambda_i|$$

or

$$\mathbf{\Lambda}^+ = \mathbf{\Lambda}^- + |\mathbf{\Lambda}|.$$

Then

$$\begin{aligned}\mathbf{F}^+ &= \mathbf{G}^+\mathbf{U} = \mathbf{P}\left(\mathbf{\Lambda}^- + |\mathbf{\Lambda}|\right)\mathbf{P}^{-1}\mathbf{U} \\ &= \left(\mathbf{G}^- + |\mathbf{G}|\right)\mathbf{U} = \mathbf{G}^-\mathbf{U} + |\mathbf{G}|\mathbf{U} = \mathbf{F}^- + |\mathbf{F}|\end{aligned}$$

so $|\mathbf{G}| = \mathbf{G}^+ - \mathbf{G}^-$ and $|\mathbf{F}| = |\mathbf{G}|\mathbf{U} = \mathbf{F}^+ - \mathbf{F}^-$. Using the *consistency condition*, $\mathbf{F} = \mathbf{F}^+ + \mathbf{F}^-$ then $\mathbf{F}^- = (\mathbf{F} - |\mathbf{F}|)/2$, $\mathbf{F}^+ = (\mathbf{F} + |\mathbf{F}|)/2$ and $|\mathbf{G}| = \mathbf{P}|\mathbf{\Lambda}|\mathbf{P}^{-1}$. Substituting into (101) then

$$\begin{aligned}\mathbf{U}_j^{n+1} = \mathbf{U}_j^n &- \frac{\Delta t}{2\Delta x}\left(\mathbf{F}_{j+1}^n - \mathbf{F}_{j-1}^n\right) \\ &+ \frac{\Delta t}{2\Delta x}\left(|\mathbf{F}|_{j+1}^n - 2|\mathbf{F}|_j^n + |\mathbf{F}|_{j-1}^n\right).\end{aligned} \tag{102}$$

In (102) the artificial viscosity term is proportional to the second derivative of the absolute flux value, $|\mathbf{F}|$ and is a centered first-order scheme. This upwind scheme is a conservative scheme which can be written in the following form

$$\mathbf{U}_j^{n+1} = \mathbf{U}_j^n - \frac{\Delta t}{\Delta x}\left(\mathbf{f}_{j+1/2}^n - \mathbf{f}_{j-1/2}^n\right) \tag{103}$$

where

$$\mathbf{f}_{j+1/2}^n = \frac{1}{2}\left(\mathbf{F}_{j+1}^n + \mathbf{F}_j^n\right) + \frac{1}{2}\left(|\mathbf{F}|_{j+1}^n - |\mathbf{F}|_j^n\right). \tag{104}$$

For the system of equations given by (**??**)

$$|\mathbf{G}| = \mathbf{P}|\mathbf{\Lambda}|\mathbf{P}^{-1} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$$

$$= \frac{1}{2c}\begin{bmatrix} |u - c/\sqrt{2}|\left(u\sqrt{2} + c\right) & \sqrt{2}|u + c/\sqrt{2}| - \\ -|u + c/\sqrt{2}|\left(u\sqrt{2} - c\right) & \sqrt{2}|u + c/\sqrt{2}| \\ -|u + c/\sqrt{2}|\left(u + c/\sqrt{2}\right)\left(u\sqrt{2} - c\right) & |u + c/\sqrt{2}|\left(u\sqrt{2} + c\right) \\ -|u - c/\sqrt{2}|\left(-u + c/\sqrt{2}\right)\left(u\sqrt{2} + c\right) & +|u - c/\sqrt{2}|\left(-u\sqrt{2} + c\right) \end{bmatrix}.$$

Results for this scheme are illustrated in Figure **??**(h). Similar results were obtained by Bermudez and Vazquez[6]. This scheme has also produced an entropy violating solution at the dam site.

The error is more dramatic than the results shown in Figure **??**(g). This would be due to the greater artificial diffusion introduced by the Steger and Warming scheme, than by the Bermudez and Vazquez scheme. There is a fix for the entropy violating solution. The modification involves adding a small value to the eigenvalues [12]. This is reported to provide a smooth transition at the sonic point at the expense of introducing a small error.

There is a fix for the entropy violating solution. One modification involves redefining the eigenvalues using[12]

$$\lambda^{\pm} = \frac{\lambda \pm \sqrt{\lambda^2 + \delta}}{2}$$

where δ is a small number. This provides a smooth transition at the expense of introducing a small error.

Equation (102) assumes that $\mathbf{F} = \mathbf{AU}$. This is valid if $\mathbf{F}(\mathbf{U})$ is a homogeneous function of degree one. The shallow water wave equations is a homogeneous function of order two, so that $\mathbf{F} \neq \mathbf{AU}$. The form of flux splitting used by Steger and Warming is not appropriate here (see, for example Hirsch[50], p. 417). It is possible to construct a matrix, $\mathbf{G}$ that does satisfy $\mathbf{F} = \mathbf{GU}$ (Bermudez and Vazquez[7]). For (**??**)

$$\mathbf{G} = \mathbf{P}\mathbf{\Lambda}\mathbf{P}^{-1} = \begin{bmatrix} 0 & 1 \\ -u^2 + gh/2 & 2u \end{bmatrix}, \ \mathbf{P} = \begin{bmatrix} 1 & 1 \\ u + c/\sqrt{2} & u - c/\sqrt{2} \end{bmatrix}$$

and

$$\mathbf{\Lambda} = \begin{bmatrix} u + c/\sqrt{2} & 0 \\ 0 & u - c/\sqrt{2} \end{bmatrix}$$

which are used in (111) to estimate $|\mathbf{F}| = |\mathbf{G}|\mathbf{U}$.

Results for this scheme are illustrated in Figure **??**(h). Similar results were obtained by Bermudez and Vazquez[6]. This scheme has also produced an entropy violating solution at the sonic point, that is, at the transition from subcritical to supercritical flows. The error is more dramatic than the results shown in Figure **??**(g). This is due to greater diffusion introduced by the Steger and Warming scheme, than by the Bermudez and Vazquez scheme. There is a fix for the entropy violating solution. The modification involves adding a small value to the eigenvalues (Buning and Steger[12]). This is reported to provide a smooth transition at the sonic point at the expense of introducing a small error.

Alternatively, the flux can be decomposed into two components such that (Yang *et al.*[133])

$$\mathbf{U}_t + \left(\mathbf{A}^+ + \mathbf{A}^-\right)\mathbf{F}_x = 0$$

where $\mathbf{A}^+ + \mathbf{A}^- = \mathbf{I}$. With the intercell flux given by

$$\mathbf{f}_{j+1/2} = \mathbf{A}^+_{j-1/2}\mathbf{F}_{j+1} + \mathbf{A}^-_{j+1/2}\mathbf{F}_j$$

where $\mathbf{A}^{\pm} = \mathbf{M}\mathbf{\Lambda}^{\pm}\mathbf{M}^{-1}$, $\mathbf{\Lambda}^{\pm} = \text{diag}(\lambda_i^{\pm})$ and $\lambda_i^{\pm} = (1 \pm \text{sgn}(\lambda_i))/2$ for $i = 1, 2$. This scheme has been used recently by Jin and Fread[59]. The solution of the test problems using this scheme is illustrated in Figure **??**(i). This scheme also produces an entropy violating solution.

8.2 Steger and Warming Scheme

In flux-vector splitting methods, the flux vector $\mathbf{F}$ is split according to the signs of the eigenvalues of $\mathbf{A}$. An example attributed to Steger and Warming[109] employed the *consistency* decomposition of the flux such that $\mathbf{F} = \mathbf{F}^+ + \mathbf{F}^-$. Therefore, (**??**) becomes

$$\frac{\partial \mathbf{U}}{\partial t} + \frac{\partial \mathbf{F}^+}{\partial x} + \frac{\partial \mathbf{F}^-}{\partial x} = 0 \tag{105}$$

where $\partial \mathbf{F}^+/\partial x$ is discretized using backward differencing and $\partial \mathbf{F}^-/\partial x$ is discretized using forward differencing so that

$$\mathbf{U}_j^{n+1} = \mathbf{U}_j^n - \frac{\Delta t}{\Delta x}\left[\left(\mathbf{F}_j^+ - \mathbf{F}_{j-1}^+\right)^n + \left(\mathbf{F}_{j+1}^- - \mathbf{F}_j^-\right)^n\right]. \tag{106}$$

where $\mathbf{F}^- = (\mathbf{F} - |\mathbf{F}|)/2$ and $\mathbf{F}^+ = (\mathbf{F} + |\mathbf{F}|)/2$. Substituting into (101) then

$$\mathbf{U}_j^{n+1} = \mathbf{U}_j^n - \frac{\Delta t}{2\Delta x}\left(\mathbf{F}_{j+1}^n - \mathbf{F}_{j-1}^n\right) + \frac{\Delta t}{2\Delta x}\left(|\mathbf{F}|_{j+1}^n - 2|\mathbf{F}|_j^n + |\mathbf{F}|_{j-1}^n\right). \tag{107}$$

Assuming $\mathbf{F} = \mathbf{A}\mathbf{U}$[109], then $|\mathbf{F}| = |\mathbf{A}|\mathbf{U}$, where $|\mathbf{A}| = \mathbf{M}|\mathbf{\Lambda}|\mathbf{M}^{-1}$,

$$\mathbf{M} = \frac{1}{2c}\begin{bmatrix} 1 & -1 \\ u+c & -u+c \end{bmatrix}, \quad \mathbf{M}^{-1} = \begin{bmatrix} -u+c & 1 \\ -u-c & 1 \end{bmatrix}$$

and

$$\mathbf{\Lambda} = \begin{bmatrix} u+c & 0 \\ 0 & u-c \end{bmatrix}$$

and the numerical flux is given by

$$\mathbf{f}_{j+1/2}^n = \frac{1}{2}\left(\mathbf{F}_{j+1}^n + \mathbf{F}_j^n\right) - \frac{1}{2}\left(|\mathbf{A}|_j^n \mathbf{U}_j^n - |\mathbf{A}|_{j+1}^n \mathbf{U}_{j+1}^n\right). \tag{108}$$

The results using this scheme are illustrated in Figure **??**(g). At the original location of the dam, $x = 1,000m$ there is a discontinuity in the simulated water depth, h and velocity u. At this shock, which has speed S entropy $\lambda_l > S > \lambda_r$ is violated, where λ_l is the speed of the characteristics to the left of the shock and λ_r is the speed of the characteristics to the right of the shock. Shocks are only formed when characteristics converge whereas, in this case they diverge. Here, a *expansive shock* has formed which is non-physical. This is a common problem with upwind schemes and is due to the inability of upwind schemes in defining the upwind direction at a *sonic point*[51]. That is when *critical flow*, $\lambda = 0$m/s occurs and there is a change from subcritical to supercritical flow or *visa versa*. At a sonic point the eigenvalues of the Jacobian vanish. In the example where $h_0 = 5$m flow is always subcritical, sonic points are not encountered and the solution is continuous along the rarefaction fan. In the second example, where $h_0 = 0.1$m a transition between subcritical and supercritical flow occurs. In this case the sonic point occurs at the original location of the dam.

The flux given by (108) can be considered as simple average flux at the cell interface with the addition of a diffusion term such that

$$\mathbf{f}^n_{j+1/2} = \frac{1}{2}\left(\mathbf{F}^n_{j+1} + \mathbf{F}^n_j\right) - \frac{1}{2}\mathbf{D}(\mathbf{U}_{j+1}, \mathbf{U}_j)$$

where

$$\mathbf{D}(\mathbf{U}_{j+1}, \mathbf{U}_j) = |\mathbf{A}|^n_j \mathbf{U}^n_j - |\mathbf{A}|^n_{j+1}\mathbf{U}^n_{j+1}.$$

Now for a stationary shock, $\mathbf{F}_{j+1} = \mathbf{F}_j$, however this does not imply $\mathbf{F}^+_{j+1} - \mathbf{F}^-_j = \mathbf{F}^+_j - \mathbf{F}^-_j$, roughly because $\mathbf{U}_j = -\mathbf{U}_{j+1}$ does not imply $|\mathbf{U}_j|\mathbf{U}_j = |\mathbf{U}_{j+1}|\mathbf{U}_{j+1}$. For nearby values of $\mathbf{U}_j$ and $\mathbf{U}_{j+1}$, $\mathbf{U}_j = -\mathbf{U}_{j+1}$ implies that $\mathbf{U}_j$ and $\mathbf{U}_{j+1}$ are near 0. It follows that a flux-vector splitting scheme does not give a good resolution near sonic points such that $a_k = 0$, since there may exist nearby states $\mathbf{U}_j$, $\mathbf{U}_{j+1}$ such that $\mathrm{sgn}a_k(\mathbf{U}_j) \neq \mathrm{sgn}a_k(\mathbf{U}_{j+1})$. This happens for stationary contact discontinuities ion the case of Euler equations and obviously the shallow water wave equations.

Also, after rearrangement, we can write Steger and Warming's viscosity term as

$$\mathbf{D}(\mathbf{U}_{j+1}, \mathbf{U}_j) = \frac{1}{2}(|\mathbf{A}|^n_j - |\mathbf{A}|^n_{j+1})(\mathbf{U}_{j+1} - \mathbf{U}_j) + \frac{1}{2}(|\mathbf{A}|_{j+1} - |\mathbf{A}|_j)(\mathbf{U}_j + \mathbf{U}_{j+1}).$$

A scheme satisfying

$$\mathbf{D}(\mathbf{U}_{j+1}, \mathbf{U}_j) = |\mathbf{A}\left(\frac{\mathbf{U}_j + \mathbf{U}_{j+1}}{2}\right)|(\mathbf{U}_{j+1} - \mathbf{U}_j) + O(|\mathbf{U}_j - \mathbf{U}_{j+1}|).$$

is called an upstream scheme by Harten, Lax and van Leer[45].

Thus, the above scheme is of upstream form except near sonic points.

Recalling that, for any $\alpha \in \mathcal{R}$

$$\mathbf{A}(\alpha\mathbf{U}) = \mathbf{A}(\mathbf{U})$$

and

$$\mathbf{A}^{\pm}(\alpha\mathbf{U}) = \mathbf{A}\pm(\mathbf{U}) \quad \text{and} \quad \mathbf{F}^{\pm}(\alpha\mathbf{U}) = \alpha\mathbf{F}^{\pm}(\mathbf{U}).$$

Using Euler's identity,

this implies

$$\mathbf{F}^{\pm}(\mathbf{U}) = \frac{\partial\mathbf{F}^{\pm}(\mathbf{U})}{\partial\mathbf{U}} \cdot \mathbf{U} \equiv \mathbf{A}_{\pm}(\mathbf{U})\mathbf{U}.$$

However, $\mathbf{A}_+(\mathbf{U}) = \partial\mathbf{F}^+(\mathbf{U})/\partial\mathbf{U}$ is not equal to $\mathbf{A}^+(\mathbf{U})$, though $\mathbf{A}_+(\mathbf{U})\mathbf{U} = \mathbf{A}^+(\mathbf{U})\mathbf{U}$. Also, the notations may be misleading since $\mathbf{A}_+(\mathbf{U})$ and $\mathbf{A}_-(\mathbf{U})$ do not necessarily have positive or negative eigenvalues, and one will have to check their sign.

Steger and Warming's decomposition is also not continuous in general at a sonic point[77].

8.3 Yang et al.

Alternatively, the flux can be decomposed into two components such that[133]

$$\frac{\partial \mathbf{U}}{\partial t} + \left(\mathbf{A}^+ + \mathbf{A}^-\right)\frac{\partial \mathbf{F}}{\partial x} = 0$$

where $\mathbf{A}^+ + \mathbf{A}^- = \mathbf{I}$. Discretizing

$$\mathbf{U}_j^{n+1} = \mathbf{U}_j^n - \frac{\Delta t}{\Delta x}\left(\mathbf{A}^+\right)_{j-1/2}^n\left(\mathbf{F}_j^n - \mathbf{F}_{j-1}^n\right) - \frac{\Delta t}{\Delta x}\left(\mathbf{A}^-\right)_{j+1/2}^n\left(\mathbf{F}_{j+1}^n - \mathbf{F}_j^n\right)$$

where $\mathbf{A}^\pm = \mathbf{M}\mathbf{\Lambda}^\pm\mathbf{M}^{-1}$, $\mathbf{\Lambda}^\pm = \mathrm{diag}(\lambda_i^\pm)$ and $\lambda_i^\pm = (1 \pm \mathrm{sgn}(\lambda_i))/2$ for $i = 1, 2$.

The solution of the test problems using this scheme is illustrated in Figure **??**(i). This scheme also produces an entropy violating solution. Alternatively, the flux can be decomposed into two components such that[133]

$$\frac{\partial \mathbf{U}}{\partial t} + \left(\mathbf{A}^+ + \mathbf{A}^-\right)\frac{\partial \mathbf{F}}{\partial x} + \mathbf{S} = 0$$

where $\mathbf{A}^+ + \mathbf{A}^- = \mathbf{I}$. Expanding

$$\frac{\partial \mathbf{U}}{\partial t} + \mathbf{A}^+\frac{\partial \mathbf{F}}{\partial x} + \mathbf{A}^-\frac{\partial \mathbf{F}}{\partial x} + \mathbf{S} = 0.$$

Discretizing

$$\begin{aligned}\mathbf{U}_j^{n+1} = \mathbf{U}_j^n &- \frac{\Delta t}{\Delta x}\left(\mathbf{A}^+\right)_{j-1/2}^n\left(\mathbf{F}_j^n - \mathbf{F}_{j-1}^n\right) \\ &- \frac{\Delta t}{\Delta x}\left(\mathbf{A}^-\right)_{j+1/2}^n\left(\mathbf{F}_{j+1}^n - \mathbf{F}_j^n\right) - \Delta t\mathbf{S}_j^n\end{aligned}$$

where

$$\begin{aligned}\mathbf{A}^+\left(\mathbf{U}_{j-1/2}^n\right) &= \mathbf{A}^+\left(\frac{\mathbf{U}_j^n + \mathbf{U}_{j-1}^n}{2}\right), \\ \mathbf{A}^-\left(\mathbf{U}_{j+1/2}^n\right) &= \mathbf{A}^-\left(\frac{\mathbf{U}_{j+1}^n + \mathbf{U}_j^n}{2}\right) \quad \text{and} \\ \mathbf{A}^+\left(\mathbf{U}_{j-1/2}^n\right) + \mathbf{A}^-\left(\mathbf{U}_{j-1/2}^n\right) &= \mathbf{I}.\end{aligned}$$

In terms of the fluxes

$$\mathbf{f}_{j+1/2}^n = \left(\mathbf{A}^-\right)_{j+1/2}^n\left(\mathbf{F}_{j+1}^n - \mathbf{F}_j^n\right). \tag{109}$$

The matrix $\mathbf{A}^\pm = \mathbf{M}\mathbf{\Lambda}^\pm\mathbf{M}^{-1}$ where $\mathbf{\Lambda}^\pm = \mathrm{diag}(\lambda_i^\pm)$ and $\lambda_i^\pm = (1 \pm \mathrm{sgn}(\lambda_i))/2$ for $i = 1, 2$. Since

$$\mathbf{M} = \frac{1}{2c}\begin{bmatrix} 1 & -1 \\ u+c & -u+c \end{bmatrix}$$

,

$$\mathbf{M}^{-1} = \begin{bmatrix} -u+c & 1 \\ -u-c & 1 \end{bmatrix} \quad \text{and}$$

$$\mathbf{\Lambda}^{\pm} = \begin{bmatrix} \lambda_1^{\pm} & 0 \\ 0 & \lambda_2^{\pm} \end{bmatrix}.$$

Expanding

$$\begin{aligned} \mathbf{A}^{+} &= \frac{1}{2c} \begin{bmatrix} -u\left(\lambda_1^{+} - \lambda_2^{+}\right) + c\left(\lambda_1^{+} + \lambda_2^{+}\right) & \lambda_1^{+} - \lambda_2^{+} \\ \left(c^2 - u^2\right)\left(\lambda_1^{+} - \lambda_2^{+}\right) & u\left(\lambda_1^{+} - \lambda_2^{+}\right) + c\left(\lambda_1^{+} + \lambda_2^{+}\right) \end{bmatrix} \\ &= \begin{bmatrix} a_{11}^{+} & a_{12}^{+} \\ a_{21}^{+} & a_{22}^{+} \end{bmatrix} \end{aligned}$$

and

$$\begin{aligned} \mathbf{A}^{-} &= \frac{1}{2c} \begin{bmatrix} -u\left(\lambda_1^{-} - \lambda_2^{-}\right) + c\left(\lambda_1^{-} + \lambda_2^{-}\right) & \lambda_1^{-} - \lambda_2^{-} \\ \left(c^2 - u^2\right)\left(\lambda_1^{-} - \lambda_2^{-}\right) & u\left(\lambda_1^{-} - \lambda_2^{-}\right) + c\left(\lambda_1^{-} + \lambda_2^{-}\right) \end{bmatrix} \\ &= \begin{bmatrix} a_{11}^{-} & a_{12}^{-} \\ a_{21}^{-} & a_{22}^{-} \end{bmatrix}. \end{aligned}$$

For supercritical flows, $\lambda_1^{+} = \lambda_2^{+} = 1$, and $\lambda_1^{-} = \lambda_2^{-} = 0$ then

$$\mathbf{A}^{+} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad \text{and} \quad \mathbf{A}^{-} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

and (95) is being solved. For supercritical flows in the negative x-direction , $\lambda_1^{+} = \lambda_2^{+} = 0$, and $\lambda_1^{-} = \lambda_2^{-} = 0$ then

$$\mathbf{A}^{+} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad \mathbf{A}^{-} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

and (96) is being solved. For subcritical flow, $\lambda_1^{+} = \lambda_2^{-} = 1$, and $\lambda_1^{-} = \lambda_2^{+} = 0$ then

$$\begin{aligned} \mathbf{A}^{+} &= \frac{1}{2c} \begin{bmatrix} c - u & 1 \\ c^2 - u^2 & 1 + u \end{bmatrix} \quad \text{and} \\ \mathbf{A}^{-} &= \frac{1}{2c} \begin{bmatrix} u + c & -1 \\ -c^2 + u^2 & -u + 1 \end{bmatrix}. \end{aligned}$$

All these matrices satisfy the condition $\mathbf{A}^{+} + \mathbf{A}^{-} = \mathbf{I}$.

8.4 Harten Scheme

Another first-order scheme which attempts to combine (95) and (96) is described by Harten[42].

Since the Jacobian $\mathbf{A}(\mathbf{U}) = \partial\mathbf{F}/\partial\mathbf{U}$, from the *mean-value theorem*[110] [9] it follows that

$$\frac{\mathbf{F}(\mathbf{U}_{j-1}^{n}) - \mathbf{F}(\mathbf{U}_{j}^{n})}{\mathbf{U}_{j-1}^{n} - \mathbf{U}_{j}^{n}} = \mathbf{A}(\overline{\mathbf{U}}), \quad \mathbf{U}_{j-1}^{n} < \overline{\mathbf{U}}_{j}^{n} < \mathbf{U}_{j}^{n}$$

[9] Mean Value Theorem:
If $f(x)$ is a continuous function on the closed interval $[a, b]$ and is at least C^1 in the interval, then there is at least one point c, $a \le c \le b$ where $\frac{df}{dc} = \frac{f(b) - f(a)}{b - a}$.

satisfies the Rankine-Hugoniot condition, is used to establish whether the characteristics emanate from the left or the right. This is essentially the linearization of Jacobian so that $\Delta\mathbf{F} = \mathbf{A}\Delta\mathbf{U}$. Defining an indicator function

$$\delta(\mathbf{V},\mathbf{W}) = \begin{cases} 1 & \dfrac{\mathbf{F}(\mathbf{V})-\mathbf{F}(\mathbf{W})}{\mathbf{V}-\mathbf{W}} \geq 0 \\ 0 & \dfrac{\mathbf{F}(\mathbf{V})-\mathbf{F}(\mathbf{W})}{\mathbf{V}-\mathbf{W}} < 0 \end{cases}$$

then the fluxes in (103) can be expressed as

$$\mathbf{f}^n_{j-1/2} = \mathbf{F}^n_{j-1} + \left[1-\delta^n_{j-1/2}\right]\left(\mathbf{F}^n_j - \mathbf{F}^n_{j-1}\right) \tag{110}$$

which is in conservative form. Harten upwind scheme becomes

$$\begin{aligned}\mathbf{U}^{n+1}_j = \mathbf{U}^n_j + \frac{\Delta t}{\Delta x}\Big\{&\mathbf{F}^n_{j-1} + \left[1-\delta^n_{j-1/2}\right]\left(\mathbf{F}^n_j - \mathbf{F}^n_{j-1}\right) \\ &-\mathbf{F}^n_j - \left[1-\delta^n_{j+1/2}\right]\left(\mathbf{F}^n_{j+1} - \mathbf{F}^n_j\right)\Big\}\end{aligned}$$

A great deal is known about the mathematical structure of the governing equations and their solution (Stoker[112]). One family of schemes which attempt to exploit these properties are *upwind* schemes. In these schemes, space discretizations are consistent with the propagation properties of the problem. The space differencing depends on the sign of the eigenvalues. A number of first-order *flux-vector splitting* upwind schemes have been considered. These are by Steger and Warming[109]), Bermudez and Vazquez[7]), Yang *et al.*[133]) and Harten[42].

In the flux-vector splitting method attributed to Steger and Warming[109], $\mathbf{F} = \mathbf{F}^+ + \mathbf{F}^-$. Then the governing equation becomes $\mathbf{U}_t + (\mathbf{F}^+)_x + (\mathbf{F}^-)_x = 0$. To explicitly consider the direction of the flow, $(\mathbf{F}^+)_x$ is discretized using backward differencing and $(\mathbf{F}^-)_x$ is discretized using forward differencing so that $\mathbf{F}^- = (\mathbf{F} - |\mathbf{F}|)/2$ and $\mathbf{F}^+ = (\mathbf{F} + |\mathbf{F}|)/2$. The intercell flux becomes

$$\mathbf{f}_{j+1/2} = \frac{1}{2}\left(\mathbf{F}_{j+1} + \mathbf{F}_j\right) + \frac{1}{2}\left(|\mathbf{F}|_{j+1} - |\mathbf{F}|^n_j\right). \tag{111}$$

Assuming that $\mathbf{F} = \mathbf{A}\mathbf{U}$ (Steger and Warming[109]), then $|\mathbf{F}| = |\mathbf{A}|\mathbf{U}$, where $|\mathbf{A}| = \mathbf{M}|\mathbf{\Lambda}|\mathbf{M}^{-1}$ with

$$\mathbf{M} = \frac{1}{2c}\begin{bmatrix} 1 & -1 \\ u+c & -u+c \end{bmatrix} \quad \text{and} \quad \mathbf{\Lambda} = \begin{bmatrix} u+c & 0 \\ 0 & u-c \end{bmatrix}.$$

The results using this scheme are illustrated in Figure **??**(g) and are typical of those produced by first-order upwind schemes, producing an entropy violating solution at the sonic point. This is due to the inability of upwind schemes in defining the upwind direction. At a sonic point the eigenvalues of theJacobian vanish.

9 Godunov-Type Schemes

Godunov[40] proposed a numerical method for gas dynamics where the exact solution to the Riemann initial-value problem on breaking up of an initial discontinuity.

Since then many numerical schemes that incorporate this solution, or an approximation of it followed in other areas of computational fluid mechanics, which are now referred to as *Godunov-type schemes*[45]. An alternative approach to solving the shallow water wave equations for problems containing discontinuous solutions, is by solving the *Riemann* problem in a *Godunov-type scheme*. In Godunov-type schemes, the numerical solution is considered piecewise constant in each mesh cell $\left[x_{j-1/2}, x_{j+1/2}\right]$. The cell interface at $x_{j+1/2}$ separates two constant states $\mathbf{U}_l$ and $\mathbf{U}_r$ which form the Riemann problem. The Riemann problem for a general hyperbolic system is the following initial-value problem

$$\frac{\partial \mathbf{U}}{\partial t} + \frac{\partial \mathbf{F}(\mathbf{U})}{\partial x} = 0 \tag{112}$$

with

$$\mathbf{U}(x,0) = \begin{cases} \mathbf{U}_l & x < 0 \\ \\ \mathbf{U}_r & x \geq 0. \end{cases} \tag{113}$$

The steps involved in the Godunov scheme given a set of initial conditions, shown in Figure 49

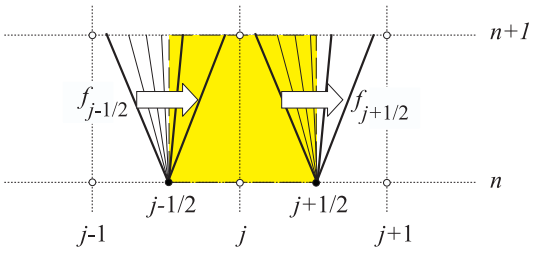


Figure 48: Waves emanating from cell interface in Godunov's method.

are;

- *averaging* or *reconstruction step*,
- *solving the local Riemann problem* or *evolution step* and
- *re-averaging* or *cell-averaging step* to obtain the nodal dependent variables.

9.1 Averaging

In the traditional Godunov scheme, the numerical solution is considered piecewise constant in each mesh cell $\left[x_{j-1/2}, x_{j+1/2}\right]$. The cell interface at $x_{j+1/2}$ separates two constant states $\mathbf{U}_l$ and $\mathbf{U}_r$ which form the Riemann problem. For each cell interface $x_{j+1/2}$, $\mathbf{U}_l$ and $\mathbf{U}_r$ are estimated from

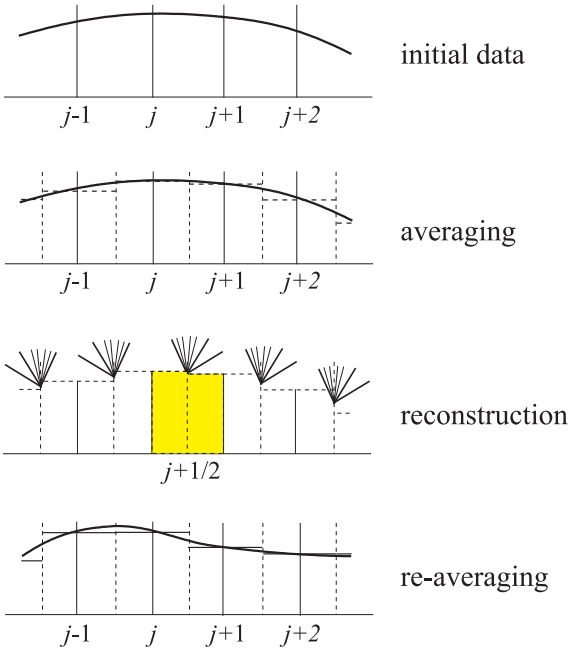


Figure 49: Steps in Godunov's scheme.

the reconstruction polynomial $\mathbf{U}_j(x) = \mathbf{P}_j(x)$. The polynomial used to estimate the states $\mathbf{U}_l$ and $\mathbf{U}_r$ determine the accuracy of the scheme. For Godunov's first-order method, the profile is approximated by piecewise constant $\mathbf{P}_j(x) = \mathbf{U}_j$, therefore $\mathbf{U}_l = \mathbf{U}_j$ and $\mathbf{U}_r = \mathbf{U}_{j+1}$. The intercell flux $\mathbf{f}_{j+1/2}$ is calculated using an exact or approximate solution to the Riemann problem given the initial states $\mathbf{U}_l$ and $\mathbf{U}_r$.

The intercell flux is then used to update the conserved quantities at the nodes. A second-order scheme is produced if linear interpolation $\mathbf{P}_j(x) = \mathbf{U}_j + \mathbf{s}_j(x - x_j)$ is used to estimate the left and right states. Other well-known schemes using higher-order reconstruction of the solution are van Leer's MUSCL, third-order accuracy *Piecewise Parabolic Method* (*PPM*) Collela and Woodward[19] and arbitrary high-order *Essentially Non-Oscillatory* (*ENO*) (Harten[43]) schemes. $\mathbf{U}_l$ and $\mathbf{U}_r$ at the cell interface are limited in all the higher-order schemes to avoid oscillations in the solution near discontinuities. With these values, the Riemann problem is solved using an exact or approximate first-order Riemann solver. The gain in accuracy is achieved by using higher-order averaging to estimate the left and right states in the Riemann problem.

9.2 Re-averaging

The re-averaging process involves solving the discrete explicit approximation of the integral form of the governing equation which leads to

$$\mathbf{U}_j^{n+1} = \mathbf{U}_j^n - \frac{\Delta t}{\Delta x}(\mathbf{f}_{j+1/2}^n - \mathbf{f}_{j-1/2}^n). \tag{114}$$

The intercell flux $\mathbf{f}_{j+1/2}$ is calculated using an exact or approximate solution to the Riemann problem given the initial states $\mathbf{U}_l$ and $\mathbf{U}_r$. The intercell flux at $x = (j+1/2)\Delta x$ is used to update the conserved quantities at the nodes.

Since we are interested in problems with discontinuities, shock waves and contact surfaces it is more appropriate to use the integral form of (**??**)

$$\oint (\mathbf{U}dx - \mathbf{F}dt) = 0 \tag{115}$$

which accepts discontinuities in the dependent variables. Evaluating the integral in the control volume ABCD of dimensions Δx and Δt shown in Figure 50 produces the scheme

$$\mathbf{U}_i^{n+1} = \mathbf{U}_i^n - \frac{\Delta t}{\Delta x}\left(\mathbf{f}_{i+1/2}^n - \mathbf{f}_{i-1/2}^n\right) \tag{116}$$

where $\mathbf{U}_i^n$ denotes the discrete value of $\mathbf{U}$ in cell i at time level n and $\mathbf{E}_{i+1/2}^n$ is he intercell flux. Godunov-type methods utilize the solution of the local Riemann problem with data, $\mathbf{U}_l = \mathbf{U}_i^n$ and $\mathbf{U}_r = \mathbf{U}_{i+1}^n$ to define numerically the intercell flux, $\mathbf{E}_{i+1/2}^n$.

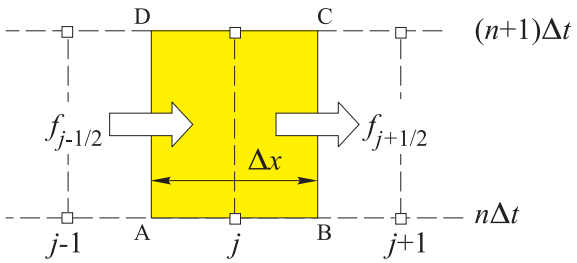


Figure 50: Control volume for cell j and the intercell fluxes.

The Riemann problem for (**??**) is the initial value problem in the domain -$\infty < x < \infty$, $t > 0$ with the data

$$\mathbf{U}(x,0) = \begin{cases} \mathbf{U}_l & x \le 0 \\ \\ \mathbf{U}_r & x > 0. \end{cases}$$

For the shallow water wave equations, (**??**) the Riemann problem can be decomposed into three waves, see Figure **??**. The middle wave is always a contact discontinuity and the left and right waves can be either shocks or rarefaction fan. Therefore, there are only four possible wave combinations involving discontinuous contact and shocks waves and rarefaction fan which are continuous. In the region between these waves h and u are always constant but there is a discontinuity in v across the contact discontinuity.

The first-order Godunov method uses the state along the t-axis ($x = 0$) to define the intercell flux, $\mathbf{E}_{i+1/2}^n$ for the solution of the Riemann problem with data, $\mathbf{U}_i^n$ and $\mathbf{U}_{i+1}^n$.

9.3 Evolution Step

The evolution step involves the use of an exact or approximate Riemann solver to calculate the inter-cell flux.

Due to the complexity of the exact Riemann solver (see section 9.4), there have been many attempts to develop approximate Riemann solvers. These have been found to be very efficient and accurate. Examples of first-order approximate Riemann solvers are; Roe's scheme[101], utilized by Glaister[38] and Jha *et al.*[56] for the shallow water wave equations, Osher's P-ordering scheme[29][93], which has been used by Zhao *et al.*[138][137] to solve the shallow water wave equations and the *Harten, Lax and van Leer* (*HLL*) scheme[45], which has been modified by Fraccarollo and Torro[33], which has recently been used by Mingham and Causon[82] and Zoppou and Roberts[139]. Examples of second and higher-order schemes are the *Weighted Average Flux* (*WAF*) scheme[33], which has been used by Zoppou and Roberts[140] and the third-order *Piecewise Parabolic Method* [19], which has been used by Savic and Holly[103].

The Riemann problem has uniform initial conditions on an infinite spatial domain, except for a single jump discontinuity. The Riemann problem centered on $x = x_0$ and $t = t_0$ has the following initial conditions

$$\mathbf{U}(x, t_0) = \begin{cases} \mathbf{U}_l & x < x_0 \\ \\ \mathbf{U}_r & x > x_0 \end{cases}.$$

For convenience, $x_0 = 0$ and $t_0 = 0$.

Numerical methods using Riemann solvers require only the flux along $x = 0$. The Godunov intercell numerical flux

$$\mathbf{f}_{j+1/2} = \mathbf{f}(\mathbf{U}_{j+1/2}(0))$$

in which $\mathbf{U}_{j+1/2}(0)$ is the exact similarity solution $\mathbf{U}_{j+1/2}(x/t)$ of the Riemann problem

$$\frac{\partial \mathbf{U}}{\partial t} + \frac{\partial \mathbf{F}(\mathbf{U})}{\partial x} = 0 \tag{117}$$

with

$$\mathbf{U}(x, 0) = \begin{cases} \mathbf{U}_l & x < 0 \\ \mathbf{U}_r & x > 0 \end{cases}$$

evaluated at $x/t = 0$.

Godunov scheme first evolves the piecewise constant solution by solving a successive sequence of non-interacting Riemann problems. The solution at time level $t + \Delta t$, can be expressed in terms of the Riemann solver, $R(x, t, u_l, u_r)$ so that

$$u(x, t + \Delta t) = R\left(\frac{x - x_{j+1/2}}{\Delta t}; u_j(t), u_{j+1}(t)\right), \quad x_j \leq x \leq x_{j+1}. \tag{118}$$

This solution is then projected back into the space of piecewise constant grid functions, see Figure **??**(a) so that

$$u_j(t+\Delta t) = \frac{1}{\Delta x}\int_{x_{j-1/2}}^{x_{j+1/2}} u(\xi, t+\Delta t)\, d\xi, \quad x_{j-1/2} \le x \le x_{j+1/2}. \tag{119}$$

Integration of the governing one-dimensional scalar equation

$$\frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x} = 0 \tag{120}$$

over a typical cell $[x_{j-1/2}, x_{j+1/2}] \times [t, t+\Delta t]$ yields

$$u_j(t+\Delta t) = u_j(t) + \frac{\Delta t}{\Delta x}\left(f(R(0^+; u_{j-1}(t), u_j(t)) - f(R(0^-; u_j(t), u_{j+1}(t)))\right). \tag{121}$$

This shows the upwind property of the Godunov scheme. Namely, if the characteristics speeds throughout the relevant neighbouring cells, $[x_{j-1}, x_{j+1/2}]$ are all positive then (121) becomes

$$u_j(t+\Delta t) = u_j(t) - \frac{\Delta t}{\Delta x}(f(u_j(t)) - f(u_{j-1}(t))$$

alternatively if the characteristic speeds are negative then

$$u_j(t+\Delta t) = u_j(t) - \frac{\Delta t}{\Delta x}(f(u_{j+1}(t)) - f(u_j(t)).$$

More complex situations occur when there is a mixture of both right-going and left-going waves. In this case, the computation of Godunov's numerical flux in (121) requires use to distinguish between the left and right-going waves inside the Riemann problem.

Roe (????) solved the Riemann problem (117) approximately by introducing the Jacobian matrix

$$\mathbf{A}(\mathbf{U}) = \frac{\partial \mathbf{F}}{\partial \mathbf{U}}$$

and using the *chain rule* so that (117) can be written as

$$\frac{\partial \mathbf{U}}{\partial t} + \mathbf{A}(\mathbf{U})\frac{\partial \mathbf{U}}{\partial x} = 0.$$

Roe's approach replaces the Jacobian matrix $\mathbf{A}(\mathbf{U})$ by a constant Jacobian matrix

$$\widetilde{\mathbf{A}} = \widetilde{\mathbf{A}}(\mathbf{U}_l, \mathbf{U}_r)$$

so that

$$\frac{\partial \mathbf{U}}{\partial t} + \widetilde{\mathbf{A}}\frac{\partial \mathbf{U}}{\partial x} = 0$$

is solved with

$$\mathbf{U}(x,0) = \begin{cases} \mathbf{U}_l & x < 0 \\ \\ \mathbf{U}_r & x > 0 \end{cases}.$$

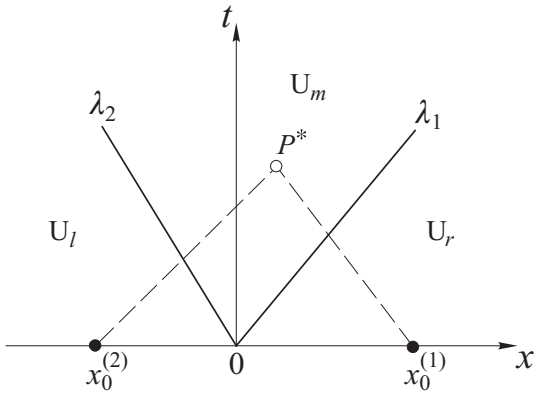


Figure 51: Structure of the solution of the Riemann problem for a 2×2 linear system of first-order equations with constant coefficients.

The approximate problem results from replacing the original non-linear conservation laws by a linearized system with constant coefficients but the initial data is retained.

Recalling from Section 2???? the Riemann problem for a system of m linear first-order hyperbolic equations can be written as m independent Riemann problems for the linear advection problem

$$\frac{\partial R_i}{\partial t} + \lambda_i \frac{\partial R_i}{\partial x} = 0$$

where

$$R_i(x,0) = \begin{cases} R_{l_i} = \mathbf{I}_i \cdot \mathbf{u}_l & x < 0 \\ \\ R_{r_i} = \mathbf{I}_i \cdot \mathbf{u}_r & x > 0 \end{cases}$$

in which $\mathbf{I}_i$ are the *left eigenvectors* of $\mathbf{A}$. Using superposition, the Riemann problem for a system of two first-order equations is shown in Figure 51 where λ_i separate regions of constant state. Returning to the conservative variables $\mathbf{U} = \mathbf{MR}$ which can be written as

$$\begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix} = \begin{bmatrix} M_{1,1} & M_{1,2} & \cdots & M_{1,m} \\ M_{2,1} & M_{2,2} & \cdots & M_{2,m} \\ \vdots & & & \vdots \\ M_{m,1} & M_{m,2} & \cdots & M_{m,m} \end{bmatrix} \begin{bmatrix} R_1 \\ R_2 \\ \vdots \\ R_m \end{bmatrix}$$

or

$$\begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix} = R_1 \begin{bmatrix} M_{1,1} \\ M_{2,1} \\ \vdots \\ M_{m,1} \end{bmatrix} + R_2 \begin{bmatrix} M_{1,2} \\ M_{2,2} \\ \vdots \\ M_{m,2} \end{bmatrix} + \cdots + R_m \begin{bmatrix} M_{1,m} \\ M_{2,m} \\ \vdots \\ M_{m,m} \end{bmatrix}$$

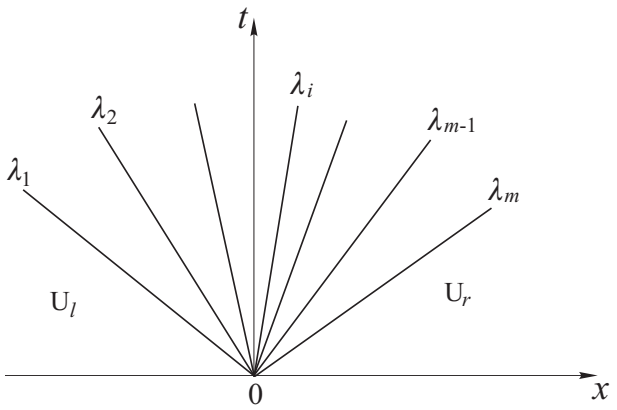


Figure 52: Structure of the solution of the Riemann problem for a $m \times m$ linear system of first-order equations with constant coefficients.

or more succinctly

$$\mathbf{U}(x,t) = \sum_{i=1}^{m} R_i(x,t)\mathbf{M}^{(i)}$$

where $\mathbf{M}^{(i)}$ is the ith column vector of $\mathbf{M}$.[10]

This means that the function $R_i(x,t)$ is the coefficient of $\mathbf{M}^{(i)}$ is an eigenvector expansion of the vector $\mathbf{U}$. But according to (**??**), $R_i(x,t) = R_i^{(0)}(x - \lambda_i t)$ then

$$\mathbf{U}(x,t) = \sum_{i=1}^{m} R_i^{(0)}(x - \lambda_i)\mathbf{M}^{(i)}. \tag{122}$$

Therefore, given a point (x,t) in the $x-t$ plane, the solution $\mathbf{U}(x,t)$ at this point depends only on the initial data at the m points $x_0^{(i)} = x - \lambda_i$. These are the intersections of the characteristics of speed λ_i with the x-axis. The solution (122) for $\mathbf{U}$ can be seen as the *superposition* of m waves, each of which is *advected independently without change in shape*. The ith wave has shape $R_i^{(0)}(x)\mathbf{M}^{(i)}$ and propagates with speed λ_i.

[10] Eigenvalues:

The eigenvalues λ_i of a matrix $\mathbf{A}$ are the solution of the characteristic polynomial

$$|\mathbf{A} - \lambda\mathbf{I}| = \det(\mathbf{A} - \lambda\mathbf{I}) = 0$$

where $\mathbf{I}$ is the identity matrix. The eigenvalues of the coefficient matrix $\mathbf{A}$ of a system of the form (???) are called the eigenvalues of the system. For this system the eigenvalues represent speeds of propagation of information. A right eigenvector of an $n \times n$ matrix $\mathbf{A}$ corresponding to an eigenvalue λ_i of $\mathbf{A}$ is a vector $\mathbf{K}^{(i)} = [K_1^{(i)}, K_2^{(i)}, \cdots, K_n^{(i)}]^T$ satisfying $\mathbf{A}\mathbf{K}^{(i)} = \lambda_i\mathbf{K}^{(i)}$, DEFINE superscript. Similarly, a left eigenvector of a matrix $\mathbf{A}$ corresponds to an eigenvalue λ_i of $\mathbf{A}$ is a vector $\mathbf{L}^{(i)} = [L_1^{(i)}, L_1^{(i)}, \cdots, L_1^{(i)}]$ such that $\mathbf{L}^{(i)}\mathbf{A} = \lambda_i\mathbf{L}^{(i)}$.

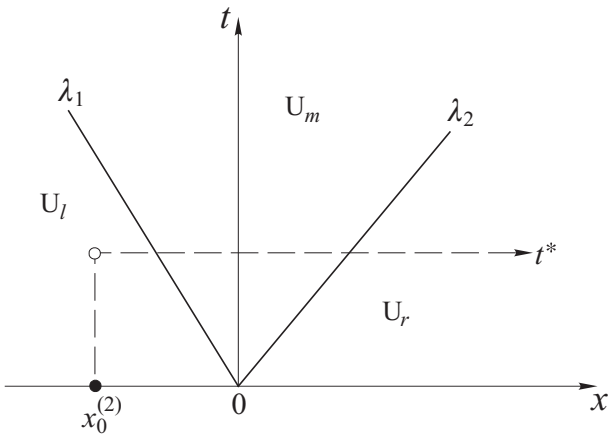


Figure 53: Solution of the Riemann problem for a 2×2 linear system of first-order equations with constant coefficients found by travelling along the line of constant time, $t = t^*$.

Determine the intermediate state for the linearised equations of Gas Dynamics

$$\frac{\partial \mathbf{U}}{\partial t} + \mathbf{A}\frac{\partial \mathbf{U}}{\partial x} = 0$$

with

$$\mathbf{U} = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \equiv \begin{bmatrix} \rho \\ u \end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix} 0 & \rho_0 \\ a^2/\rho_0 & 0 \end{bmatrix}.$$

by solving the Riemann problem. The eigenvalues of the system are

$$\lambda_1 = -a, \quad \lambda_2 = a$$

and the corresponding right eigenvectors are

$$\mathbf{M}^{(1)} = \begin{bmatrix} \rho_0 \\ -1 \end{bmatrix}, \quad \mathbf{M}^{(2)} = \begin{bmatrix} \rho_0 \\ a \end{bmatrix}.$$

First we decompose the left data state $\mathbf{U}_l = [\rho_l, u_l]^T$ in terms of the right eigenvectors according to (133), namely

$$\mathbf{U}_l = \begin{bmatrix} \rho_l \\ u_l \end{bmatrix} = \alpha_1 \begin{bmatrix} \rho_0 \\ -a \end{bmatrix} + \alpha_2 \begin{bmatrix} \rho_0 \\ a \end{bmatrix}.$$

Solving for the unknown coefficients α_1 and α_2 we obtain

$$\alpha_1 = \frac{a\rho_l - \rho_0 u_l}{2a\rho_0} \quad \text{and} \quad \alpha_2 = \frac{a\rho_l + \rho_0 u_l}{2a\rho_0}.$$

Similarly, by expanding the right-hand data $\mathbf{U}_r = [\rho_r, u_r]^T$ in terms of the eigenvectors and solving for the coefficients β_1 and β_2 we obtain

$$\beta_1 = \frac{a\rho_r - \rho_0 u_r}{2a\rho_0} \quad \text{and} \quad \beta_2 = \frac{a\rho_r + \rho_0 u_r}{2a\rho_0}.$$

Now by using (136) we find the solution in the *intermediate region* as

$$\mathbf{U}^* = \begin{bmatrix} \rho^* \\ u^* \end{bmatrix} = \beta_1 \begin{bmatrix} \rho_0 \\ -a \end{bmatrix} + \beta_2 \begin{bmatrix} \rho_0 \\ a \end{bmatrix}.$$

After some algebraic manipulations we obtain the solution explicitly as

$$\rho^* = (\rho_l + \rho_r)/2 - (u_r - u_l)\rho_0/(2a) \quad \text{and} \tag{123}$$

$$u^* = (u_l + u_r)/2 - (\rho_r - \rho_l)a/(2\rho_0). \tag{124}$$

9.4 Exact Riemann Solver

In the exact Riemann solver use is made of the fact that all the possible waves for the shallow water wave equation that can be connected to state $\mathbf{U}_l$ are shown in Figure 54. The region surrounding $\mathbf{U}_l$ is divided into four disjoint open regions; I, II, III and IV. Depending on the status of the adjoining state, $\mathbf{U}_r$ it can be connected to state, $\mathbf{U}_l$ by either a 1-shock, S_1, a 2-shock, S_2, a 1-rarefaction fan, R_1 or a 2-rarefaction fan, R_2. Because S_1 and R_1 represent *back waves* and S_2 and R_2 are

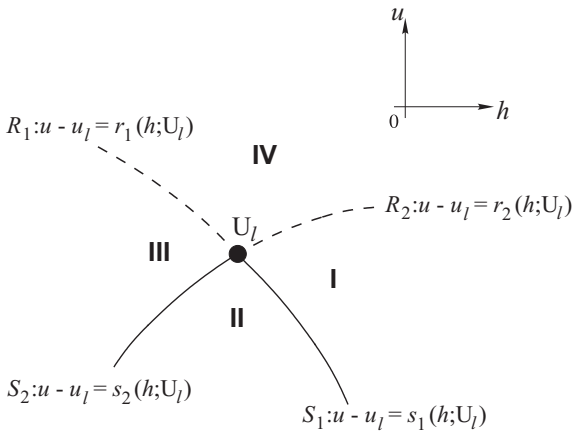


Figure 54: Waves connecting state $\mathbf{U}_l$.

front waves, the Riemann problem will always involve either an S_1 or R_1 and either an S_2 or R_2 wave. Depending in which region $\mathbf{U}_r$ lies, a different Riemann problem must be solved. These are depicted in Figure 57 for zone I, Figure 58 for zone II, Figure 59 for zone III, and in Figure 60 for zone IV. The characteristics and the depth and velocity profile for each case is also shown in their respective figure for subcritical flows.

Denoting all the waves associated with the eigenvalue, λ_1, S_1 and R_1 as *back waves* and those associated with the other eigenvalue for the shallow water wave equations, λ_2, S_2 and R_2 as *front waves*. For convenience, the solution to the the Riemann problem involves travelling from the left

Table 4: Wave structure connecting the left state, $\mathbf{U}_l$ with the right state, $\mathbf{U}_r$ located in the four zones representing the possible solution to the Riemann problem.

Zone	**Left Wave**	**Right Wave**
I	shock (S_1)	shock (S_2)
II	shock (S_1)	rarefaction (R_2)
III	rarefaction (R_1)	shock (S_2)
IV	rarefaction (R_1)	rarefaction (R_2)

state clockwise to the right state. Since $\lambda_2 < \lambda_1$ then the solution of the Riemann problem involves travelling along S_1 or R_1 then along S_2 and R_2.

For an arbitrary flux function $\mathbf{F}(\mathbf{U})$ the following relationship between the shock speed, S and the states $\mathbf{U}_l$ and $\mathbf{U}_r$ across the shock, called the *Rankine-Hugoniot shock condition* condition must be satisfied[74]

$$\mathbf{F}(\mathbf{U}_l) - \mathbf{F}(\mathbf{U}_r) = S(\mathbf{U}_l - \mathbf{U}_r). \tag{125}$$

For (**??**) with $\mathbf{S}^T = [0, 0]$ this gives the following shock speeds

$$S = \frac{u_l^2 h_l - u_r^2 h_r + g/2\left(h_l^2 - h_r^2\right)}{u_l h_l - u_r h_r}$$

and

$$S = \frac{u_l h_l - u_r h_r}{h_l - h_r}.$$

Eliminating S between these expression, provides the relationship

$$u_r - u_l = \pm\sqrt{\frac{g}{2}\frac{\left[h_r^2 - h_l^2\right]\left[h_r - h_l\right]}{h_r h_l}}$$

across the shock that joins the two states, u_l and u_r.

Using $\mathbf{U}_l$, $\mathbf{U}_r$ and $\mathbf{U}_m$, the intercell flux $\mathbf{f}_{j+1/2}$ at $x = (j + 1/2)\Delta x$ can be calculated.

This is used in (93), which is the re-averaging step, to advance the solution in time.

In defining the Hugonoit locus above, the question of whether a given discontinuity is physically relevant has not been considered. Lax[70] proposed a simple entropy condition for system of equations that are genuinely nonlinear. This condition is known as *Lax entropy condition*, which

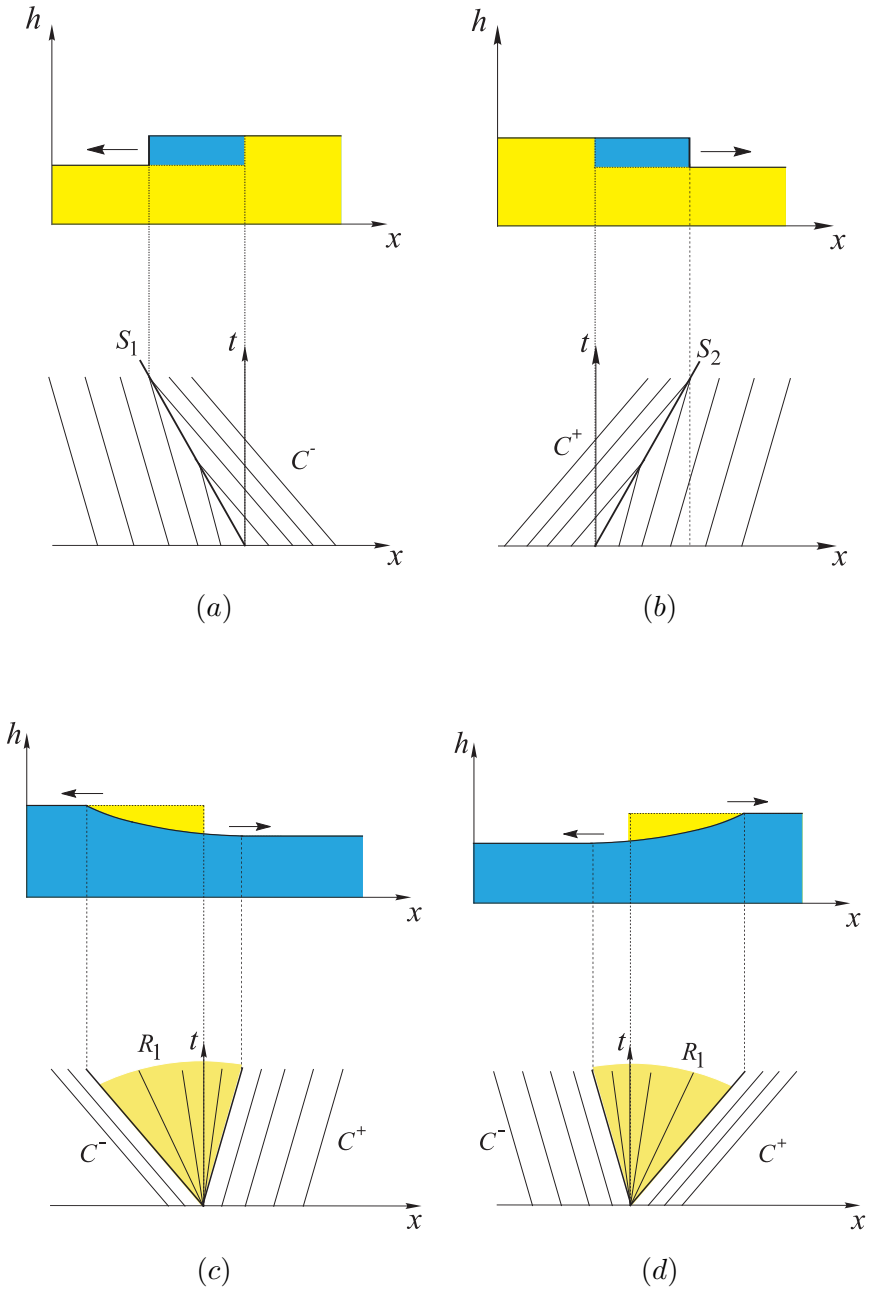


Figure 55: The family of shocks and rarefaction fans that are possible for the shallow water wave equations (*a*) backward shock S_1, (*b*) forward shock S_2, (*c*) backward rarefaction fan R_1 and (*d*) forward rarefaction fan R_2.

will ensure that non-physical solutions will not appear states that a jump in the p-th field is admissible only if

$$\lambda_p(\mathbf{U}_r) < S_1 < \lambda_p(\mathbf{U}_l)$$

Therefore, shocks must satisfy the entropy conditions, see Figure 56; for the 1-shock

$$S_1 < \lambda_1(\mathbf{U}_l) \quad \lambda_1(\mathbf{U}_r) < S_1 < \lambda_2(\mathbf{U}_r) \tag{127}$$

while for the 2-shock

$$\lambda_1(\mathbf{U}_r) < S_2 \quad \lambda_1(\mathbf{U}_l) < S_2 < \lambda_2(\mathbf{U}_l). \tag{128}$$

If these conditions are not satisfied then a non-physical solution is obtained. Violation of these

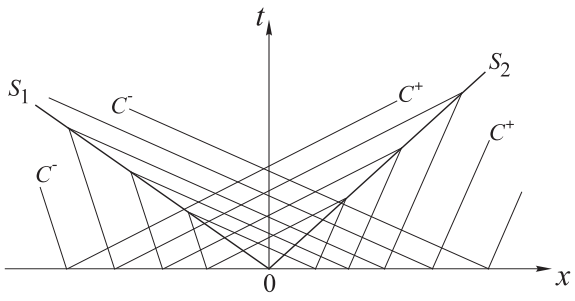


Figure 56: Entropy satisfying shock.

conditions will result in a rarefaction shock or an expansion shock.

Equations (125), (127), (128) and the Riemann invariants

$$u_m + 2\sqrt{gh_m} = u_l + 2\sqrt{gh_l} \tag{129a}$$

and

$$u_m - 2\sqrt{gh_m} = u_r - 2\sqrt{gh_r} \tag{129b}$$

provide the necessary conditions to solve all possible Riemann problems that could be encountered for the shallow water wave equations. There are 16 possible combinations to connect state $\mathbf{U}_l$ with state $\mathbf{U}_r$ and therefore to calculate the intercell flux. Most of these will involve the solution of a non-linear equation.

This set of states is called the *Hugoniot locus*. There is a Hugoniot locus through any point $\mathbf{U}_m$ that gives the set of all points $\mathbf{U}_l$ that can be connected to $\mathbf{U}_r$ by an elementary p-wave. For a general Riemann problem with arbitrary $\mathbf{U}_l$ and $\mathbf{U}_r$, the solution consists of two discontinuities travelling with speeds λ_1 and λ_2, with a new constant state, $\mathbf{U}_m$ in between. The location of $\mathbf{U}_m$ in the pase plane must be where the 1-wave locus through $\mathbf{U}_l$ intersects the 2-wave locus through $\mathbf{U}_r$. In each case we travel from $\mathbf{U}_l$ to $\mathbf{U}_r$ by first going in the direction r_1 and then in the direction r_2.

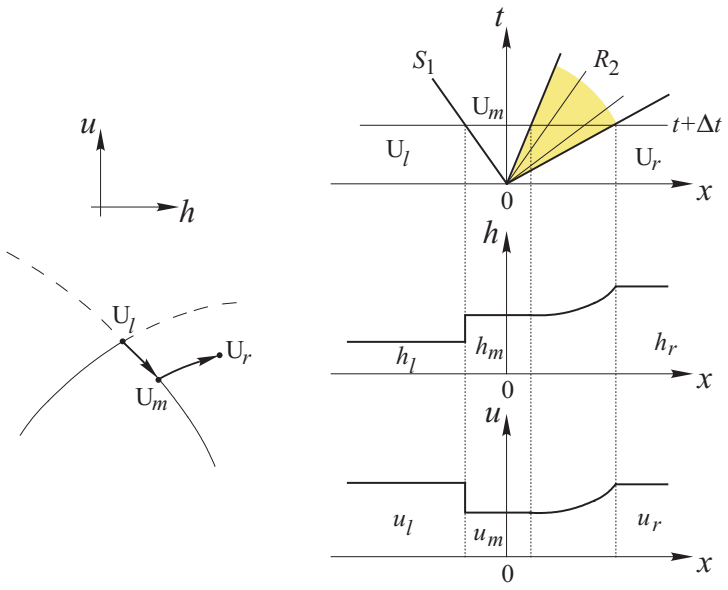


Figure 57: Waves connecting state $\mathbf{U}_l$ when $\mathbf{U}_r$ lies in zone I.

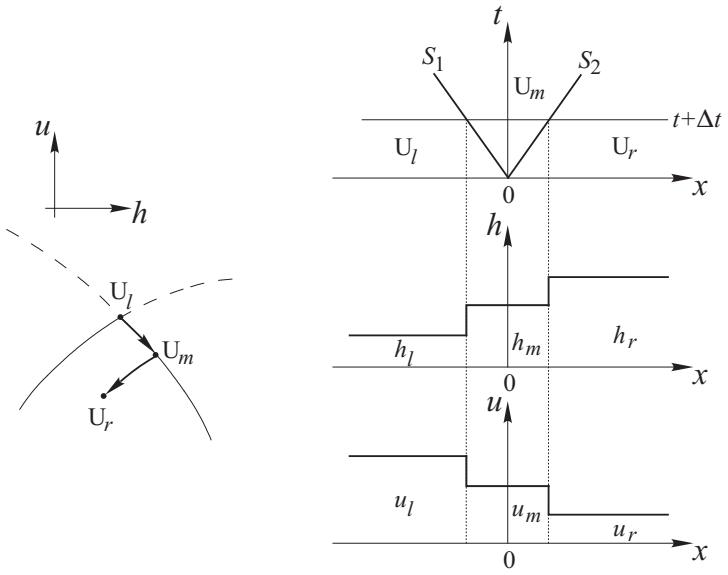


Figure 58: Waves connecting state $\mathbf{U}_l$ when $\mathbf{U}_r$ lies in zone II.

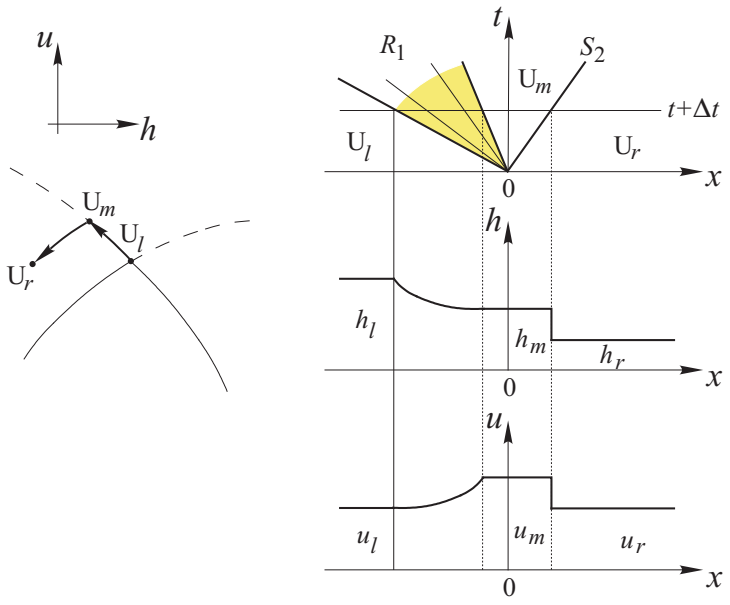


Figure 59: Waves connecting state $\mathbf{U}_l$ when $\mathbf{U}_r$ lies in zone III.

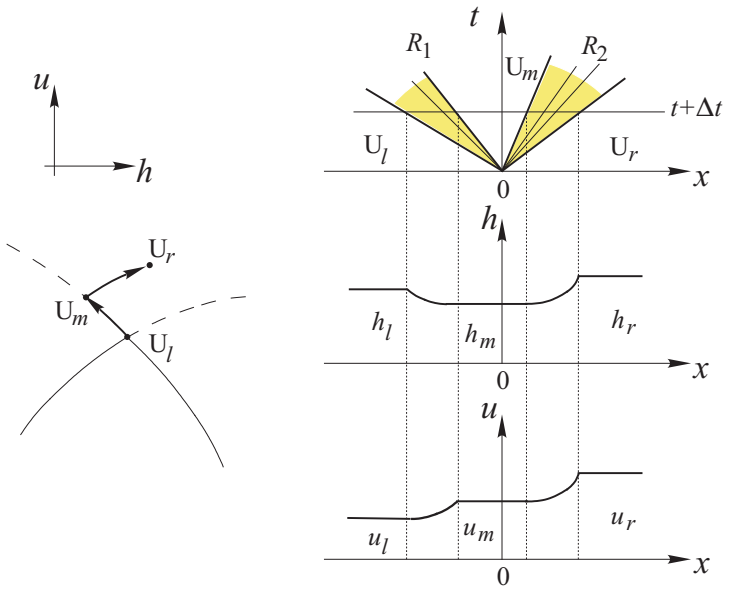


Figure 60: Waves connecting state $\mathbf{U}_l$ when $\mathbf{U}_r$ lies in zone IV.

This is required by the fact that $\lambda_1 < \lambda_2$ since clearly the jump between $\mathbf{U}_l$ and $\mathbf{U}_m$ must travel slower than the jump between $\mathbf{U}_m$ and $\mathbf{U}_r$ if we are to obtain a unique solution.

The simple wave region, which is characterized by a rarefaction fan consists of straight characteristics and curved characteristics emanating from the constant state region surrounding the simple wave region. Along the curved characteristics $u \pm 2c = \alpha^{\pm}$. For the R_1 wave along the curved C^+ characteristic $u + 2c = \alpha^+$, see Figure 61. For the R_2 wave along the curved C^- characteristic $u - c =$ constant, see Figure 62.

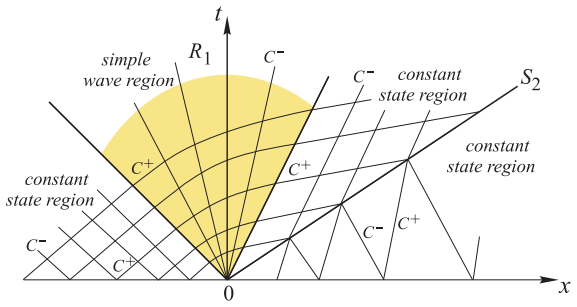


Figure 61: Characteristics in zone III.

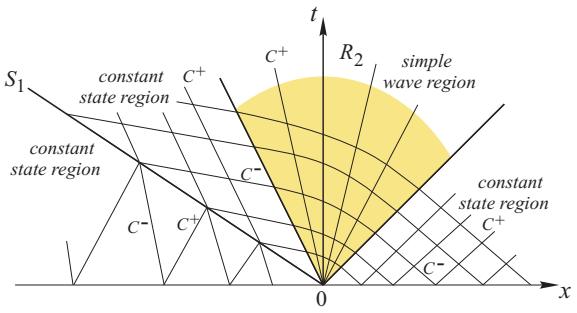


Figure 62: Characteristics in zone I.

Two shocks occur in zone II. The characteristics for this case are shown in Figure 63.

In zone IV there are two rarefaction fans, see Figure 64. In the constant state region the quantities, $u + 2c$ along C^+ and $u - 2c$ along C^- are equal.

If $\mathbf{U}_r$ lies in zone I then from Figure 57 the waves separating the states are

$$U_l \overset{S_1}{\to} U_m \overset{R_2}{\to} U_r$$

The Riemann problem involves solving the 1-shock, S_1 wave followed by the R_2 rarefaction fan. $\mathbf{U}_r$ lies in zone I if $h_r \geq h_l$ and

$$-2\left(\sqrt{gh_r} - \sqrt{gh_l}\right) \leq u_l - u_r < \sqrt{\frac{g}{2}\frac{\left[h_l^2 - h_r^2\right]\left[h_l - h_r\right]}{h_l h_r}}$$

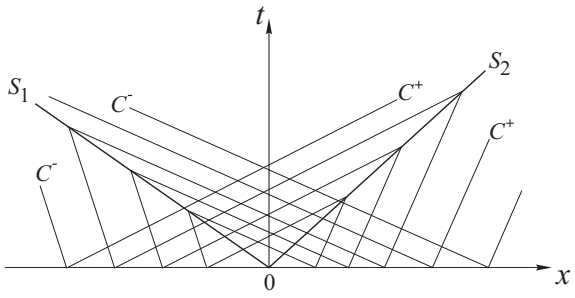


Figure 63: Characteristics in zone II.

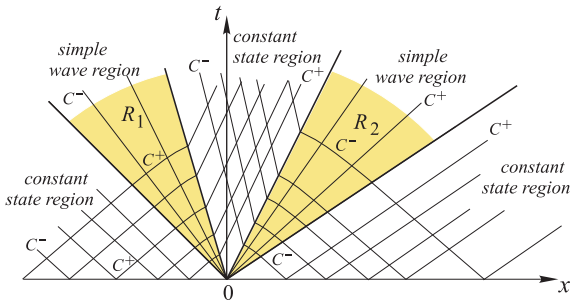


Figure 64: Characteristics in zone IV.

In this region

$$u_l - u_m = \sqrt{\frac{g}{2}\frac{\left[h_m^2 - h_l^2\right]\left[h_m - h_l\right]}{h_m h_l}}$$

and

$$u_r - u_m = 2\left(\sqrt{gh_r} - \sqrt{gh_m}\right)$$

should be solved.

$\mathbf{U}_r$ lies in zone II when

$$u_r - u_l \geq \sqrt{\frac{g}{2}\frac{\left[h_m^2 - h_l^2\right]\left[h_m - h_l\right]}{h_m h_l}}$$

then from Figure 58 the states

$$U_l \overset{S_1}{\rightarrow} U_m \overset{S_2}{\rightarrow} U_r$$

and the following equations must be solved

$$u_l - u_m = \sqrt{\frac{g}{2}\frac{\left[h_m^2 - h_l^2\right]\left[h_m - h_l\right]}{h_m h_l}} \tag{130}$$

with

$$u_m - u_r = \sqrt{\frac{g}{2}\frac{\left[h_m^2 - h_r^2\right]\left[h_m - h_r\right]}{h_m h_r}} \tag{131}$$

for h_m using some nonlinear solver. Substituting h_m into (130) or (131) will give h_m and the intermediate state, $\mathbf{U}_m$.

If $\mathbf{U}_r$ lies in zone III if $h_r < h_l$ and

$$-2\left(\sqrt{gh_l} - \sqrt{gh_r}\right) \leq u_l - u_r < \sqrt{\frac{g}{2}\frac{\left[h_l^2 - h_r^2\right]\left[h_l - h_r\right]}{h_l h_r}}.$$

Then from Figure 59 the states

$$U_l \overset{R_1}{\to} U_m \overset{S_2}{\to} U_r$$

and the following equations must be solved

$$u_m - u_r = \sqrt{\frac{g}{2}\frac{\left[h_m^2 - h_r^2\right]\left[h_m - h_r\right]}{h_m hr}}$$

with

$$u_l - u_m = -2\left(\sqrt{gh_l} - \sqrt{gh_m}\right).$$

In zone IV, then the states are

$$U_l \overset{R_1}{\to} U_m \overset{R_2}{\to} U_r.$$

from Figure 60. $\mathbf{U}_r$ lies in zone IV when $h_r > h_l$ and $u_l - u_r < -2\left(\sqrt{gh_l} - \sqrt{gh_r}\right)$ or $h_r \leq h_l$ and $u_l - u_r < -2\left(\sqrt{gh_r} - \sqrt{gh_l}\right)$. If either of these conditions are satisfied, then the following equations must be solved

$$u_l - u_m = -2\left(\sqrt{gh_l} - \sqrt{gh_m}\right)$$

with

$$u_r - u_m = 2\left(\sqrt{gh_r} - \sqrt{gh_m}\right).$$

There are four possible outcomes for zone I and III, three for zone II and five for zone IV. For zone I, the four possibilities are depicted in Figure **??**. The cases are;

- $\mathbf{f}_{j+1/2} = \mathbf{f}(\mathbf{U}_m)$,
- $S_1 \geq 0 \to \mathbf{f}_{j+1/2} = \mathbf{f}(\mathbf{U}_l)$,
- $u_r + \sqrt{gh_r} \leq 0 \to \mathbf{f}_{j+1/2} = \mathbf{f}(\mathbf{U}_r)$ and
- $u_m + \sqrt{gh_m} < 0 < u_r + \sqrt{gh_r} \to u = (u_r - 2c_r + 2x/t)/3, c = (-u_r + 2c_r + x/t)/3$ then $\mathbf{f}_{j+1/2} = \mathbf{f}(u, h)$.

For zone II the three possibilities are shown in Figure **??**. These result in;

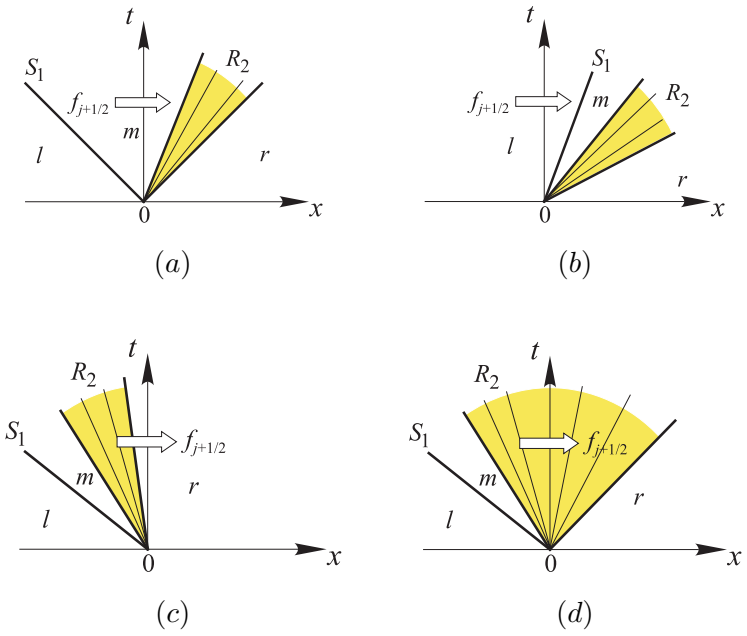


Figure 65: Estimating the correct flux $f_{j+1/2}$ in zone I for the Riemann problem.

- $\mathbf{f}_{j+1/2} = \mathbf{f}(\mathbf{U}_m)$,

- $S_1 \geq 0 \rightarrow \mathbf{f}_{j+1/2} = \mathbf{f}(\mathbf{U}_l)$ and

- $S_2 \leq 0 \rightarrow \mathbf{f}_{j+1/2} = \mathbf{f}(\mathbf{U}_r)$.

For zone III, there are four possible outcomes. From Figure **??**;

- $\mathbf{f}_{j+1/2} = \mathbf{f}(\mathbf{U}_m)$,

- $u_l - \sqrt{gh_l} \geq 0 \rightarrow \mathbf{f}_{j+1/2} = \mathbf{f}(\mathbf{U}_l)$,

- $S_2 \leq 0 \rightarrow \mathbf{f}_{j+1/2} = \mathbf{f}(\mathbf{U}_r)$ and

- $u_l - \sqrt{gh_l} < 0 < u_m - \sqrt{gh_m} \rightarrow u = (u_l + 2c_l + 2x/t)/3, c = (u_l + 2c_l - x/t)/3$ then $\mathbf{f}_{j+1/2} = \mathbf{f}(u, h)$.

Finally in zone IV there are five possible outcomes, which are shown in Figure **??**. For

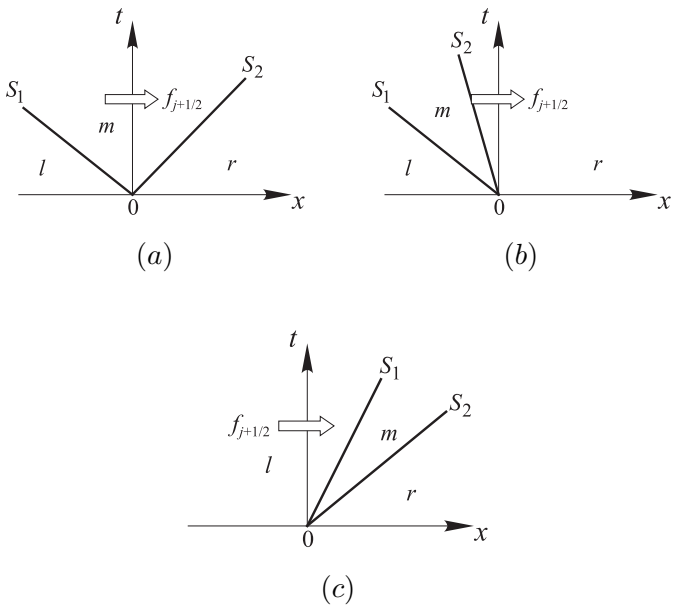


Figure 66: Estimating the correct flux $f_{j+1/2}$ in zone II for the Riemann problem.

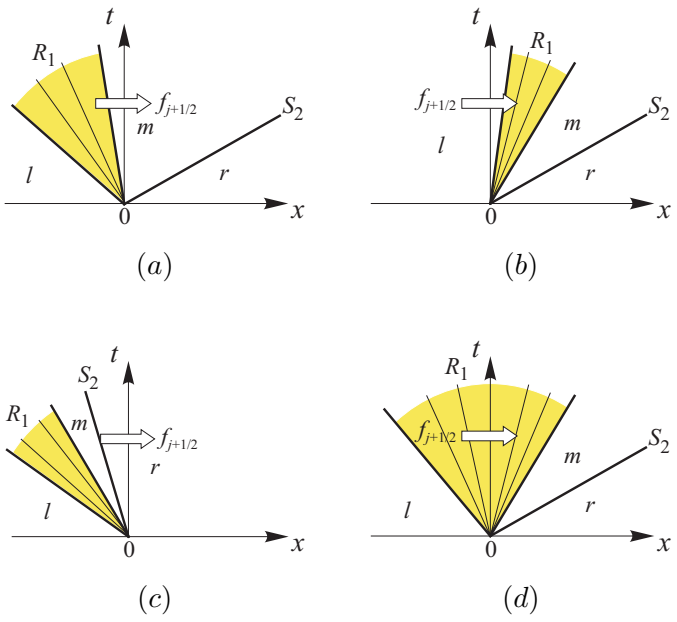


Figure 67: Estimating the correct flux $f_{j+1/2}$ in zone III for the Riemann problem.

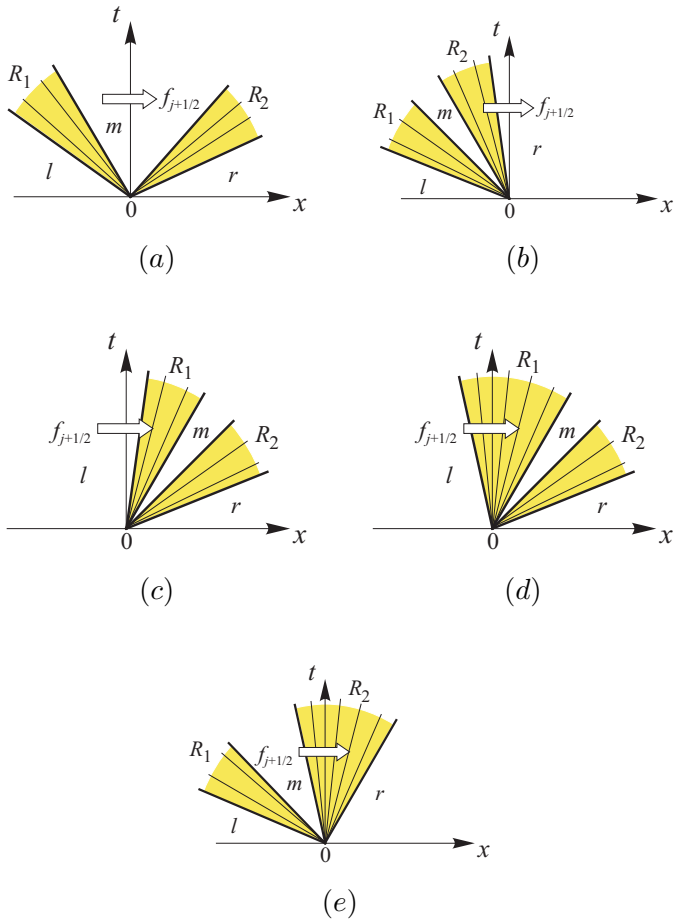


Figure 68: Estimating the correct flux $f_{j+1/2}$ in zone IV for the Riemann problem.

- $\mathbf{f}_{j+1/2} = \mathbf{f}(\mathbf{U}_m)$,
- $u_r + \sqrt{gh_r} \leq 0 \rightarrow \mathbf{f}_{j+1/2} = \mathbf{f}(\mathbf{U}_r)$,
- $u_l - \sqrt{gh_l} \geq 0 \rightarrow \mathbf{f}_{j+1/2} = \mathbf{f}(\mathbf{U}_l)$,
- $u_l - \sqrt{gh_l} < 0 < u_m - \sqrt{gh_m} \rightarrow u = (u_l + 2c_l + 2x/t)/3, c = (u_r + 2c_r - x/t)/3$ then $\mathbf{f}_{j+1/2} = \mathbf{f}(u, h)$.

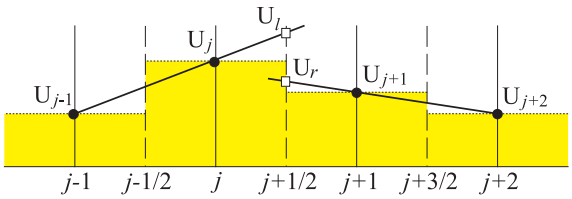


Figure 69: Piecewise linear MUSCL reconstruction using three consecutive cells.

9.5 The Riemann Problem for Constant Coefficient System

The Riemann problem for the hyperbolic, constant coefficient system is a special initial value problem

$$\frac{\partial \mathbf{U}}{\partial t} + \mathbf{A}\frac{\partial \mathbf{U}}{\partial x} = 0 \quad -\infty < x < \infty,\ t > 0, \tag{132}$$

with

$$\mathbf{U}(x,0) = \mathbf{U}^{(0)}(x) = \begin{cases} \mathbf{U}_l & \text{if} \quad x < 0 \\ \mathbf{U}_r & \text{if} \quad x > 0 \end{cases}.$$

The system is strictly hyperbolic with ordered real distinct eigenvalues

$$\lambda_1 < \lambda_2 < \ldots < \lambda_m.$$

The structure of the solution of the Riemann problem (132) in the $x-t$ plane is depicted in Figure 52. It consists of m waves emanating from the origin, one for each eigenvalue λ_i. Each wave i carries a jump discontinuity in $\mathbf{U}$ propagating with speed λ_i. Naturally, the solution to the left of the λ_1 wave is simply the initial data $\mathbf{U}_l$ and to the right of the λ_m wave is $\mathbf{U}_r$. The task at hand is to find the solution in the wedge between the λ_1 and λ_m waves. As the eigenvectors $\mathbf{M}^{(1)}, \mathbf{M}^{(2)}, \ldots, \mathbf{M}^{(m)}$ are linearly independent, we can expand the data $\mathbf{U}_l$, constant left state, and $\mathbf{U}_r$, constant right state, as linear combinations of the set $\mathbf{M}^{(1)}, \mathbf{M}^{(2)}, \ldots, \mathbf{M}^{(m)}$, that is

$$\mathbf{U}_l = \sum_{i=1}^{m} \alpha_i \mathbf{M}^{(i)}, \quad \mathbf{U}_r = \sum_{i=1}^{m} \beta_i \mathbf{M}^{(i)} \tag{133}$$

with constant coefficients α_i, β_i for $i = 1, 2, \ldots, m$. Formally, the solution of the initial value problem (132) is given by (122) in terms of the initial data $R_i^{(0)}(x)$ for the characteristic variables and the right eigenvectors $\mathbf{M}^{(i)}$. Note that each of the expansions in (133) is a special case of (122). In terms of the characteristic variables we have m scalar Riemann problems for

$$\frac{\partial R_i}{\partial t} + \lambda_i \frac{\partial R_i}{\partial x} = 0$$

with the initial data obtained by comparing (133) with (122), that is

$$R_i^{(0)}(x) = \begin{cases} \alpha_i & \text{if} \quad x < 0 \\ \beta_i & \text{if} \quad x > 0 \end{cases}$$

for $i = 1, 2, \ldots, m$. From the previous result, see equation (**??**), we know that the solution of these scalar Riemann problems are given by

$$R_i(x,t) = R_i^{(0)}(x - \lambda_i t) = \begin{cases} \alpha_i & \text{if} \quad x - \lambda_i t < 0 \\ \beta_i & \text{if} \quad x - \lambda_i t > 0 \end{cases} . \tag{134}$$

We can thus write the final solution in terms of the original variables as

$$\mathbf{U}(x,t) = \sum_{i=I+1}^{m} \alpha_i \mathbf{M}^{(i)} + \sum_{i=I}^{I} \beta_i \mathbf{M}^{(i)} \tag{135}$$

where the integer $I = I(x,t)$ is the maximum value of the sub-index i for which $x - \lambda_i t > 0$.

The solution for a 2×2 system. As an example consider the Riemann problem for a general 2×2 linear system. From the origin $(0,0)$ in the (x,t) plane there will be two waves traveling with speeds that are equal to the characteristic speeds λ_1 and λ_2 ($\lambda_1 < \lambda_2$) see Figure 51. The solution to the left of $dx/dt = \lambda_1$ is simply the data state $\mathbf{U}_l = \alpha_1 \mathbf{M}^{(1)} + \alpha_2 \mathbf{M}^{(2)}$ and to the right of $dx/dt = \lambda_2$ the solution is the constant data state $\mathbf{U}_r = \beta_1 \mathbf{M}^{(1)} + \beta_2 \mathbf{M}^{(2)}$. The wedge between the λ_1 and λ_2 waves is usually called the *star region* and the solution there is denoted by $\mathbf{U}^*$; its value is due to the passage of two waves emerging from the origin of the initial discontinuity. From the point $P^*(x,t)$ we trace back the characteristics with speed λ_1 and λ_2. These are parallel to those passing through the origin. The characteristics through P^* pass through the initial points $x_0^{(2)} = x - \lambda_2 t$ and $x_0^{(1)} = x - \lambda_1 t$. The coefficients in the expansion (135) for $\mathbf{U}(x,t)$ are thus determined. The solution at point P^* has the form (135). It is a question of choosing the correct coefficients α_i or β_i. Select a time t^* and a point x_l to the left of the slowest wave so $\mathbf{U}(x_l, t^*) = \mathbf{U}_l$, see Figure 53. The solution at the starting point (x_l, t^*) is obviously

$$\mathbf{U}_l = \mathbf{U}_l = \sum_{i=1}^{2} \alpha_i \mathbf{M}_i = \alpha_1 \mathbf{M}^{(1)} + \alpha_2 \mathbf{M}^{(2)}$$

i.e. all coefficients are α's, that is, the point (x_l, t^*) lies to the left of every wave. As we move to the right of (x_l, t^*) on the horizontal line $t = t^*$ we cross the wave $dx/dt = \lambda_i$, hence $x - \lambda_1 t$ changes from negative to positive, see (134), and therefore the coefficient α_1 above changes to β_1. Thus the solution in the entire star region, between the λ_1 and λ_2 waves, is

$$\mathbf{U}^*(x,t) = \beta_1 \mathbf{M}^{(1)} + \alpha_2 \mathbf{M}^{(2)}. \tag{136}$$

As we continue moving right and cross the λ_2 wave the value $x - \lambda_2 t$ changes from negative to positive and hence the coefficient α_2 in (135) and (136) change to β_2, i.e. the solution to the right of the fastest wave of speed λ_2 is

$$\mathbf{U}^*(x,t) = \beta_1 \mathbf{M}^{(1)} + \beta_2 \mathbf{M}^{(2)}.$$

From equation (133) it is easy to see that the jump in $\mathbf{U}$ across the whole wave structure in the solution of the Riemann problem is

$$\Delta\mathbf{U} = \mathbf{U}_r - \mathbf{U}_l = (\beta_1 - \alpha_1)\mathbf{M}^{(1)} + \cdots + (\beta_m - \alpha_m)\mathbf{M}^{(m)}.$$

It is an eigenvector expansion with coefficients that are strengths of the waves present in the Riemann problem. The wave strength of wave i is $\beta_i - \alpha_i$ and the jump in $\mathbf{U}$ across wave i, denoted by

$$(\Delta\mathbf{U})_i = (\beta_i - \alpha_i)\mathbf{M}^{(i)}.$$

When solving the Riemann problem, sometimes it is more useful to expand the total jump $\Delta\mathbf{U} = \mathbf{U}_r - \mathbf{U}_l$ in terms of the eigenvectors and unknown wave strengths $\beta_i - \alpha_i$.

Let $\Delta\mathbf{R} = \mathbf{R}_r - \mathbf{R}_l$ or $\Delta\mathbf{U} = \mathbf{U}_r - \mathbf{U}_l$. Let $\Delta R_i = R_{r_i} - R_{l_i}$ be the jump in the ith characteristic variable. The quantity ΔR_i is often called the *strength* or *amplitude* of the ith wave. The ith wave creates a jump ΔR_i in the ith characteristic variable. Note that $\Delta R_1 + \Delta R_2 + \cdots + \Delta R_n = \Delta\mathbf{R} = \mathbf{R}_r - \mathbf{R}_l$.

9.6 Roe's Solver

The construction of Roe's scheme is based on a local linearization that requires the definition of an analytic Jacobian[101]. For this discrete representation of the system, it is necessary that the condition

$$\Delta\mathbf{F} = \mathbf{J}\Delta\mathbf{U}. \tag{137}$$

For a uniform rectangular channel, $\mathbf{F}$, and $\mathbf{U}$ given by (**??**) and (**??**) produces

$$\mathbf{J} = \frac{\partial\mathbf{F}}{\partial\mathbf{U}} = \begin{bmatrix} 0 & 1 \\ g\dfrac{A}{B} - \dfrac{Q^2}{A^2} & \dfrac{2Q}{A} \end{bmatrix}$$

and

$$\mathbf{J}\Delta\mathbf{U} = \begin{bmatrix} \Delta Q \\ \left(g\dfrac{A}{B} - \dfrac{Q^2}{A^2}\right)\Delta A + \dfrac{2Q}{A}\Delta Q \end{bmatrix}$$

which is equal to

$$\Delta\mathbf{F} = \begin{bmatrix} \Delta Q \\ \left(g\dfrac{A}{B} - \dfrac{Q^2}{A^2}\right)\Delta A + \dfrac{2Q}{A}\Delta Q \end{bmatrix}$$

and (137) is satisfied. However, as pointed out by Garcia-Navarro and Vazquez-Cendon[35] this is not the case for a non-uniform channel. For a non-prismatic channel, $\mathbf{J}\Delta\mathbf{U}$ is identical to that for

a uniform channel. Unfortunately,

$$\Delta\mathbf{F} = \begin{bmatrix} \Delta Q \\ \left(g\dfrac{A}{B} - \dfrac{Q^2}{A^2}\right)\Delta A + \dfrac{2Q}{A}\Delta Q - \dfrac{gA^2}{2B^2}\Delta B \end{bmatrix}$$

where the last term in the second row is due to the fact that $\partial B/\partial x \neq 0$ and (137) is not satisfied. It does however satisfy

$$\Delta\mathbf{F} = \mathbf{J}\Delta\mathbf{U} + \mathbf{V} \tag{138}$$

where

$$\mathbf{V} = \begin{bmatrix} 0 \\ -\dfrac{gA^2}{2B^2}\Delta B \end{bmatrix}.$$

Therefore, for non-uniform channels the fluxes must be corrected and the source term is also modified, as shown in (**??**).

A non-prismatic channel is defined as a channel of arbitrary shape. A non-uniform channel could be a prismatic or non-prismatic channel which changes shape width distance, therefore $\partial A/\partial x|_{h=\text{constant}} \neq 0$.

For a uniform channel Roe's approach replaces the Jacobian in (**??**) by a constant Jacobian matrix $\overline{\mathbf{A}}$=$\overline{\mathbf{A}}\,(\mathbf{U}_r, \mathbf{U}_l)$ which satisfies the Rankine-Hugoniot condition

$$\overline{\mathbf{A}}\,(\mathbf{U}_r, \mathbf{U}_l)\,(\mathbf{U}_r - \mathbf{U}_l) = \mathbf{F}(\mathbf{U}_r) - \mathbf{F}(\mathbf{U}_l) \tag{139}$$

for any finite value of $\mathbf{U}_r - \mathbf{U}_l$. To do this, the parameter vector $\mathbf{Q} = [\sqrt{h}, \sqrt{h}u]^T$ is used to find two matrices, $\overline{\mathbf{B}}$ and $\overline{\mathbf{C}}$ that satisfy $\mathbf{U}_r - \mathbf{U}_l = \overline{\mathbf{B}}(\mathbf{Q}_r - \mathbf{Q}_l)$ and $\mathbf{F}(\mathbf{U}_r) - \mathbf{F}(\mathbf{U}_l) = \overline{\mathbf{C}}(\mathbf{Q}_r - \mathbf{Q}_l)$. This results in the following constant Jacobian matrix

$$\overline{\mathbf{A}}\,(\mathbf{U}_r, \mathbf{U}_l) = \overline{\mathbf{C}}\overline{\mathbf{B}}^{-1} = \begin{bmatrix} 0 & 1 \\ \overline{c}^2 - \overline{u}^2 & 2\overline{u} \end{bmatrix}$$

in which

$$\overline{u} = \frac{\sqrt{h_l}u_l + \sqrt{h_r}u_r}{\sqrt{h_l} + \sqrt{h_r}}, \overline{c} = \sqrt{g\left(\frac{h_l + h_r}{2}\right)} \quad \text{and} \quad \overline{h} = \sqrt{h_l h_r}.$$

This is by far the most common method of approximating the Jacobian. The eigenvalues and right eigenvectors of the Jacobian $\overline{\mathbf{A}}\,(\mathbf{U}_r, \mathbf{U}_l)$ are given by

$$\overline{\lambda}_1 = \overline{u} - \overline{c}, \overline{\lambda}_1 = \overline{u} - \overline{c}, \overline{\mathbf{e}}_1 = [1, \overline{u} - \overline{c}]^T \quad \text{and} \quad \overline{\mathbf{e}}_1 = [1, \overline{u} + \overline{c}]^T.$$

The structure of the solution of the Riemann problem consists of two shocks as depicted in Figure 70 for subcritical flow.

It consists of two waves emanating from the origin, one for each eigenvalue $\overline{\lambda}_1$ and $\overline{\lambda}_2$. As the eigenvectors are independent then

$$\mathbf{U}_r - \mathbf{U}_l = \Delta\mathbf{U} = \sum_{i=1}^{2} \overline{\alpha}_i \overline{\mathbf{e}}_i \tag{140}$$

which is solved for $\overline{\alpha}_i$, the *wave strengths* to give

$$\overline{\alpha}_1 = \frac{-(\bar{u} - \bar{c})(h_r - h_l) + (uh)_r - (uh)_l}{2\bar{c}}$$

and

$$\overline{\alpha}_2 = \frac{(\bar{u} + \bar{c})(h_r - h_l) - (uh)_r + (uh)_l}{2\bar{c}}.$$

Roe and Pike[102] provide the following $O(\Delta x^2)$ approximation for the wave strengths

$$\overline{\alpha}_1 - \frac{1}{2}\left[h_r - h_l - \overline{h}\frac{u_r - u_l}{\overline{c}}\right]$$

and

$$\overline{\alpha}_2 - \frac{1}{2}\left[h_r - h_l + \overline{h}\frac{u_r - u_l}{\overline{c}}\right]$$

Substituting (140) into (139) and making use of the eigenvalues of $\overline{\mathbf{A}}\,(\mathbf{U}_r, \mathbf{U}_l)$ gives

$$\mathbf{F}(\mathbf{U}_r) - \mathbf{F}(\mathbf{U}_l) = \Delta\mathbf{F} = \overline{\mathbf{A}}\Delta\mathbf{U} = \sum_{i=1}^{2} \overline{\lambda}_i \overline{\alpha}_i \overline{\mathbf{e}}_i. \tag{143}$$

From the Rankine-Hugoniot condition this implies that the shock speed $S_i = \overline{\lambda}_i$. Although this will overestimate the wave speed, it enhances the stability of the numerical scheme[33]. Equation (143) can be applied across each shock to calculate the intercell flux $\mathbf{f}^n_{j+1/2}(\mathbf{U}_r, \mathbf{U}_l)$. The three possible cases are illustrated in Figure 70.

The intercell flux can be estimated using

$$\mathbf{f}^n_{j+1/2} = \mathbf{F}(\mathbf{U}_l) + \sum_{\overline{\lambda}_i \le 0} \overline{\lambda}_i \overline{\alpha}_i \overline{\mathbf{e}}_i,$$

or

$$\mathbf{f}^n_{j+1/2} = \mathbf{F}(\mathbf{U}_r) - \sum_{\overline{\lambda}_i > 0} \overline{\lambda}_i \overline{\alpha}_i \overline{\mathbf{e}}_i$$

where $\sum_{\overline{\lambda}_i > 0}$ denotes the sum of shock with eigenvalues greater than zero only. Taking the average of these expressions then

$$\mathbf{f}^n_{j+1/2} = \frac{1}{2}\left(\mathbf{F}(\mathbf{U}_r) + \mathbf{F}(\mathbf{U}_l)\right) - \frac{1}{2}\sum_{i=1}^{2} |\overline{\lambda}_i| \overline{\alpha}_i \overline{\mathbf{e}}_i. \tag{144}$$

Any of these expressions can be used in (103) to update $\mathbf{U}$.

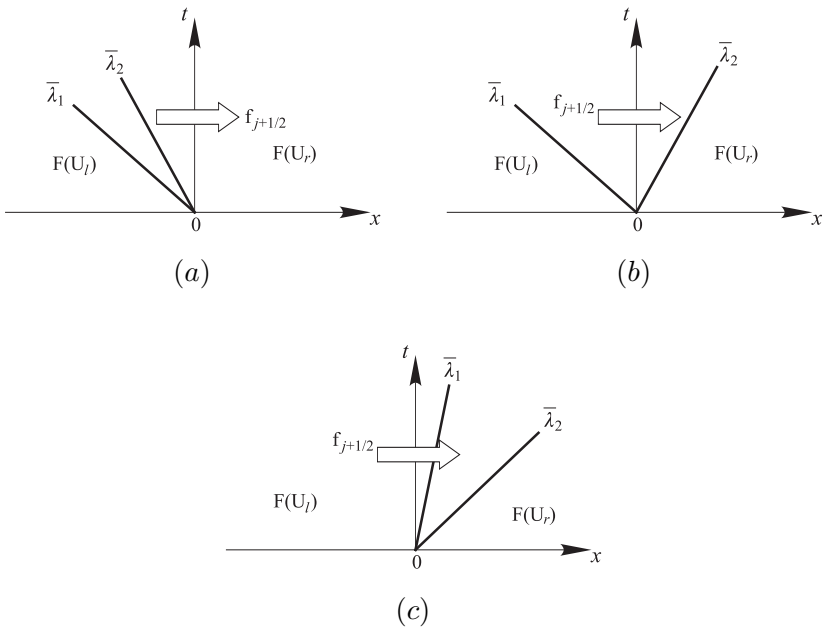


Figure 70: Three possible wave configurations of Roe's approximate Riemann solver.

For a non-uniform channels the intercell flux must be corrected. From (138) (143)

$$\begin{aligned}\sum_{i=1}^{2}\left(\overline{\lambda}_i\overline{\alpha}_i\overline{\mathbf{e}}_i\right)_{j+1/2} &= \Delta\mathbf{F}_{j+1/2}-\mathbf{V}_{j+1/2}\\ &= \mathbf{F}(\mathbf{U}_r)_{j+1/2}-\mathbf{F}(\mathbf{U}_l)_{j+1/2}-\mathbf{V}_{j+1/2}.\end{aligned} \tag{145}$$

Therefore, for (144) the corrected flux becomes

$$\mathbf{f}^n_{j+1/2}=\frac{1}{2}\left(\mathbf{F}_{j+1}+\mathbf{F}_j-\mathbf{V}_{j+1/2}\right)-\frac{1}{2}\sum_{i=1}^{2}\left(|\overline{\lambda}_i|\overline{\alpha}_i\overline{\mathbf{e}}_i\right)_{j+1/2}$$

with

$$\mathbf{f}^n_{j-1/2}=\frac{1}{2}\left(\mathbf{F}_j+\mathbf{F}_{j-1}+\mathbf{V}_{j-1/2}\right)-\frac{1}{2}\sum_{i=1}^{2}\left(|\overline{\lambda}_i|\overline{\alpha}_i\overline{\mathbf{e}}_i\right)_{j-1/2}.$$

Results for the hypothetical example using this scheme are illustrated in Figures 70 and 70.

The problems associated with the assumption that the Riemann problem consists of two shock waves producing a non-physical stationary shock in the region of the rarefaction fan occur for this scheme as well. Substituting (143) and (140) into (144) then the numerical flux can be written in the simplified form as

$$\mathbf{f}^n_{j+1/2}=\frac{1}{2}\left(\mathbf{F}(\mathbf{U}_r)+\mathbf{F}(\mathbf{U}_l)\right)-\frac{1}{2}|\overline{\mathbf{A}}(\mathbf{U}_l,\mathbf{U}_r)|(\mathbf{U}_r-\mathbf{U}_l).$$

At a stationary discontinuity $\overline{\mathbf{A}}(\mathbf{U}_l, \mathbf{U}_r) = 0$ and from (143) implies that $\mathbf{F}(\mathbf{U}_r) = \mathbf{F}(\mathbf{U}_l)$ and the intercell flux $\mathbf{f}^n_{j+1/2} = \mathbf{F}(\mathbf{U}_l)$ which is independent of the requirement that for a physically realistic solution the entropy condition should also be satisfied.

Roe's approximate Riemann solver can be modified so as to avoid entropy violating solutions. *Entropy fixes* can be found in Harten and Hyman[42] and Roe and Pike[102]. Roe's approximate Riemann solver can be modified so as to avoid entropy violating solution. *Entropy fixes* can be found in Harten and Hyman[44], Roe and Pike[102] and Shu and Osher[107]. The Roe and Pike entropy fix was successfully employed by Jha *et al.*[57] in their implementation of Roe's method for estimating the interface flux.

Which corresponds to adding viscosity near sonic points. The parameter δ is empirical and problem dependent.

An important drawback of the Roe scheme is that it admits non-physical stationary jumps, those in which energy increases across the jump, because it cannot properly distinguish a shock from a rarefaction wave that contains a sonic point. Therefore, an entropy correction is needed to ensure entropy-satisfying solutions, energy dissipating in the case of the Saint-Venant equations.

A function ψ is introduced which is the entropy correction to the moduli of the eigenvalues, $|\widetilde{a}^k_{j+1/2}|$ of the Jacobian matrix $\mathbf{A}$. It can either take the form

$$\psi(\widetilde{a}) = \max(\delta, |\widetilde{a}|) \tag{146}$$

or the smoother form[44]

$$\psi(\widetilde{a}) = \begin{cases} |\widetilde{a}| & |\widetilde{a}| \geq \delta \\ \dfrac{(\widetilde{a}^2 + \delta^2)}{2\delta} & |\widetilde{a}| < \delta \end{cases} \tag{147}$$

where δ is a small positive number whose value must be determined for each individual problem. It is usually set constant between 0.1 and 1.

Following Harten and Hyman[44], Delis *et al.*[26] suggest the following formula for δ that eliminates trial procedure for its value and is calculated independently for each computational cell

$$\delta_{j+1/2} = \max(0, \widetilde{a}_{j+1/2} - a_j, a_{j+1} - \widetilde{a}_{j+1/2}) \tag{148}$$

$$\delta_{j-1/2} = \max(0, \widetilde{a}_{j-1/2} - a_{j-1}, a_j - \widetilde{a}_{j-1/2}). \tag{149}$$

Using the above formulas, the entropy correction is enforced only when entropy violating discontinuity appears. These corrections can be viewed as amplifications of the numerical dissipation for the scheme's numerical flux. So that (144) becomes

$$\mathbf{f}^n_{j+1/1} = \frac{1}{2}(\mathbf{F}(\mathbf{U}_r) - \mathbf{F}(\mathbf{U}_l)) - \frac{1}{2}\sum_{i=1}^{2} \psi(\widetilde{\lambda}_i)\widetilde{\alpha}_i\widetilde{\mathbf{e}}_i.$$

Monthe *et al.*[85] use the following entropy fix

Results for the hypothetical example using this scheme are illustrated in Figure **??**(b).

A non-physical stationary shock, in the region of the rarefaction fan has been produced by this scheme.

Dry bed Problem

In Roe's scheme, the approximate Riemann solver contains a priori $p-1$ intermediate states between $\mathbf{U}_l$ and $\mathbf{U}_r$ separated by the signal velocities α_i.

9.7 Osher's scheme

Osher's approach is based on a flux splitting approach[93] where

$$\mathbf{F}(\mathbf{U}) = \mathbf{F}^+(\mathbf{U}) + \mathbf{F}^-(\mathbf{U}) \tag{150}$$

in which

$$\frac{\partial \mathbf{F}^+}{\partial \mathbf{U}} = \mathbf{A}^+(\mathbf{U}) \quad \text{and} \quad \frac{\partial \mathbf{F}^-}{\partial \mathbf{U}} = \mathbf{A}^-(\mathbf{U}). \tag{151}$$

For the Riemann problem, with the initial data $\mathbf{U}_l$ and $\mathbf{U}_r$, then (150) becomes

$$\mathbf{f}_{j+1/2} = \mathbf{F}^+(\mathbf{U}_l) + \mathbf{F}^-(\mathbf{U}_r). \tag{152}$$

The intercell flux consists of contributions from fluxes propagating both upstream and downstream. Using the integral relationships

$$\int_{\mathbf{U}_l}^{\mathbf{U}r} \mathbf{A}^-(\mathbf{U})\, d\mathbf{U} = \mathbf{F}^-(\mathbf{U}_r) - \mathbf{F}^-(\mathbf{U}_l)$$

and

$$\int_{\mathbf{U}_l}^{\mathbf{U}r} \mathbf{A}^+(\mathbf{U})\, d\mathbf{U} = \mathbf{F}^+(\mathbf{U}_r) - \mathbf{F}^+(\mathbf{U}_l)$$

along with (151) then (152) can be expressed as

$$\mathbf{f}_{j+1/2} = \mathbf{F}(\mathbf{U}_l) + \int_{\mathbf{U}_l}^{\mathbf{U}r} \mathbf{A}^-(\mathbf{U})\, d\mathbf{U} \tag{153}$$

or

$$\mathbf{f}_{j+1/2} = \mathbf{F}(\mathbf{U}_r) - \int_{\mathbf{U}_l}^{\mathbf{U}r} \mathbf{A}^+(\mathbf{U})\, d\mathbf{U}.$$

By taking the average of these quantities

$$\mathbf{f}_{j+1/2} = \frac{1}{2}\left(\mathbf{F}(\mathbf{U}_l) + \mathbf{F}(\mathbf{U}_r)\right) - \frac{1}{2}\int_{\mathbf{U}_l}^{\mathbf{U}r} |\mathbf{A}(\mathbf{U})|\, d\mathbf{U}.$$

These are similar to the intercell fluxes estimated in the Roe's scheme. It differs however, in the integration path that is used. It is assumed that the integration path consists of two sub-curves, I_l and I_r, shown in Figure 71 which pass through sonic points S_l where $\lambda_l = 0\text{m/s}$ and S_r

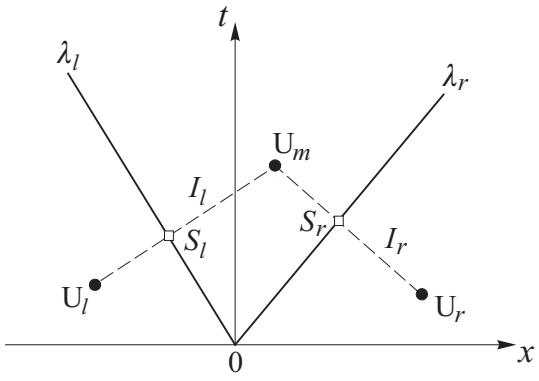


Figure 71: Integration paths, I_l and I_r, sonic points, S_l and S_r, and intersection point, m used in the solution of the shallow water wave equations using Osher's P-ordering approximate Riemann solver.

where $\lambda_r = 0$m/s. These integration paths are curves associated with the set of right eigenvectors (Toro[119]). Consider Osher's P-ordering and using the one-sided flux expression (153) in which

$$\int_{\mathbf{U}_l}^{\mathbf{U}r} \mathbf{A}^-(\mathbf{U})\, d\mathbf{U} = \int_{\mathbf{U}_l}^{\mathbf{U}m} \mathbf{A}^-(\mathbf{U})\, d\mathbf{U} + \int_{\mathbf{U}m}^{\mathbf{U}r} \mathbf{A}^-(\mathbf{U})\, d\mathbf{U}$$

then there are 16 possible outcomes, but only 12 are physically realistic. All physically realistic estimates for Osher's intercell flux, $\mathbf{f}_{j+1/2}$ for the one-dimensional shallow water wave equation using P-ordering, are given in Table 5.

where $\mathbf{F}_{S_l}$ and $\mathbf{F}_{S_r}$ are fluxes associated with the sonic points S_l and S_r respectively.

The quantity, $\mathbf{U}_m = (h_m, u_m)^T$ is obtained using the Riemann invariants (129)

$$u_m + 2\sqrt{gh_m} = u_l + 2\sqrt{gh_l} \quad \text{along} \quad I_l$$

and

$$u_m - 2\sqrt{gh_m} = u_r - 2\sqrt{gh_r} \quad \text{along} \quad I_r$$

which can be solved for

$$u_m = \frac{u_l + u_r}{2} + \sqrt{gh_l} - \sqrt{gh_r} \tag{155a}$$

and

$$h_m = \frac{\left(u_l + 2\sqrt{gh_l} - u_r + 2\sqrt{gh_r}\right)^2}{16g}. \tag{155b}$$

The Riemann invariants are also used to obtain a solution at the sonic points. In Osher's scheme, sonic points S_l, where $\lambda_l = 0$m/s and S_r, where $\lambda_r = 0$m/s are explicitly used in the approximate Riemann solver and therefore this scheme excludes expansion shocks. This is not the case for Roe's

Table 5: Intercell flux, $\mathbf{F}_{j+1/2}$ estimated using Osher's P-ordering approximate Riemann solver for physically realistic flow conditions.

Condition	$u_l - c_l \geq 0$ $u_r + c_r \geq 0$	$u_l - c_l \geq 0$ $u_r + c_r \leq 0$	$u_l - c_l \leq 0$ $u_r + c_r \leq 0$	$u_r + c_r \geq 0$ $u_l - c_l \leq 0$
$u_m - c_m \geq 0$	$\mathbf{F}_l$	$\mathbf{F}_l + \mathbf{F}_r - \mathbf{F}_{S_l}$	$\mathbf{F}_{S_l} + \mathbf{F}_r - \mathbf{F}_{S_r}$	$\mathbf{F}_{S_l}$
$u_m - c_m \leq 0$ $u_m + c_m \geq 0$	$\mathbf{F}_l + \mathbf{F}_m$ $-\mathbf{F}_{S_l}$	$\mathbf{F}_l + \mathbf{F}_m - \mathbf{F}_{S_l}$ $+\mathbf{F}_r - \mathbf{F}_{S_r}$	$\mathbf{F}_m + \mathbf{F}_r$ $-\mathbf{F}_{S_r}$	$\mathbf{F}_m$
$u_m + c_m \leq 0$	$\mathbf{F}_l + \mathbf{F}_{S_r} - \mathbf{F}_{S_l}$	$\mathbf{F}_l + \mathbf{F}_r - \mathbf{F}_{S_l}$	$\mathbf{F}_r$	$\mathbf{F}_{S_r}$

approximate Riemann solver (see, for example Hirsch[50], p. 469). This yields

$$\begin{aligned} u_{S_l} &= \frac{u_l + 2\sqrt{gh_l}}{3}, \\ u_{S_r} &= \frac{u_r - 2\sqrt{gh_r}}{3}, \\ h_{S_l} &= \frac{u_{S_l}}{g} \end{aligned}$$

and

$$h_{S_r} = \frac{u_{S_r}}{g}.$$

These quantities are sufficient to estimate the intercell flux in Osher's P-ordering approximate Riemann solver.

Osher's P-ordering scheme has been used recently by Zhao *et al.*[138][137] to solve the two-dimensional shallow water wave equations. The results for the two test cases using Osher's *physical ordering* or P-ordering scheme are shown in Figures **??** and **??**. Although this first-order scheme has introduced some diffusion it has produced better results than any of the other first-order schemes. Despite the similarity to Roe's approximate Riemann solver, this scheme has overcome the entropy violating solution.

9.8 Harten, Lax and van Leer Scheme

The Harten, Lax and van Leer scheme is based on the scheme described by Harten *et al.*[45]. The intercell flux, $\mathbf{f}_{j+1/2}$ is evaluated by solving the approximate Riemann problem which involves only

two shocks, although two rarefaction fan could be considered separating three states; l, r and m shown in Figure 72 The intercell flux in (114) is chosen from

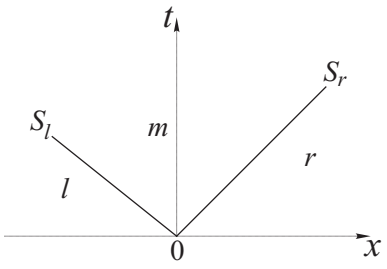


Figure 72: Waves in the one-dimensional Harten, Lax and van Leer scheme.

$$\mathbf{f}_{j+1/2} = \begin{cases} \mathbf{F}(\mathbf{U}_l) & \text{if } S_l \geq 0 \\ \mathbf{F}(\mathbf{U}_m) & \text{if } S_l < 0 < S_r \\ \mathbf{F}(\mathbf{U}_r) & \text{if } S_r \leq 0. \end{cases} \tag{157}$$

For the shallow water waves, the intermediate state variables, u_m, h_m are given by (155a) and (155b). These are used to estimate the shock speeds, S_l and S_r. Using the entropy condition, the shock speeds, S_l and S_r are estimated

$$S_l = \min\left(u_l - \sqrt{gh_l}, u_m - \sqrt{gh_m}\right) \tag{158a}$$

and

$$S_r = \max\left(u_r + \sqrt{gh_r}, u_m + \sqrt{gh_m}\right) \tag{158b}$$

which are estimates of the largest and smallest propagation speed. For left super-critical flow, $S_r < 0$. When $S_l < 0$ and $S_r > 0$ sub-critical flow exists. When $S_r \geq 0$ super-critical flow to the right occurs. Although this approach may overestimate the true wave speed for all combinations of shocks and rarefaction waves[33], Fraccarollo and Toro[33] suggest that this enhances the stability of the scheme and if the Riemann problem involves two rarefaction fans, (155a) and (155b) are the exact solutions. Experience by Fraccarollo and Toro[33] suggest that large errors in h^m do not significantly affect the computed fluxes and the numerical error in the scheme is small.

Equation (158) assumes that there exists a finite water depth everywhere. If a dry bed exists upstream $h_l = 0$, the two eigenvalues collapse into one and the system of equations is not strictly hyperbolic. Under these circumstances no shock exists and S_l represents the speed of the head of

the rarefaction fan and S_r represents the speed of the toe of the rarefaction wave. Therefore,

$$\begin{aligned} S_l &= u_r - 2\sqrt{gh_r} \\ S_m &= S_l \\ S_r &= u_r + \sqrt{gh_r} \end{aligned}$$

Similar expressions can be derived for a dry bed downstream

$$\begin{aligned} S_l &= u_l - \sqrt{gh_l} \\ S_r &= u_l + 2\sqrt{gh_l} \\ S_m &= S_r \end{aligned}$$

The approximation of the wave speeds given by (158) are used to solve for the flux in the intermediate state, $\mathbf{F}(\mathbf{U}_m)$ by satisfying the integral form of the conservation laws

$$\oint [\mathbf{U}\, dx - \mathbf{F}\, dt] = 0 \tag{159}$$

which is valid across a discontinuity. This equation is solved using a control volume approach.

Consider, for example subcritical flow shown in Figure 73. At time t, $-x_l = -S_l t$ and $x_r = S_r t$.

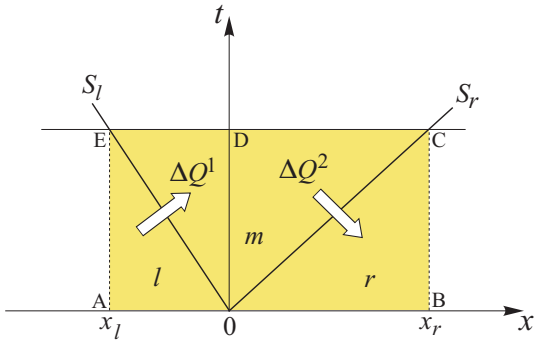


Figure 73: Control volume used in the first-order Harten, Lax and van Leer scheme to solve the integral form of the conservative law for subcritical flow.

Evaluating the integral in the rectangular control volume AODE gives $x_l(\mathbf{U}_l - \mathbf{U}_m) = t(\mathbf{f}(\mathbf{U}_m) - \mathbf{f}(\mathbf{U}_l))$ or $\mathbf{U}_m = \mathbf{U}_l + (\mathbf{f}(\mathbf{U}_m) - \mathbf{f}(\mathbf{U}_l))/S_l$. Similarly for the control volume OBCD gives; $\mathbf{U}_m = \mathbf{U}_r + (\mathbf{f}(\mathbf{U}_r) - \mathbf{f}(\mathbf{U}_m))/S_r$. Eliminating $\mathbf{U}_m$ from these equations gives

$$\mathbf{F}(\mathbf{U}_m) = \frac{S_r\mathbf{F}(\mathbf{U}_l) - S_l\mathbf{F}(\mathbf{U}_r) + S_l S_r(\mathbf{U}_r - \mathbf{U}_l)}{S_r - S_l} \tag{160}$$

This expression is also valid for $S_r \leq 0$ and $S_l \geq 0$. This is all the information required to estimate the intercell flux, $\mathbf{f}_{j+1/2}$ from (157). For the test problem, the results from this first-order approximate Riemann solver is shown in Figure **??**(d). There is reasonable resolution of the

shock and rarefaction fan. More importantly the problems with the inability to satisfy the entropy condition at the dam, common to many of the other schemes has almost been eliminated in this scheme.

Harten, Lax and van Leer[45] developed a first-order approximate Riemann solver where it was assumed that the solution of the Riemann problem consisted of two waves separating three constant states. For the one-dimensional shallow water wave equation this is correct. For other equations or higher dimensional shallow water wave equation, this is not correct. It ignores the existence of intermediate waves. They also assumed that estimates of the lower and upper bounds of the speeds of these non-linear waves, S_l and S_r are known *a-priori*. The Harten, Lax and van Leer scheme was improved by Toro *et al.*[120] by restoring the contact surface. Given the initial data h_l, h_r, u_l, u_r, v_l and v_r the Riemann problem for the augmented one-dimensional shallow water wave equations has the structure shown in Figure 74 in which $S_l \leq 0$ and $S_r \geq 0$. Which has been called the Harten, Lax and van Leer Contact ($HLLC$) scheme in the literature.

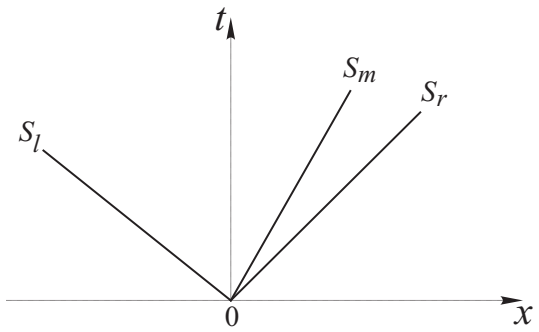


Figure 74: Wave structure of the split two-dimensional approximate Riemann problem used in the first-order Harten, Lax and van Leer Contact scheme, where S_l and S_r are either shocks or rarefaction waves and S_m is a contact discontinuity

The solution for h and u is unaffected by v and thus the complete solution of the Riemann problem for the two-dimensional problem is identical to the solution of the one-dimensional problem.

Since the velocity component v only changes in the intermediate region where $S_c = u^*$, the solution to v is

$$v(x,t) = \begin{cases} v_l & \text{if } x/t \leq S_m \\ \\ v_r & \text{otherwise.} \end{cases}$$

Therefore, the intercell flux for the primary variable v is

$$\mathbf{f}_{j+1/2} = \begin{cases} \mathbf{f}_1 v_l & \text{if } S_m \geq 0 \\ \\ \mathbf{f}_1 v_r & \text{otherwise} \end{cases}$$

where $\mathbf{f}_1$ is the intercell flux estimated using (157) for the conserved quantity uh.

For the dry bed problem when $h_l = 0$m and $h_r \neq 0$, $S_m = S_l$ and $S_m = S_r$ when $h_r = 0$m and $h_l \neq 0$m.

9.9 Weighted Average Flux Scheme

Toro developed a second-order accuracy Godunov-type method which uses a solution averaged over space and time, called the *weighted average flux* scheme.

The local Riemann problem, (113) for the augmented one-dimensional problem, consists of three waves with speeds $a^{(k)}_{j+1/2}$ separating four constant states. In Figure 75 these are denoted by

$$\mathbf{U}^{(k)}_{j+1/2}, \quad 1 \leq k \leq 4 \quad \text{with} \quad \mathbf{U}^{(1)}_{j+1/2} \equiv \mathbf{U}_l \quad \text{and} \quad \mathbf{U}^{(4)}_{j+1/2} \equiv \mathbf{U}_r.$$

The intercell flux, $\mathbf{f}_{j+1/2}$ is estimated by integrating along AB in Figure 75, from $x = -\Delta x/2$ to $x = \Delta x/2$ at time $t = \Delta t/2$ so that

$$\mathbf{f}_{j+1/2} = \frac{1}{\Delta x_j} \int_{-\Delta x_j/2}^{\Delta x_j/2} \mathbf{f}\left(\mathbf{U}^*\right) \, dx \tag{161}$$

$\mathbf{U}^* = \mathbf{U}(2x/\Delta t, \mathbf{U}_l, \mathbf{U}_r)$ is the solution of the Riemann problem with data $\mathbf{U}_l, \mathbf{U}_r$ evaluated at the half time level and over an regular grid, Δx. The integration could be performed exactly, but this would be computationally expensive. In practice it is sufficient to represent all waves in the

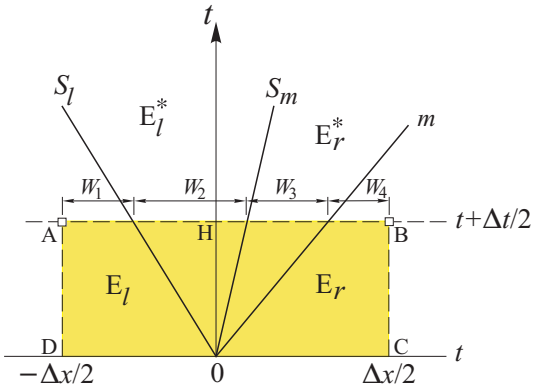


Figure 75: Weights used in the second-order weighted average flux scheme

solution as single straight ray waves, then (161) becomes a weighted average of the fluxes

$$\mathbf{f}_{j+1/2} = \sum_{k=1}^{4} W^{(k)} \mathbf{f}^{(k)}_{j+1/2} \tag{162}$$

where the coefficients W_k or weights are the geometric extent of the constant states. The flux $\mathbf{f}^{(k)}_{j+1/2}$ is the value of the flux function $\mathbf{f}$ in the region k of the local Riemann problem with data $\mathbf{U}_j, \mathbf{U}_{j+1}$.

The region $k = 1$ refers to $\mathbf{U}_l$, on the left side of the wave speed S_l in the (x, t) diagram, $k = 2$ to the region between wave speeds S_l and S_m and so on up to $k = 4$, which refers to state $\mathbf{U}_r$. $S^{(k)}$ is the speed of wave k between states k and $k + 1$ of the Riemann problem. The weights W_k can be written in terms of the Courant-Friedrichs-Lewy numbers $\nu^{(k)}_{j+1/2} = S^{(k)}_{j+1/2}\Delta t/\Delta x$, $k = 1, 2, 3, 4$ with $S^{(k)}_{j+1/2}$ the speed of wave k as follows

$$W^{(k)} = (\nu^{(k)}_{j+1/2} - \nu^{(k-1)}_{j+1/2})/2 \tag{163}$$

in which

$$\nu^{(0)}_{j+1/2} = -1, \quad \nu^{(4)}_{j+1/2} = 1, \quad W^{(k)} \geq 0 \; \forall \; k \quad \text{and} \quad \sum_{k=1}^{4} W^{(k)} = 1.$$

Substituting (163) into (162)

$$\mathbf{f}_{j+1/2} = \frac{(\mathbf{F}(\mathbf{U}_l) + \mathbf{F}(\mathbf{U}_r))}{2} - \frac{1}{2}\sum_{k=1}^{3} \nu^{(k)}_{j+1/2}\Delta\mathbf{f}^{(k)}_{j+1/2}$$

where

$$\Delta\mathbf{f}^{(k)}_{j+1/2} = \mathbf{f}^{(k+1)}_{j+1/2} - \mathbf{f}^{(k)}_{j+1/2}$$

is the flux jump across wave k.

These weights produce a scheme that is second-order accurate in space and time and like all second-order schemes introduces spurious oscillations.. In practice, the weights (163) are replaced by

$$W^{(k)} = (\phi^{(k)}_{j+1/2} - \phi^{(k-1)}_{j+1/2})/2$$

with

$$\phi^{(0)}_{j+1/2} = -1, \quad \phi^{(4)}_{j+1/2} = 1.$$

The functions $\phi^{(k)}_{j+1/2} = \text{sgn}(S^{(k)}_{j+1/2})\psi^{(k)}_{j+1/2}$, where $\text{sgn}(a) = a/|a|$ are *total variation diminishing limiter functions*, used to control the spurious oscillations associated with second-order schemes. One possible choice for the limiter, $\psi^{(k)}_{j+1/2}$ is

$$\psi^{(k)}_{j+1/2} = \begin{cases} 1 - 2(1 - |S^{(k)}_{j+1/2}|) & \text{if } r^{(k)}_j > 2 \\ 1 - r^{(k)}_j(1 - |S^{(k)}_{j+1/2}|) & \text{if } 1 \leq r^{(k)}_j \leq 2 \\ |S^{(k)}_{j+1/2}| & \text{if } 1/2 \leq r^{(k)}_j \leq 1 \\ 1 - 2r^{(k)}_j(1 - |S^{(k)}_{j+1/2}|) & \text{if } 0 \leq r^{(k)}_j \leq 1/2 \\ 1 & \text{if } r^{(k)}_j \leq 0 \end{cases}$$

which is related to the well known *superbee* limiter. Another choice for a limiter is

$$\psi_{j+1/2}^{(k)} = \begin{cases} |S_{j+1/2}^{(k)}| & \text{if } r_j^{(k)} > 1 \\ r_j^{(k)} & \text{if } 0 \le r_j^{(k)} \le 1 \\ 1 & \text{if } r_j^{(k)} \le 0 \end{cases}$$

which is based on the *minmod* limiter. The limiter

$$\psi_{j+1/2}^{(k)} = \begin{cases} 1 - 2r_j^{(k)} \left(1 - |S_{j+1/2}^{(k)}|\right) / (1 + r_j^{(k)}) & \text{if } r_j^{(k)} \ge 0 \\ 1 & \text{if } r_j^{(k)} \le 0 \end{cases}$$

is based on the van Leer limiter.

The effect of these limiters as well as several other forms that can be found in the literature are the same. They add sufficient dissipation to the scheme to guarantee monotonicity near steep gradient, while retaining second-order accuracy in smooth regions. These limiter functions depend on the local Courant-Friedrichs-Lewy numbers $\nu_{j+1/2}^{(k)}$ and a local flow parameter

$$r_{j+1/2}^{(k)} = \frac{\Delta q_{\text{upwind}}}{\Delta q_{\text{local}}}$$

where q is some suitable variable. Here, Δq_{local} is the local jump in q given by

$$\Delta q_{\text{local}} = q_{j+1/2}^{(k+1)} - q_{j+1/2}^{(k)}$$

and Δq_{upwind} is the upwind jump

$$\Delta q_{\text{upwind}} = \begin{cases} q_{j-1/2}^{(k+1)} - q_{j-1/2}^{(k)} & \text{if } S_{j+1/2}^{(k)} \ge 0 \\ q_{j+3/2}^{(k+1)} - q_{j+3/2}^{(k)} & \text{if } S_{j+1/2}^{(k)} < 0. \end{cases}$$

If $\Delta q_{\text{local}} = 0$ then $r_{j+1/2}^{(k)} = 1$. For most waves q can be taken as either the water depth or flow velocity for the shallow water wave equations. Fraccarollo and Toro[33] recommend the use of $q = u$ as the better choice for two-dimensional shallow water problems. Toro[118] used $q = h$ for the shocks and $q = v$ for the contact discontinuity. These are the choices used here for the solution of the two-dimensional shallow water wave equation using the second-order scheme. Note that $r_{j+1/2}^{(k)} \approx 1$ corresponds to gradients in q that vary slowly in space, and $r_{j+1/2}^{(k)} \approx 0$ corresponds to rapid increase in the magnitude of the gradient q in the direction of wave propagation. In the extreme case $r_{j+1/2}^{(k)} = 0 \ \forall \ k$, in which case the second-order intercell flux at boundary $j + 1/2$ reduces to the Godunov first-order intercell flux.

The second-order total variation diminishing intercell flux becomes

$$\mathbf{f}_{j+1/2} = \frac{\mathbf{F}(\mathbf{U}_l) + \mathbf{F}(\mathbf{U}_l)}{2} - \frac{1}{2} \sum_{k=1}^{3} \text{sgn}(S_{j+1/2}^{(k)}) \psi_{j+1/2}^{(k)} \Delta \mathbf{f}_{j+1/2}^{(k)}. \tag{164}$$

Estimates of the fluxes $\mathbf{f}^{(2)}$ and $\mathbf{f}^{(3)}$ are required by (164), see Figure 75. Since the solution for h and u are unaffected by v, then the fluxes across the contact discontinuity are identical except for the flux involving the variable v. The fluxes across the contact discontinuity are given by

$$\mathbf{f}^{(2)} = \begin{bmatrix} \mathbf{f}_1^{(2)} \\ \mathbf{f}_2^{(2)} \\ \mathbf{f}_1^{(2)} v_l \end{bmatrix} \quad \text{and} \quad \mathbf{f}^{(3)} = \begin{bmatrix} \mathbf{f}_1^{(2)} \\ \mathbf{f}_2^{(2)} \\ \mathbf{f}_1^{(2)} v_r \end{bmatrix}$$

where $\mathbf{f}_1^{(2)}$ and $\mathbf{f}_2^{(2)}$ are the vector components of the flux calculated by satisfying (115) over the control volume ABCD shown in Figure 75 for subcritical flow. At time $\Delta t/2$, $\Delta x = S_l \Delta t$ and $\Delta x = S_r \Delta t$. Evaluating (115) over the control volume AH0D gives $\mathbf{U}^* = \mathbf{U}_l + (\mathbf{f}^{(2)} - \mathbf{f}^{(1)})/S_l$ and over the control volume HBC0, $\mathbf{U}^* = \mathbf{U}_r + (\mathbf{f}^{(4)} - \mathbf{f}^{(3)})/S_r$. Eliminating $\mathbf{U}^*$ yields

$$\mathbf{f}^{(2)} = \mathbf{f}^{(3)} = \frac{S_r \mathbf{f}^{(1)} - S_l \mathbf{f}^{(4)} + S_l S_r (\mathbf{U}_r - \mathbf{U}_l)}{S_r - S_l} \tag{165}$$

which is also satisfied for supercritical flow. These fluxes along with $\mathbf{f}_j$ and $\mathbf{f}_{j+1}$ are used in (164) to estimate the intercell flux $\mathbf{f}_{j+1/2}$.

Causon *et al.*[13] use the van Leer limiter, MUSCL[124] averaging with the first-order approximate Riemann solver described by Toro[118].

9.10 Piecewise Parabolic Method

Piecewise parabolic method[19] is a higher-order extension of Godunov's method. Quartic curves are used to interpolate the dependent variables on the edges of piecewise zones. In the averaging step, parabolic interpolation is used to obtain left and right states $\mathbf{U}_l$ and $\mathbf{U}_r$ in the Riemann problem. These values are constrained so that new extrema are not produced. The domain of dependence for the zone interface is determined and the piecewise monotone parabolic function within the domain of dependence is integrated. The left and right states at a zone interface are simply the averages of the integrated quantities over the domain of dependence. The Riemann problem is solved using the left and right states at a zone interface. If the zones are equally spaced then the scheme is fourth-order accurate, otherwise it is third-order.

The piecewise parabolic method has been used by Savic and Holly[103].

Results using the Piecewise parabolic method for the test problem is shown in Figure **??**(g). In this problem, the first-order approximate Riemann solver described by Fraccarollo and Toro[33], (155a)-(160) was used. There is excellent resolution of both the rarefaction fan and the shock.

Following Collela and Woodward[19], let $x_{j+1/2}$ be the boundary between the jth and the $j+1st$ zones on the computational grid, see Figure 76. Within each zone, the function $\psi(x)$, which represents either the primary variables is approximated by a second-order polynomial

$$\psi(x) = ax^2 + bx + c.$$

The coefficients, a, b and c are determined by satisfying

$$\psi_j = \frac{1}{\Delta x}\int_{x_{j-1/2}}^{x_{j+1/2}} \psi(x)\,dx = \frac{1}{\Delta x}\int_{x_{j-1/2}}^{x_{j+1/2}} (ax^2 + bx + c)\,dx.$$

Two additional conditions are required to uniquely define the unknown coefficients. These are provided by the values $\psi(x)$ at both ends of the interval. These are

$$\psi(x_{j+1/2}) = \psi_{R,j} \quad \text{and} \quad \psi(x_{j-1/2}) = \psi_{L,j}$$

see Figure 76, which are unknown. From the above conditions the polynomial representation within the interval $x_{j-1/2}$ and $x_{j+1/2}$ is

$$\psi(x) = \psi_{L,j} + \frac{x - x_{j-1/2}}{\Delta x}\left[\Delta\psi_j + \psi_{6,j}\left(1 - \frac{x - x_{j-1/2}}{\Delta x_j}\right)\right] \tag{166}$$

where

$$\Delta\psi = \psi_{R,j} - \psi_{L,j} \quad \text{and} \quad \psi_{6,j} = 6\left[\psi_j - \frac{(\psi_{R,j} + \psi_{L,j})}{2}\right].$$

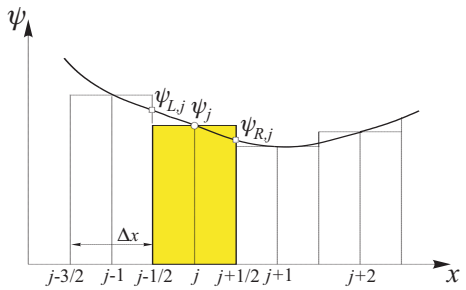


Figure 76: Interpolation step of the piecewise parabolic method method.

The value $\psi_{R,j}$ are obtained using an auxiliary integral function

$$A(x_{j+1/2}) = \int_0^{x_{j+1/2}} \psi(x')\,dx'$$

which is simply the area under the piecewise constant zones. Therefore

$$A(x_{j+1/2}) = \int_0^{x_{j+1/2}} \psi(x')\,dx' = \sum_{i\le j} \psi_i^n \Delta x.$$

The integral function $A(x)$ is approximated by a quartic interpolation polynomial passing through the points $j - 3/2, j - 1/2, j + 1/2, j + 3/2$ and $j + 5/2$, see Figure 76. The quartic polynomial is

$$A(x) = a(x - x_{j+1/2})^4 + b(x - x_{j+1/2})^3 + c(x - x_{j+1/2})^2 + d(x - x_{j+1/2}) + e.$$

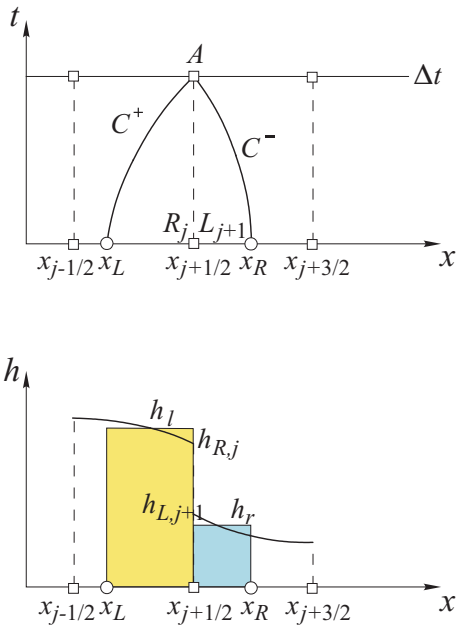


Figure 77: Interpolation step of the piecewise parabolic method method for subcritical flow.

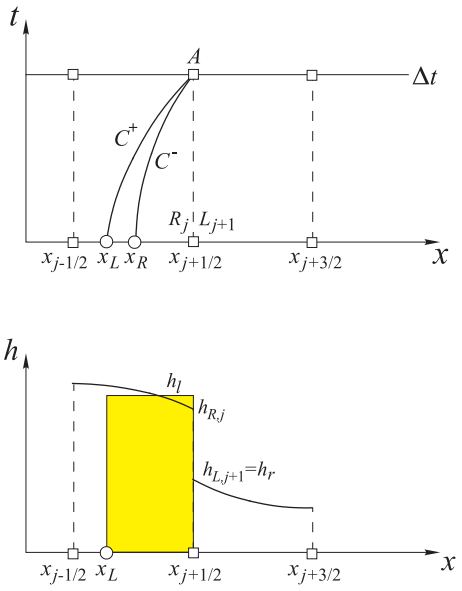


Figure 78: Interpolation step of the piecewise parabolic method method for supercritical flow in the $x-$direction.

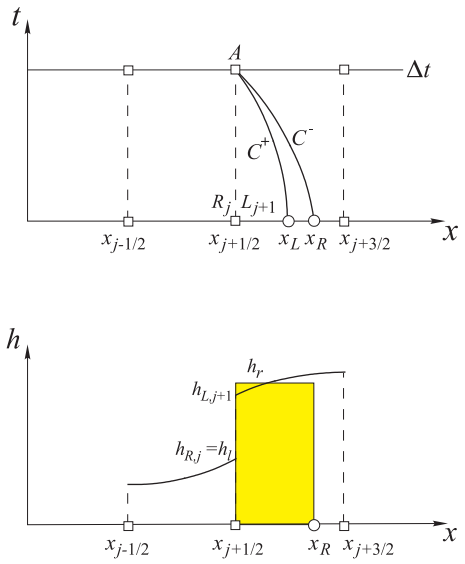


Figure 79: Interpolation step of the piecewise parabolic method method for supercritical flow in the negative $x-$direction.

The function value $\psi_{R,j}$ is obtained by differentiating $A(x)$ so that

$$\left.\frac{dA}{dx}\right|_{x=x_{j+1/2}} = \psi^n(x_{j+1/2}) = \psi_{R,j}.$$

The expression for $\psi^n(x_{j+1/2})$, in terms of ψ_j^n is given by

$$\psi^n(x_{j+1/2}) = \psi_{R,j} = \frac{1}{2}\left(\psi_j^n + \psi_{j+1}^n\right) - \frac{1}{6}\left(\delta\psi_{j+1}^n - \delta\psi_j^n\right) \tag{167}$$

where

$$\delta\psi_j^n = \frac{\psi_{j+1}^n - \psi_{j-1}^n}{2}.$$

This provides the necessary information to define the interpolating parabola.. However near extrema and sharp gradients the interpolating parabola may introduction artificial oscillations in the solution. Two modifications have been incorporated in the interpolation algorithm to produce monotone solutions. This involves two separate steps. The first is replacing $\delta\psi_j^n$ with

$$\delta\psi_j^n = \begin{cases} \min\left(\begin{array}{c} |\delta\psi_j^n|, 2|\delta\psi_{j+1}^n - \delta\psi_j^n|, \\ 2|\delta\psi_j^n - \delta\psi_{j+1}^n|)\mathrm{sgn}(\delta\psi_j^n) \end{array}\right) & \text{if } \left(\psi_{j+1}^n - \psi_j^n\right)\left(\psi_j^n - \psi_{j-1}^n\right) > 0 \\ 0 & \text{if } \left(\psi_{j+1}^n - \psi_j^n\right)\left(\psi_j^n - \psi_{j-1}^n\right) \le 0 \end{cases}$$

in (167) which ensures that $\psi^n(x_{j+1/2}) \in \left[\min(\psi_j^n, \psi_{j+1}^n), \max(\psi_j^n, \psi_{j+1}^n)\right]$ and produces steeper representations of discontinuities. Only when $|\Delta\psi_j| \ge |\psi_{6,j}|$, which ensures that an overshoot has

not occurred, then $\psi_{L,j} = \psi_{R,j-1} = \psi^n(x_{j+1/2})$ and the solution is continuous at $x_{j+1/2}$. If this condition fails to hold, to ensure that quadratic is monotone

$$\begin{aligned}
\psi_{L,j} &= \psi_{R,j} = \psi_j^n && \text{if } \left(\psi_{R,j} - \psi_j^n\right)\left(\psi_j^n - \psi_{L,j}\right) > 0 \\
\psi_{L,j} &= 3\psi_j^n - \psi_{R,j} && \text{if } (\psi_{R,j} - \psi_{L,j})\left(\psi_j^n - \frac{\psi_{R,j} + \psi_{L,j}}{2}\right) > \frac{(\psi_{R,j} - \psi_{L,j})^2}{6} \\
\psi_{R,j} &= 3\psi_j^n - \psi_{L,j} && \text{if } (\psi_{R,j} - \psi_{L,j})\left(\psi_j^n - \frac{\psi_{R,j} + \psi_{L,j}}{2}\right) < -\frac{(\psi_{R,j} - \psi_{L,j})^2}{6}.
\end{aligned}$$

The monotone constraints on each interval $(x_{j-1/2}, x_{j+1/2})$ introduces the discontinuity at zone edges. Once $\psi_{L,j}$ and $\psi_{R,j}$ are known for all internal computational nodes, then the local quadratic equation, given by (166) can be defined for $\psi(x)$. This function is integrated over the domain of dependence of A, see Figure 77. For subcritical flow the

$$\psi_l(x_{j+1/2}) = \frac{1}{x_{j+1/2} - x_L} \int_{x_L}^{x_{j+1/2}} \psi(x)\, dx$$

and

$$\psi_r(x_{j+1/2}) = \frac{1}{x_R - x_{j+1/2}} \int_{x_{j+1/2}}^{x_R} \psi(x)\, dx$$

which represent averages of the quantities that could move across the interface $x_{j+1/2}$. These quantities can be evaluated explicitly, and are given by

$$\psi_l = \psi_{L,j} + \frac{1}{1 - \xi_L}\left[(\Delta\psi_j + \psi_{6,j})\frac{1 - \xi_L^2}{2} - \psi_{6,j}\frac{1 - \xi_L^3}{3}\right]$$

for zone j left of the discontinuity and

$$\psi_r = \psi_{L,j+1} + (\Delta\psi_{j+1} + \psi_{6,j+1})\frac{\xi_R^2}{2} - \psi_{6,j+1}\frac{\xi_R^3}{3}$$

for zone $j + 1$ to the right of the discontinuity, where

$$\xi_L = \frac{x_L - x_{j-1/2}}{\Delta x_j}, \xi_R = \frac{x_R - x_{j+1/2}}{\Delta x_{j+1}} \quad \text{and} \quad \Delta x_j = x_{j+1/2} - x_{j-1/2}.$$

If the flow is supercritical in the $x-$direction, see Figure 78 x_R is located outside the zone $j+1$. In this case the C^- characteristic is moved back to $x_{j+1/2}$ and $\psi_r = \psi_{L,j+1}$. Similarly for supercritical flow in the negative $x-$direction, see Figure 79, $\psi_l = \psi_{R,j}$. These averages are used in a Riemann solver to establish the flux at the zone interface. The final step involves the re-mapping of the solution of the Riemann problem to the nodal values which is accomplished using

$$\mathbf{U}_j^{n+1} = \mathbf{U}_j^n - \frac{\Delta t}{\Delta x}\left(\mathbf{F}_{j+1/2}^{n+1} - \mathbf{F}_{j-1/2}^{n+1}\right).$$

Results using the piecewise parabolic method for the two test problems are shown in Figures **??**(f) and **??**(f). In this problem the first-order approximate Riemann solver described by Fraccarollo

and Toro[33], (157)-(160) was used. There is excellent resolution of both the rarefaction fan and the shock. This is despite the use of a first-order Riemann solved and assuming that the characteristics are straight. Savic and Holly[103] use trapezoidal rule to integrate along the characteristics. This was necessary because they included source terms in the governing equations.

Results using the piecewise parabolic method for the test problem are shown in Figure **??**(g). In this problem, the first-order Harten-Lax-van Leer approximate Riemann solver described by Fraccarollo and Toro[33], (155a)-(160) was used to solve the Riemann problem. There is excellent resolution of both the rarefaction fan and the shock, with the shock resolved over the fewest computational nodes.

10 Splitting Techniques

Many practical problems involve more than a single spatial dimension as well as source and sink terms. Instead of developing multi-dimensional numerical schemes and incorporating sink and source terms in the underlying numerical scheme, a *splitting* strategy can be employed. In splitting methods the problem is broken down into pieces corresponding to the different processes and numerical methods appropriate for each separate piece is applied independently. This technique is also used to split multi-dimensional problems into a sequence of one-dimensional problems, called *dimensional splitting* and for dealing with reaction or source terms, often called *operator splitting*. For example, the solution of the non-homogeneous equation

$$\frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x} = S(u) \tag{168a}$$

can be split into the solution of the homogeneous equation

$$\frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x} = 0 \tag{168b}$$

to provide $u^{(1)}$ followed by the solution of the ordinary differential equation

$$\frac{du^{(1)}}{dt} = S(u^{(1)}) \tag{168c}$$

for the source term, to provide a complete solution to (168a). This is an example of operator splitting

Consider the general partial differential equation of the form

$$\frac{du}{dt} = (\mathcal{A} + \mathcal{B})u$$

where the two linear operators, $\mathcal{A}$ and $\mathcal{B}$ may be spatial differential operators in independent co-ordinates or representing different physical or chemical processes. Assume that $\mathcal{A}$ and $\mathcal{B}$ do not depend on t, which considerably simplifies the problem. Then

$$\frac{d^j u}{dt^j} = (\mathcal{A} + \mathcal{B})^j u.$$

The solution at time $t + \Delta t$ given the solution at time t, can be expressed using the Taylor series as

$$\begin{aligned} u(x, t + \Delta t) &= u(x, t) + \Delta t(\mathcal{A} + \mathcal{B})u(x, t) + \frac{\Delta t^2}{2}(\mathcal{A} + \mathcal{B})^2 u(x, t) + \cdots \\ &= \left(I + \Delta t(\mathcal{A} + \mathcal{B}) + \frac{\Delta t^2}{2}(\mathcal{A} + \mathcal{B})^2 + \cdots \right) u(x, t) \\ &= \sum_{j=0}^{\infty} \frac{\Delta t^j}{j!}(\mathcal{A} + \mathcal{B})^j u(x, t), \end{aligned} \tag{169}$$

where I is the identity operator, can be formally written as

$$u(x, t + \Delta t) = \exp(\Delta t(\mathcal{A} + \mathcal{B}))u(x, t).$$

The various splitting methods approximate $\exp(\Delta t(\mathcal{A} + \mathcal{B}))$ to the desired accuracy.

10.1 Fractional splitting

The simplest *fractional splitting* method involves computing

$$u^{(1)}(x, t + \Delta t) = \exp(\Delta t \mathcal{A}) u(x, t), \tag{170a}$$

followed by

$$u^{(2)}(x, t + \Delta t) = \exp(\Delta t \mathcal{B}) u^{(1)}(x, t + \Delta t) \tag{170b}$$

over each time step Δt.

Expanding equation (170) using the Taylor series, then

$$\begin{aligned} u^{(2)}(x, t + \Delta t) &= \left(I + \Delta t \mathcal{B} + \frac{\Delta t^2}{2} \mathcal{B}^2 + \cdots \right) \left(I + \Delta t \mathcal{A} + \frac{\Delta t^2}{2} \mathcal{A}^2 + \cdots \right) u(x, t) \\ &= \left(I + \Delta t (\mathcal{A} + \mathcal{B}) + \frac{\Delta t^2}{2} (\mathcal{A}^2 + 2\mathcal{A}\mathcal{B} + \mathcal{B}^2) + \cdots \right) u(x, t). \end{aligned} \tag{171}$$

Equation (171) is identical to equation (169) up to terms Δt^2 if

$$(\mathcal{A} + \mathcal{B})^2 = \mathcal{A}^2 + \mathcal{A}\mathcal{B} + \mathcal{B}\mathcal{A} + \mathcal{B}^2$$

in equation (169) is equivalent to

$$\mathcal{A}^2 + 2\mathcal{A}\mathcal{B} + \mathcal{B}^2$$

in equation (171). This will only occur when $\mathcal{A}$ and $\mathcal{B}$ *commute*, i.e. $\mathcal{A}\mathcal{B} = \mathcal{B}\mathcal{A}$ which is generally not the case. Therefore, (170) is only first-order accurate. Higher-order fractional splitting have been proposed. In many of these schemes the sequence of operations is symmetric over a single time step. Consider the following approximation

$$\begin{aligned} u(x, t + \Delta t) \approx \frac{1}{2} &\left(\exp(\Delta t \mathcal{A}) \exp(\Delta t \mathcal{B}) \right. \\ &\left. + \exp(\Delta t \mathcal{B}) \exp(\Delta t \mathcal{A}) \right) u(x, t) + O(\Delta t^3). \end{aligned} \tag{172}$$

The second-order accuracy of this scheme is obtained from the symmetry of the operations. Unfortunately this scheme requires twice the computational effort per time step than the first-order scheme.

10.2 Strang splitting

A slight modification of the splitting idea will yield second-order accuracy quite generally, provided that each sub-problem is solved to at least second-order accuracy. The approach is often called *Strang splitting*. In second-order Strang splitting, the operator

$$u(x, t + \Delta t) = \exp(\Delta t (\mathcal{A} + \mathcal{B}))$$

is approximated by

$$u(x, t + \Delta t) \approx \exp(\Delta t/2\mathcal{A}) \exp(\Delta t\mathcal{B}) \exp(\Delta t/2\mathcal{A}) \tag{173}$$

over each time step Δt.

Expanding the product using the Taylor series, then

$$\begin{aligned}
\exp(\frac{\Delta t}{2}\mathcal{A}) \exp(\Delta t\mathcal{B}) \exp(\frac{\Delta t}{2}\mathcal{A}) &= \left(I + \frac{\Delta t}{2}\mathcal{A} + \frac{\Delta t^2}{8}\mathcal{A}^2 + \cdots\right) \\
&\left(I + \Delta t\mathcal{B} + \frac{\Delta t^2}{2}\mathcal{B}^2 + \cdots\right)\left(I + \frac{\Delta t}{2}\mathcal{A} + \frac{\Delta t^2}{8}\mathcal{A}^2 + \cdots\right) \\
&= I + \Delta t(\mathcal{A} + \mathcal{B}) + \frac{\Delta t^2}{2}(\mathcal{A}^2 + \mathcal{A}\mathcal{B} + \mathcal{B}\mathcal{A} + \mathcal{B}^2) + O(\Delta t^3)
\end{aligned}$$

which is consistent with equation (169) with all terms up to $O(\Delta t^3)$. Second-order accuracy is achieved even when $\mathcal{A}$ and $\mathcal{B}$ do not commute. The $O(\Delta t^3)$ term is not generally correct unless $\mathcal{A}\mathcal{B} = \mathcal{B}\mathcal{A}$.

Second-order Strang splitting is computationally more efficient than using (172). Employing second-order Strang splitting could be only marginally more expensive than using first-order fractional splitting if one realizes that the process over several time steps can be consolidated. For example, consider the following operations over two time steps

$$\begin{aligned}
u(x, t + 2\Delta t) &= (\exp(\Delta t/2\mathcal{A}) \exp(\Delta t\mathcal{B}) \exp(\Delta t/2\mathcal{A}) \\
&\exp(\Delta t/2\mathcal{A}) \exp(\Delta t\mathcal{B}) \exp(\Delta t/2\mathcal{A}))\, u(x, t) + O(\Delta t^3) \\
&= \exp(\Delta t/2\mathcal{A}) \exp(\Delta t\mathcal{B}) \exp(\Delta t\mathcal{A}) \exp(\Delta t\mathcal{B}) \exp(\Delta t/2\mathcal{A}) u(x, t) + O(\Delta t^3).
\end{aligned}$$

The half time step operations need only be calculated at the beginning and end of the computational period, $2\Delta t$ in this case. Therefore, if the solution is not required at every time step, the computational effort required by this second-order splitting strategy can be significantly reduced.

It is possible to construct a numerical scheme using Strang splitting to any order provided the individual operations are performed to at least the same accuracy. In addition, the homogeneous and non-homogeneous operations can be separated. Thereby, the most appropriate method could be applied to the homogeneous and non-homogeneous terms. Crandall and Majda [23] have demonstrated that at least for the first-order splitting and Strang splitting algorithms will always converge to unique weak solutions satisfying the entropy condition.

11 Boundary Conditions

Boundary conditions are imposed on a problems because; we are only interested in a finite region, computer resources do not permit the simulation of the problem over the whole physical domain.

.

The boundary treatment can be one order lower in the global sense compared to the interior treatment without degrading the order or overall global accuracy. However, if the boundary condition is two or more orders lower in accuracy, it degrades the overall global accuracy. Some related material on the issue of numerical boundary treatment can be found in a paper by Blottner (1982).

So far we have only studied methods for updating the cell average $\mathbf{U}_j^{n+1}$ assuming that we have at least the neighboring cell values $\mathbf{U}_{j-1}^n$ and $\mathbf{U}_{j+1}^n$. In practice we are dealing with a finite domain and the neighboring information is not available for the first and last cells. The simplest way to overcome this problem is to extend the computational domain and introduce *ghost cells* that provide the necessary information to calculate the cell values for the first and last cells. This approach avoids developing special techniques for estimating the flux at these boundaries and it decoupled the boundary conditions from the choice of numerical method that is applied to solve the problem.

Suppose that a one-dimensional computational domain $[0, L]$ is subdivided into $N-1$ cells so that each cells has a width $\Delta x = L/(N-1)$. For numerical schemes which use $\mathbf{U}_{j+1}^n$and $\mathbf{U}_j^n$ to estimate the numerical flux, $\mathbf{f}_{j+1/2}^n$ then a single ghost cell is required at the upstream and downstream boundary conditions. With higher-order schemes two or more ghost cells are required. The information provided at these ghost cells is governed by the flow conditions and the boundary condition imposed at the extremity of the computational domain. The *reflective boundary condition* and the *transmissive boundary condition* OTHERS are described below.

The application of boundary condition in this scheme is straight forward. Assuming that the computational domain $[0, L]$ is subdivided into M cells so that $i = 1, ..., M$ lie within the computational domain. A single fictitious cell is required next to each *reflective* or *transmissive* boundary condition. These are denoted by $i = 0$ for the left boundary, and for the right boundary $i = M + 1$.

The shallow water wave equations are hyperbolic system of equations with two families of characteristics. Therefore, two boundary conditions are required for its solution. In addition, at internal points where the shallow water wave equations are not valid, internal boundary conditions are required.

11.1 Transmissive Boundary Condition

Now suppose $x = 0$ is a solid wall. If we are solving the equations only for $x > 0$, then we can specify values in the ghost cells according to

$$\mathbf{U}_0^n = r\mathbf{U}_1^n, \quad \mathbf{U}_{-1}^n = r\mathbf{U}_2^n$$

where r represents the reflection operator which simply negates the velocity while preserving the mass.

For the transmissive boundary condition at $x = 0$

$$[\mathbf{W}_0^n = \mathbf{W}_1^n$$

where, $\mathbf{W}$ can be the vector of conservative or primitive variables. For the reflective boundary condition at $x = L$

$$h_{M+1}^n = h_M^n \quad u_{M+1}^n = -u_M^n \ .$$

These boundary conditions imply that the boundary lies mid-way between the nodes M and $M+1$.

Some schemes, like second-order schemes require two false nodes. Similar to the first-order scheme two fictitious cells are required next to each *reflective* or *transmissive* boundary condition. These are denoted by $i = -1$ and $i = 0$ for the left boundary, and for the right boundary $i = M+1$ and $i = M+2$. For the transmissive boundary condition at $x = 0$

$$\mathbf{W}_0^n = \mathbf{W}_1^n$$
$$\mathbf{W}_{M+1}^n = \mathbf{W}_M^n$$

where, $\mathbf{W}$ can be the vector of conservative or primitive variables. For the reflective boundary condition at $x = L$

$$h_{M+1}^n = h_M^n \qquad u_{M+1}^n = -u_M^n$$
$$h_{M+2}^n = h_{M-1}^n \quad u_{M+2}^n = -u_{M-1}^n.$$

There are three types of boundary conditions; *actual* boundary conditions along which external influences on the computational domain are specified, *artificial* boundary conditions and *internal* boudary conditions. The need for artificial boundary conditions arises when the spatial domain of the problem is *unbounded*. In this case, a numerical treatment usually requires the introduction of an *artificial* boundary in order to make the computational domain finite. It is desirable that this boundary condition have the property that waves incident on this boundary from within the computational domain are transmitted through the boundary without any reflection. There hase been a considerable amount of work to devise boundary conditions that reduce the amount of spurious reflections. These boundary conditions are referred to as *non-reflective* boundary conditions. They have been also called, *radiating*, *absorbing*, *silent*, *transmitting*, *transparent*, *open*, *free-space* and *one-way* boundary conditions. (**??**)

Regions or boundaries where the governing equations are not applicable can be represented by *internal* boundary conditions. For example, the junction of several tributaries is usually represented as an internal boundary condition.??????MORE

The Saint Venant equations describing one-dimensional unsteady flow in open channels are of the hyperbolic type. Hyperbolic equations require two boundary and initial conditions for their solution. These boundary conditions are used to define external influences on the flow within the reach. Boundary conditions can be provided as either a time history of water levels or discharge,

or a relationship between these parameters at a particular location. For practical problems this information is generally available at a gauging station. Unfortunately, there are many practical problems where a reach of a stream only contains one gauging station. An additional boundary condition is required to guarantee a unique and well-posed solution to the Saint Venant equations. The computational domain can be extended so that the influence of the boundary condition does not affect the flow within the reach of interest. This is not a practical solution, requiring a large amount of computing resources. Under these circumstances it is preferable to introduce an artificial boundary to limit the area of computation.

Since computer restrictions put a limit on the size of a region which can be handled, placing the boundaries a long way from the region of computational interest is impractical. Thus the need for boundary conditions which act as an infinite region of free space is desirable. The primary requirement of such a boundary is the ability to absorb waves incident upon it rather than reflect them. Artificial boundaries are introduced to limit the computational domain. At these boundaries non-physical reflections may occur, limiting the accuracy of the results in this case. It is therefore necessary to use boundary conditions that adsorb waves incident on the artificial boundary condition to eliminate these reflections.

Reflections not only degrade time accuracy of transient solutions but may also inhibit their convergence to steady-state. Spurious reflections from the boundary condition might contaminate the interior solution and the real modifications of the wave would be obscured.

What is needed is a boundary condition which allows phenomena generated in the domain of interest to pass through the boundaries without undergoing significant distortion and without influencing the interior solution.. Unfortunately, perfectly absorbing boundary conditions do not exist[28] In practice, one would like to minimize the amplitude of waves reflected from these artificial boundaries.

This can be achieved using a relationship between the dependant variables The use of a rating curve as the boundary condition is only valid if these is a unique relationship between the dependent variables. A weir would be an example where such a relationship would be valid.

Examples are; (*i*) Sommerfeld radiation condition[92](also known as open or non-reflecting), (*ii*) adsorbing boundary conditions, and (*iii*) damping "sponge" layers near the boundaries. A lateral sponge layer is designed to minimize spurious reflections. The formulation selectively damps out the Riemann invariants (characteristic variables) associated with the reflected waves inside the sponge layer.

To include the sponge layer, the domain is simply expanded. Open boundary conditions are specified at a point.[60]

The boundary effects can appear as reflections. These are always extraneous to the real physical problem being studied and their elimination is desirable. Adsorbing boundary was achieved through a gradual elimination of the wave amplitude in a strip along the boundary of the grid. A class of methods derived more recently is based on replacing the wave equation on the boundary grid by a

one way equation which allows energy to propagate in the outward direction only. This method has proven effective for waves which impinge normal to the boundary, however, the method degrades for other angles of incidence. It is also highly dependent on the wave equation and the spatial discretisation method[65].

The correct boundary condition depends on the external solution, but for many problems the external solution is not known. In such cases non-reflective boundary conditions often produce solutions with the desired behaviour.

Non-reflective in that all waves arriving at the boundary cross it without reflection.

Other schemes involve co-ordinate transformations[41]

Many solution techniques make use of data which are known at infinity, and define the computational domain large enough such that these data can be specified at the artificial boundary. Another procedure sometimes used is to map the infinite domain to a finite one, and to solve the transformed equations on this new domain.

11.2 Dirichlet Boundary Condition

For subcritical flows the Riemann invariants across the S_2 shock at the upstream boundary must satisfy

$$u_b - 2\sqrt{gh_b} = u_r - 2\sqrt{gh_r}. \tag{174}$$

This is the Dirichlet boundary condition. Therefore, if the water depth at the upstream boundary $h_b(t)$ is given, (174) can be used to calculate u_b at the boundary. Similarly if the velocity, $u_b(t)$ or the discharge, $Q_b(t) = u_b(t)A_b$ is given at the upstream boundary condition, $h_b(t)$ can be calculated using the channel properties. For supercritical flows, both $u_b(t)$ and $h_b(t)$ must be specified.

At the downstream boundary, for subcritical flow, the given boundary values along with

$$u_b + 2\sqrt{gh_b} = u_l + 2\sqrt{gh_l}$$

is used to determine the unknown dependent variable.

(**??**)

The outgoing waves, $\lambda_i \leq 0$ at $x = 0$ and $\lambda_i \geq 0$ at $x = L$, where L is the width of the computational domain, depend only on information at and within the boundaries $0 \leq x \leq L$ and do not represent a computational problem. The incoming waves, $\lambda_i \geq 0$ at $x = 0$ and $\lambda_i \leq 0$ at $x = L$, depend on data external to the computational domain. Knowledge about the external solution is necessary to provide useful boundary conditions. Alternatively it may be desirable to use so-called non-reflective or radiation boundary conditions, which have the property of minimizing reflections from outgoing waves.

(**??**)

11.3 Non-Reflective Boundary Condition

This is by far the most imp[ortant boundary condition. Do not want to pollute the computational domain.

The interest in non reflective boundary conditions commenced in the early 1970s. The application of many of these techniques to nonlinear problems is still to be done[37]. We review the practical applicability of a number of these reflective boundary conditions for the solution of the shallow water wave equation used to rout floods in open channels. The review includes; (*i*) Sommerfeld radiation condition, (*ii*) Weakly reflective boundary condition based on the Method of Characteristics, and (*iii*) The introduction of a sponge layer outside the computational domain.

There are an number of successful methods for incorporating reflective boundary conditions for linear wave like equations. There are very few examples of the application of these approaches to nonlinear equations such as the Saint Venant. In this paper a number of approaches for incorporating non-reflective boundary conditions for the solution of the St. Venant equations are examined.

Generally, the application of non-reflective boundary conditions involves either (*i*) extending the computational domain (*nonlocal*) **sponge layer** or (*ii*) solving a equation, which may be a modified governing equations at the boundary (*local*). An alternative approach may include solving the problem on an infinite domain. The approached considered here include (*i*) co-ordinate transformation, (*ii*) Sommerfeld radiation condition. (*iii*) Applying modified forms of the governing equations on the boundary (*iv*)?????

Often we have artificial computational boundaries that arise simply because we can only solve the problem on a bounded domain. At such boundaries we often want to have no incoming signal, while there may ne out-going waves that should leave the domain cleanly without generating spurious reflections at the artificial boundary polluting the computational domain. We thus want a non-reflective boundary condition. At such boundaries we can often set ghost cell values by extrapolation from the interior solution. If the ghost cell value $\mathbf{U}_N^n$ is set based on $\mathbf{U}_{N-1}^n, \mathbf{U}_{N-2}^n, \ldots,$ then the new value $\mathbf{U}_{j+1}^{n+1}$ will effectively be computed on the basis of values to the left alone, even if the formula depends on $\mathbf{U}_N^n$, and hence this reduces to some sort of upwind method. The simplest approach is to use a *zero-order extrapolation*, meaning extrapolation by a constant function. We simply set

$$\mathbf{U}_{N+1}^n = \mathbf{U}_N^n, \quad \mathbf{U}_{N+2}^n = \mathbf{U}_N^n$$

Sponge Layer

There a number of approaches which employ a sponge layer to facilitate the application of non-reflective boundary conditions using a sponge or absorbing layer surrounding the computational domain. More recent inovations have been that proposed by Karni[61] which ensures that outgoing waves do not reach the boundary or that their amplitude is negligible when they do reach the boundary so that they do not have the opportunity to reflect and therefore prevent the generation

of incoming waves from the boundary.

This is achieved within a sponge layer surrounding the computational domain. Karni's method for constructing absorbing boundary conditions is not based on specific equations and therefore it can be applied to other types of wave propagation problems.

Karni proposes two filtering schemes one to annihilate the waves that approach a boundary and the other to gradually slow down the waves so that they do not reach the boundary and therefore cannot reflect from the boundary.

Solutions to (**??**) have the form (REFERENCE)

$$\mathbf{u}(x,t) = \sum_{k=1}^{n} \alpha_k (t - x/\lambda_k)\mathbf{r}_k$$

representing a superposition of travelling waves, $\mathbf{r}_k$ of strength α_k propagating at speeds λ_k. To prevent waves from travelling through the right-hand boundary, for example, (**??**) is replaced by;

$$\mathbf{u}_t + \mathbf{J}^* \mathbf{u}_x = \mathbf{0} \tag{175}$$

so that $\mathbf{J}^*$ contains reduced positive eigenvalues in the far field. Therefore, in the far field $\mathbf{J}^*$ has the eigenvalues;

$$\lambda_1 < \lambda_2 < ... < \lambda_m < 0 < a_{m+1}\lambda_{m+1} < a_{m+2}\lambda_{m+2} < ... < a_m \lambda_n$$

where, $0 \leq a_m \leq 1, k = m+1, ..., n$. This is the sponge layer. Karni showed that when

$$\mathbf{J}^* = \mathbf{R}\mathbf{\Lambda}^*\mathbf{L}$$

which is a unique matrix that will not generate partial reflections and preserves the eigenvectors of the matrix $\mathbf{A}$, where $\mathbf{\Lambda}^* = \text{diag}(\lambda_k^*)$, λ_k^* are the modified speed of the characteristics, λ_k ie. $\lambda_k^* = a_k \lambda_k$.

To prevent waves from reaching the right-hand boundary $a_k < 1$ and $d_k > 0$ for $k = 1$.

For the combined *slowing down* and *damping* operators, the following system of equations;

$$\mathbf{u}_t + \mathbf{J}^* \mathbf{u}_x + \mathbf{D}^*(\mathbf{u} - \mathbf{u}_0) = 0 \tag{176}$$

where

$$\mathbf{J}^* = \mathbf{R}\mathbf{\Lambda}^*\mathbf{L},\ \mathbf{\Lambda}^* = \text{diag}(\lambda_k^*),\ \mathbf{D}^* = \mathbf{R}\mathbf{\Delta}^*\mathbf{L},\ \mathbf{\Delta}^* = \text{diag}(d_k\lambda_k^*)$$

$\mathbf{u}_0$ are the vector of initial conditions. The parameters λ_k^* are the modified speed of the characteristics, λ_k i.e.. $\lambda_k^* = a_k\lambda_k$ and d_k are the attenuation rates. The decay rates $d_k \geq 0$ are chosen independently of the speed coefficients.

The reduction factor, d_k has the largest value on the grid boundary and it tapers gradually towards zero within the sponge layer. This gradual tapering is necessary to eliminate reflections.

For the absorbing boundary condition the parameter a_k differs from zero only in a strip of nodes, sponge layer surrounding the numerical mesh.

The application of the above will be demonstrated in the next section for a simple scalar wave equation and in Section 3. for the quasi-linear shallow water equation. $d = 0$, and $21\Delta x \leq a_1 = 1 - 0.01(N - x/\Delta x) \leq N\Delta x$.

Sommerfeld Radiation Condition

Sommerfeld radiation condition involves the solution of the hyperbolic equation

$$\frac{\partial \phi}{\partial t} + C_\phi \frac{\partial \phi}{\partial x} = 0 \tag{177}$$

where, ϕ is any dependant variable and C_ϕ the phase velocity for the quantity ϕ, and has only one characteristic is solved on the boundary. This is the exact local boundary condition for the second-order linear wave equation.

For other equations, including nonlinear equations it is an approximate boundary condition.

The major difficulty with this condition is that the phase speed C_ϕ is not known. There have been a number of approaches used to estimate C_ϕ. Chapman[16] reviews several boundary conditions based on the Sommerfeld radiation condition for a coastal ocean model. In these models the ocean bed slope has very little influence on the propagation of waves. These include; (i) gradient method, where $C_\phi = \infty$, (ii) gravity wave, where $C_\phi = \sqrt{gy}$, (iii) partially clamped, which is identical to the gravity wave with an *ad-hoc* friction term; $-\phi/T_f$, where T_f is a friction time scale, and (iv) Orlanski[92] estimate the phase speed using a finite difference analog of (177); $C_\phi = -\partial\phi/\partial t/\partial\phi/\partial x$. Implicit forms of these boundary conditions have been implemented.

(i) Gradient

$$\phi_N^{n+1} = \phi_{N-1}^{n+1}$$

(ii) Gravity wave

$$\phi_N^{n+1} = \frac{\phi_N^n + \mu\phi_{N-1}^{n+1}}{1+\mu}$$

where $\mu = C_\phi \Delta t/\Delta x$ and $C_\phi = \sqrt{gy}$. In the case of (**??**) $C_\phi = u + \sqrt{gy}$.

(iii) Partially clamped

$$\phi_N^{n+1} = \frac{\phi_N^n \left[1 - \frac{\Delta t}{T_f}\right] + \mu\phi_{N-1}^{n+1}}{1+\mu}$$

(iv) Orlanski radiation condition

$$\phi_N^{n+1} = \frac{\phi_N^{n-1}(1-\mu) + 2\mu\phi_{N-1}^n}{1+\mu}$$

where

$$\mu = \begin{cases} 1 & \text{if } C_\phi \geq 1 \\ C_\phi & \text{if } 0 < C_\phi < 1 \quad \text{and} \\ 0 & \text{if } C_\phi \leq 0 \end{cases}$$

$$C_\phi = \frac{\phi_{N-1}^{n-1} - \phi_{N-1}^{n+1}}{\phi_{N-1}^{n+1} + \phi_{N-1}^{n-1} - 2\phi_{N-2}^{n}}.$$

Characteristic Based Methods

Weakly Reflecting Boundary Conditions

The characteristics equations for (**??**) and (**??**) are[**?**];

$$\left.\begin{array}{c} dx/dt_+ = u + c \\ (Dy/Dt)_+ + (c/g)(Du/Dt)_+ + F_+ = 0 \end{array}\right\} \quad \text{along } C_+$$

$$\left.\begin{array}{c} dx/dt_- = u - c \\ (Dy/Dt)_- - (c/g)(Du/Dt)_- + F_- = 0 \end{array}\right\} \quad \text{along } C_-$$

in which $F_\pm = \mp cS_0 \pm cS_f$, C_+ is the rightward propagating characteristic, C_- is the leftward propagating characteristic and D/Dt stands for the differentiation along the characteristic

$$\frac{D}{Dt_\pm} = \frac{\partial}{\partial t} + (u \pm c)\frac{\partial}{\partial x}.$$

Using a first-order approximation, the characteristic equations become;

$$\begin{gathered} x_N - x_L = (u + c)_L(t_N - t_L), \\ x_N - x_R = (u + c)_R(t_N - t_R), \\ y_N - y_L + (c_L/g)(u_N - u_L) + (F_+)_L(t_N - t_L) = 0 \text{ and} \\ y_N - y_R + (c_R/g)(u_N - u_R) + (F_-)_R(t_N - t_R) = 0 \end{gathered}$$

where the subscripts L and R are the position where the characteristics intersect the previous time level, $n\Delta t$ and N is the location of the intersection of the two characteristics on the current time level $(n+1)\Delta t$. For a weakly reflecting downstream boundary condition[**?**] the C_- characteristic emanates from $x = \infty$ where steady uniform conditions are assumed to exist. Therefore, $u_R = u_\infty = u_0$, $y_R = y_\infty = y_0$, and $S_f = S_0$, hence $(F_-)_R = 0$. The characteristic equations become;

$$\begin{aligned} y_N - y_L + (c_L/g)(u_N - u_L) + (F_+)_L(t_N - t_L) &= 0 \\ y_N - y_0 + (c_0/g)(u_N - u_0) &= 0. \end{aligned}$$

Eliminating y_N then

$$u_N(c_L/g - c_0/g) - c_L/gu_L + c_0/gu_0 + y_0 - y_L + (F_+)_L\Delta t_N = 0$$

therefore

$$A_{0,N} = 0, B_{0,N} = c_L/g - c_0/g \text{ and}$$
$$G_{0,N} = c_L/gu_L - c_0/gu_0 - y_0 + y_L - (F_+)_L\Delta t_N = 0$$

where the variables at L are calculated using linear interpolation at

$$x_L = x_N - (u + c)_L\Delta t.$$

This boundary condition implies that steady uniform flow exits immediately downstream of the boundary condition. This is not correct, and is therefore considered as a weakly reflecting boundary condition.

Simple Wave Approach

Thomson[117] generalized the non-reflective boundary condition described by Hedstrom[49] which is applicable to nonlinear equations. Hedstroms non-reflective boundary condition relies on the fact that simple waves do not reflect off boundaries. In the simple wave region the Riemann invariant????, V_i remains constant with time along the characteristics in the xt plane defined by $dx/dt = \lambda_i$. It is the change in amplitude which indicates a wave. The function $\mathbf{V}$ is derived by multiplying (**??**) by $\mathbf{L}$ and calling that $\mathbf{\Lambda} = \mathbf{LAR}$, then;

$$\mathbf{L}\frac{\partial \mathbf{U}}{\partial t} + \mathbf{\Lambda L}\frac{\partial \mathbf{U}}{\partial x} + \mathbf{LS} = \mathbf{0}$$

or for each wave component

$$\mathbf{l}_i\frac{\partial \mathbf{U}}{\partial t} + \lambda_i\mathbf{l}_i\frac{\partial \mathbf{U}}{\partial x} + \mathbf{l}_i\mathbf{S} = \mathbf{0} \tag{178}$$

where

$$\mathbf{l}_1 = \begin{bmatrix} -u + c & 1 \end{bmatrix} \text{ and } \mathbf{l}_2 = \begin{bmatrix} u + c & -1 \end{bmatrix}.$$

The compatability equations for $\mathbf{U}$ is given by

$$d\mathbf{U} = \frac{\partial \mathbf{U}}{\partial x}dx + \frac{\partial \mathbf{U}}{\partial t}dt$$

then equation (178) becomes

$$\mathbf{l}_i d\mathbf{U} + \mathbf{l}_i\mathbf{S}dt = \mathbf{0}.$$

Let

$$dV_i = \mathbf{l}_i d\mathbf{U} + \mathbf{l}_i \mathbf{S} dt$$

and using the compatability equation for V_i then (178) becomes

$$\frac{\partial V_i}{\partial t} + \lambda_i \frac{\partial V_i}{\partial x} = 0$$

where $\lambda_i = dx/dt$. The quantity V_i is analogous to the Riemann invariants for the shallow water equations, it remains constant along the characteristic curve C_i in the xt plane which have the slope λ_i.

Since it is required that

$$\frac{\partial V_i}{\partial t} = 0$$

on the boundary, then

$$\mathbf{l}_i \frac{\partial \mathbf{U}}{\partial t} + \mathbf{l}_i \mathbf{S} dt = 0.$$

For (**??**) on the downstream boundary, this condition requires that

$$(-u + c + y)\left[\frac{\partial u}{\partial t} + (u + c)\frac{\partial u}{\partial x}\right] + u\left[\frac{\partial y}{\partial t} + (u + c)\frac{\partial y}{\partial x}\right] - gy(S_0 - S_f) = 0$$

remains constant.

Co-ordinate Transformation

By first mapping the semi-infinite region $0 \leq x < \infty$ into the finite region $0 \leq z < 1$ and solve the problem in the $z - t$ plane. The major advantage in this approach is that ??? the known boundary condition at infinity can be used in a finite domain model. Grosch and Orszag[41] describe an algebraic and exponential map. Only the exponential map

$$z = 1 - \exp(-\frac{x}{L})$$

where L is a constant sclae factor. A major advantage of this mapping is that it yields very simple expressions for the derivatives

$$\frac{\partial f(x,t)}{\partial x} = \frac{1 - z}{L}\frac{\partial f(z,t)}{\partial z}.$$

This transformation is applied to all the spatial derivatives in the governing equations which are discretized using thev Preissmann scheme and solved in the usual maner. The solution is transformed bach into the $x - t$ space using the inverse transformation.

12 Treatment of the Source Term using Fractional Steps

There are various ways to handle source terms. Fortunately, these fall into two categories; either *unsplit* or *fractional step* methods. In the unsplit method the full solution is advance in time using a single finite difference scheme. In naive finite difference schemes, such as first-order and second-order Lax-Wendroff and MacCormack schemes the introduction of a source term is straight forward. Unfortunately, these schemes either introduce excessive diffusion or dispersion in the solution. The source term in schemes which use some form of limiter or artificial viscosity contradicts the basic philosophy of these limiters. Flux limiter and total variation diminishing-type schemes seek to prevent an increase in the total variation of the numerical solution of the partial differential equations. However, for partial differential equations containing source terms an increase in the total variation may be the physical consequence of the source term. In addition, these limiters were designed for problems with weakly non-linear or non-stiff source terms. Algorithms for treating problems containing non-linear source terms is a new field of research[67]. A standard approach, which is also used here, is to solve the homogeneous part of the governing equations followed by the solution of the non-homogeneous term. In this fractional step method the best schemes can be used to solve each sub-problem.

Riemann solvers rely on the assumption that the characteristics are straight, which is no longer satisfied when source terms appear. A remedy for the Toro's scheme for a constant bed slope is given by Watson *et al.*[129].

The approach of separating the homogeneous problem and the source term is known as *Strang splitting or fractional step* [113] method. For the shallow water wave equations this involves solving the initial value problems

$$\frac{\partial \mathbf{U}}{\partial t} + \frac{\partial \mathbf{F}}{\partial x} = 0$$

with the initial conditions

$$\mathbf{U}(x,0) = \mathbf{U}_0(x) \Longrightarrow \mathbf{U}^*(x,t) \tag{179}$$

for the homogeneous problem, followed by the solution of the ordinary differential equation

$$\frac{d\mathbf{U}}{dt} = \mathbf{S}(\mathbf{U},t) \tag{180}$$

with the initial conditions

$$\mathbf{S}(0) = \mathbf{U}^*(x,t) \Longrightarrow \mathbf{U}(x,t)$$

for the source term. The splitting can be expressed as

$$\mathbf{U}^{n+1} = \mathbf{O}^{(\Delta t)}\mathbf{H}^{(\Delta t)}(\mathbf{U}^n)$$

in which $\mathbf{H}^{(t)}$ is the *homogeneous solution operator* and $\mathbf{O}^{(t)}$ is the *ordinary differential equation operator*. Any suitable numerical scheme can be used to solve the homogeneous problem. In certain problems (179) can be evaluated analytically. In many practical problems this will not be the case

and numerical techniques are required. Care must be exercised in solving the ordinary differential equation. If the ordinary differential problems contain processes with widely different time scales this will lead to a *stiff* ordinary differential equation. These can be resolved using implicit or adaptive ordinary differential equation solvers.

Equation (180) is only first-order in time, higher-order scheme can developed. For example splitting scheme

$$\mathbf{U}^{n+1} = \mathbf{O}^{(\Delta t/2)}\mathbf{H}^{(\Delta t)}\mathbf{O}^{(\Delta t/2)}(\mathbf{U}^n)$$

is second-order in time provided that both the homogeneous and ordinary differential equation solvers are at least second-order accurate. However, Pember[94] showed that for stiff problems, the use of higher-order Strang splitting will only produce a first-order accurate scheme. The difficulty lies in establishing when the equations are stiff.

Source terms can be treated fully implicitly. For example, (180) becomes

$$\frac{d\mathbf{U}}{dt} = \mathbf{S}(\mathbf{U}^{n+1}). \tag{181}$$

Expanding $\mathbf{S}(\mathbf{U}^{n+1})$ using a Taylor series about $\mathbf{S}(\mathbf{U}^n)$ by

$$\mathbf{S}(\mathbf{U}^{n+1}) = \mathbf{S}(\mathbf{U}^n) + \Delta\mathbf{U}\frac{\partial\mathbf{S}(\mathbf{U}^n)}{\partial\mathbf{U}} + O(\Delta\mathbf{U}^2) \tag{182}$$

where $\partial\mathbf{S}(\mathbf{U})/\partial\mathbf{U}$ is the Jacobian matrix of $\mathbf{S}(\mathbf{U})$ and $\Delta\mathbf{U} = \mathbf{U}^{n+1} - (U)^n$ is the change in the conserved variables.

Discretizing (181), then

$$\frac{\mathbf{U}^{n+1} - \mathbf{U}^n}{\Delta t} = \mathbf{S}(\mathbf{U}^{n+1}).$$

Substituting (182) then

$$\frac{\mathbf{U}^{n+1} - \mathbf{U}^n}{\Delta t} = \mathbf{S}(\mathbf{U}^n) + \Delta\mathbf{U}\frac{\partial\mathbf{S}(\mathbf{U}^n)}{\partial\mathbf{U}} + O(\Delta\mathbf{U}^2).$$

Re-arranging

$$\mathbf{U}^{n+1} = \frac{\mathbf{U}^n + \Delta t\mathbf{S}(\mathbf{U}^n) - \mathbf{U}^n\Delta t\dfrac{\partial\mathbf{S}(\mathbf{U}^n)}{\partial\mathbf{U}}}{\mathbf{I} - \Delta t\dfrac{\partial\mathbf{S}(\mathbf{U}^n)}{\partial\mathbf{U}}} + O(\Delta t\Delta\mathbf{U}^2). \tag{183}$$

For example, consider

$$\mathbf{S} = \begin{bmatrix} 0 \\ -u\tau \end{bmatrix}$$

and $\partial S_2/\partial u = -\tau$ and the first-order implicit treatment of the source term is given by

$$u^{n+1} = \frac{u^n}{1 + \tau\Delta t}. \tag{184}$$

The source term only affects the momentum equation. Therefore, h does not change and the source term only effects the velocity u. This is also the case for (**??**), where the friction slope S_f, expressed in terms of the conservative quantity uh is given by the Manning equation as

$$S_f = \frac{\eta^2 uh|uh|}{h^{10/3}}.$$

In contrast to the linear friction source term, the modulus of the conserved quantity in the Manning equation precudes its treament analytically or fully implicitly.

It has been explained and proven by Ambrosi???? that a simple explicit discretization of source terms leads to numerical instabilities. In this case the source term is

$$\mathbf{S} = \begin{bmatrix} 0 \\ ghS_0 - \dfrac{\eta g uh|uh|}{h^{7/3}} \end{bmatrix}$$

for a unit width of a wide rectangular channel. Splitting procedure is used to solve the non-homogeneous governing equations. The homogeneous equations are solved followed by the solution of the ordinary differential equation

$$\frac{d(uh)}{dt} = S_2$$

at the point (t^{n+1}, x_j) using a first-order semi-implicit scheme for the fiction term, so that it is discretized as

$$(uh)^{n+1} = (uh)^n - g\Delta t \frac{\eta^2 (uh)^{n+1}|(uh)^n|}{(h^{7/3})^n}$$

which becomes

$$(uh)^{n+1} = \frac{(uh)^n}{1 + g\Delta t \dfrac{\eta^2 |(uh)^n|}{(h^{7/3})^n}} \tag{185a}$$

is a stable treatment of the friction term as the denominator is always ≥ 0, and using an explicit scheme for the bed slope

$$(uh)^{n+1} = g\Delta t h^n S_0 \tag{185b}$$

where the values u^n and h^n are obtained from the solution of the homogeneous equations. This approach has proven to be an effective and stable treatment of the nonlinear friction term and has been used throughout this book.

Most real situations problems in one-dimensional river flow include source term quantities that should be considered when using a numerical scheme. Keeping the left hand side of (**??**) discretised exactly as for the homogeneous system, terms can be added to the right hand side of the scheme accounting for the source term $\mathbf{S}$. The *pointwise* descretisation of the source term shown is

$$\mathbf{U}_j^{n+1} = \mathbf{U}_j^n - \frac{\Delta t}{\Delta x}(\mathbf{F}_{j+1/2}^n - \mathbf{F}_{j-1/2}^n) + \Delta t \mathbf{S}_j^n \tag{186}$$

where $\mathbf{S}_j^n = \mathbf{S}(x_j, \mathbf{U}_j^n)$. This is the simplest possible discretisation of the source term and is often sufficient to produce satisfactory results. In this section a method is presented for discretising

the source term, when flux-difference techniques, such as the Roe and the Harten, Lax and van Leer contact approximate Riemann solvers are employed. The method we present here by Delis *et al.*[26] is a modification of that presented by Bermudez and Vazquez[7] for the shallow water wave equations. In order to obtain the numerical source, we project the source term onto the eigenvectors of $\mathbf{A}$ as follows

$$\begin{aligned}\mathbf{S} &= \mathbf{R}\beta = \mathbf{R}(\mathbf{\Lambda}\mathbf{\Lambda}^{-1})\beta = \mathbf{R}(\mathbf{\Lambda}^{+} + \mathbf{\Lambda}^{-})\mathbf{\Lambda}^{-}\beta \\ &= \mathbf{R}(\mathbf{\Lambda}^{+}\mathbf{\Lambda}^{-1})\beta + \mathbf{R}(\mathbf{\Lambda}^{-}\mathbf{\Lambda}^{-1})\beta\end{aligned}$$

where $\mathbf{R}$ is the matrix of the eigenvectors of $\mathbf{A}$ and $\beta = \beta(x, \mathbf{U})$.

Recalling that $\mathbf{\Lambda}^{+} = (\mathbf{\Lambda} + |\mathbf{\Lambda}|)/2$ and $\mathbf{\Lambda}^{-} = (\mathbf{\Lambda} - |\mathbf{\Lambda}|)/2$ then

$$\mathbf{R}(\mathbf{\Lambda}^{+}\mathbf{\Lambda}^{-1})\beta = \mathbf{R}(\mathbf{\Lambda}\mathbf{\Lambda}^{-1} + |\mathbf{\Lambda}|\mathbf{\Lambda}^{-1})\beta/2.$$

Since $\mathbf{S} = \mathbf{R}\beta$, $\mathbf{A}^{-1} = \mathbf{R}\mathbf{\Lambda}^{-1}\mathbf{R}$ and $|\mathbf{A}| = \mathbf{R}|\mathbf{\Lambda}|\mathbf{R}^{-1}$ then

$$\begin{aligned}\mathbf{R}(\mathbf{\Lambda}^{+}\mathbf{\Lambda}^{-1})\beta &= (\mathbf{S} + |\mathbf{A}|\mathbf{A}^{-1}\mathbf{R}\beta)/2 \\ &= (\mathbf{A} + |\mathbf{A}|)\mathbf{A}^{-1}\mathbf{S}/2 \\ &= \mathbf{A}^{+}\mathbf{A}^{-1}\mathbf{S} \qquad (187)\end{aligned}$$

which makes use of $\mathbf{A}^{+} = (\mathbf{A} + |\mathbf{A}|)/\mathbf{2}$. Similarly

$$\begin{aligned}\mathbf{R}(\mathbf{\Lambda}^{-}\mathbf{\Lambda}^{-1})\beta &= (\mathbf{A} - |\mathbf{A}|)\mathbf{A}^{-1}\mathbf{S}/2 \\ &= \mathbf{A}^{-}\mathbf{A}^{-1}\mathbf{S}. \qquad (188)\end{aligned}$$

Equations (187) and (188) suggest we introduce left and right numerical source term functions given by

$$\widetilde{\mathbf{S}}^{r,l}_{i\pm 1/2} = \mathbf{A}^{\pm}_{j\pm 1/2}\mathbf{A}^{-1}_{j\pm 1/2}\mathbf{S}_{j\pm 1/2}$$

where the terms $\mathbf{S}_{j\pm 1/2}$ represent approximations of source term at the cell interface between nodes $j-1, j$ and $j, j+1$ respectively. The projection is consistent with the conservation law since we have

$$\lim_{\mathbf{U}_k \to \mathbf{U}} (\widetilde{\mathbf{S}}^{l}_{j-1/2} + \widetilde{\mathbf{S}}^{r}_{j+1/2}) = \mathbf{A}^{+}_{j-1/2}\mathbf{A}_{j-1/2}\mathbf{S}_{j-1/2} + \mathbf{A}^{-}_{j+1/2}\mathbf{A}_{j+1/2}\mathbf{S}_{j+1/2}$$

for $k = i-1, i, i+1$. Assuming that

$$\lim_{\mathbf{U}_k \to \mathbf{U}} \mathbf{S}_{j\pm 1/2} = \mathbf{S}$$

and recalling that $\mathbf{A} = \mathbf{A}^{+} + \mathbf{A}^{-}$ then

$$\lim_{\mathbf{U}_k \to \mathbf{U}} (\widetilde{\mathbf{S}}_{j-1/2} + \widetilde{\mathbf{S}}_{j+1/2}) = \mathbf{S}.$$

Therefore, it is safe to approximate $\mathbf{S}_j^n$ in (186) with $\widetilde{\mathbf{S}}_j^n$, where

$$\widetilde{\mathbf{S}}_j^n = \widetilde{\mathbf{S}}_{j+1/2}^l + \widetilde{\mathbf{S}}_{j+1/2}^r.$$

In the above upwinded formulation of the source term $\mathbf{S}$, we need to evaluate $\widetilde{\mathbf{S}}_{j\pm1/2}$ approximations, that is some average of the source term on the left and right of the cell interfaces. A rather crude average is given by

$$\widetilde{\mathbf{S}}_{j\pm1/2} = \frac{\mathbf{S}_{j\pm1} + \mathbf{S}_j}{2}.$$

Delis *et al.*[26] used the averages of the wetted area, given by

$$\widetilde{A}_{j+1/2} = \sqrt{A_{j+1}A_j}$$

and the velocity

$$\widetilde{u}_{j+1/2} = \frac{\dfrac{Q_{j+1}}{\sqrt{A_{j+1}}} + \dfrac{Q_j}{\sqrt{A_j}}}{\sqrt{A_{j+1}} + \sqrt{A_j}} \tag{189}$$

which is the *square root averaging* of Roe and assuming that the channel is locally rectangular, then

$$\widetilde{\mathbf{S}}_{j\pm1/2} = \begin{bmatrix} 0 \\ g\dfrac{\widetilde{A}_{j\pm1/2}^2}{\widetilde{B}_{j\pm1/2}^2}\dfrac{\Delta B}{\Delta x} - g\widetilde{A}_{j\pm1/2}S_0 - g\widetilde{A}_{j\pm1/2}\widetilde{u}_{j\pm1/2}\eta^2\left(\dfrac{\widetilde{A}_{j\pm1/2}}{2\widetilde{A}_{j\pm1/2}/\widetilde{B}_{j\pm1/2} + \widetilde{B}_{j+1/2}}\right)^{-4/3} \end{bmatrix} \tag{190}$$

where the arithmetic mean for the average $\widetilde{B}_{j+1/2} = (B_{j+1/2} + B_j)/2$ has been used. This kind of discretisation for $\widetilde{\mathbf{S}}_{j\pm1/2}$ was originally suggested by Glaister[39], but he used the arithmetic mean to calculate $\widetilde{A}$. Approximation (158) of the source terms has been applied by Delis and Skeels[25]. An alternative treatment of the source term can be found in Glaister[39].

Consider a system of ordinary differential equations

$$\frac{\partial \mathbf{U}}{\partial t} = \mathbf{A}\mathbf{U}$$

with N unknowns. If λ denotes the eigenvalues of the matrix $\mathbf{A}$ and $\mathbf{e}$ its eigenvectors, then we have

$$\mathbf{A}\mathbf{e} = \lambda\mathbf{e}.$$

Usually, for and $N \times N$ matrix $\mathbf{A}$, there are N linearly independent eigenvectors $\mathbf{e}$ and a matrix composed of these N eigenvectors can be written as

$$\mathbf{T} = [\mathbf{e}_1 \cdots \mathbf{e}_N]$$

and we have

$$\mathbf{T}^{-1}\mathbf{A}\mathbf{T} = \mathbf{\Lambda}$$

where $\mathbf{\Lambda}$ is the diagonal matrix composed of eigenvalues

$$\mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_N \end{bmatrix}.$$

Thus we have

$$\begin{aligned} \mathbf{T}^{-1}\frac{\partial \mathbf{U}}{\partial t} &= \mathbf{T}^{-1}\mathbf{A}\mathbf{U} = \mathbf{T}^{-1}\mathbf{A}\mathbf{I}\mathbf{U} \\ &= \mathbf{T}^{-1}\mathbf{A}(\mathbf{T}\mathbf{T}^{-1})\mathbf{U} \\ &= (\mathbf{T}^{-1}\mathbf{A}\mathbf{T})\mathbf{T}^{-1}\mathbf{U} \\ &= \mathbf{\Lambda}\mathbf{T}^{-1}\mathbf{U}. \end{aligned}$$

If we define

$$\mathbf{W} = \mathbf{T}^{-1}\mathbf{U} \tag{191}$$

then we have

$$\frac{\partial \mathbf{W}}{\partial t} = \mathbf{\Lambda}\mathbf{W}$$

and hence

$$\frac{\partial W_i}{\partial t} = \lambda_i W_i \quad i = 1, \ldots, N$$

which is an uncoupled set of first-order ordinary differential equations and thus all the knowledge about single first-order ordinary differential equations can be directly applied to the above general system of ordinary differential equations. The solution to the above is given by

$$W_i = \alpha_i \exp(\lambda_i t)$$

where the constant coefficients α_i are determined by the initial conditions. Hence, using (191) we have

$$\mathbf{U} = \sum_{i=1}^{N} \mathbf{e}_i \alpha_i \exp(\lambda_i t)$$

It should be noted that $1/\lambda$ indicates the characteristic scales that control the solution profile for the system of equations. *Stiffness* corresponds to the situation in which eigenvalues of the system differ from each other by order of magnitude, thus leading to drastically different time or length scales in the problem. Examples of such stiff problems are some situations in chemical kinetics where the rate constants of chemical reaction of, say, two different species differ by six or seven orders of magnitude.

12.1 The Ordinary Differential Equation

Algorithms for treating problems containing non-linear source terms are a new field of research (Lafon and Yee[67]). A standard approach which is also used here to solve (**??**), is to solve the homogeneous part of the governing equations and solve the non-homogeneous term separately. In this way the most appropriate schemes can be used to solve each sub-problem. This approach of separating the homogenous problem and the source term is known as the *splitting* or *fractional step* (Strang[113]) method. There are numerous techniques for solving the ordinary differential equation operator (**??**). Care must be exercised in solving the ordinary differential equation. If the ordinary differential problems contain processes with widely different time scales, this will lead to a *stiff* ordinary differential equation. These can be resolved using implicit or adaptive ordinary differential equation solvers. However, Pember[94] showed that for stiff problems, the use of higher-order Strang splitting will only produce a first-order accurate scheme. Therefore, there may not be any advantages in implementing higher-order fractional step schemes for the solution of the shallow water wave equation which contains source terms. In this study, the ordinary differential equation operator is solved using the forward Euler method for the friction term and the backward Euler method for the source term which incorporates the influence of obstructions (Stoer and Bulirsch[111]). Since a source term only occurs in the momentum equation, it only affects the velocity.

Consider the following model problem

$$\frac{du(t)}{dt} + Au(t) = 0 \tag{192}$$

where A is a constant with $\Re(A) > 0$. This has a solution of the form

$$u(t) = \exp(-At)$$

with the initial condition $u(0) = u_0$. The model solution decays exponentials so that as $x \to \infty$ then $u(t) \to 0$.

For the explicit Euler scheme

$$\frac{du(t)}{dt} \cong \frac{u^{n+1} - u^n}{\Delta t}$$

and the model problem becomes

$$\begin{aligned} u^{n+1} &= (1 - A\Delta x)u^n \\ &\vdots \\ &= (1 - A\Delta t)^{n+1} u_0 \\ &\cong \exp(-At^n). \end{aligned}$$

Which is stable when the *amplification factor*

$$|1 - A\Delta t| < 1$$

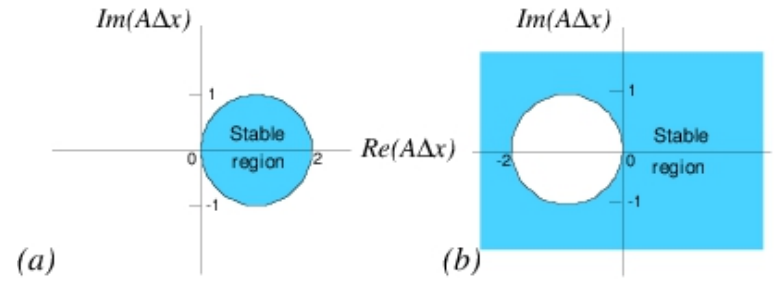


Figure 80: Stability region of the (a) explicit Euler and (b) implicit Euler schemes applied to (192).

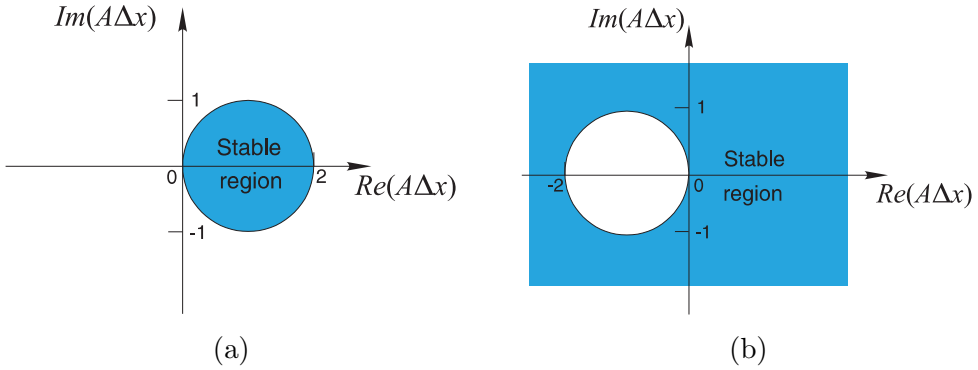


Figure 81: Stability region of the (a) explicit Euler and (b) implicit Euler schemes applied to (192).

so that the solution vanishes as $n \to \infty$ otherwise the errors grow exponentially and the solution is *divergent*. The region of stability is shown in Figure 81(a). Within this region, the solution is *convergent*, a meaningful solution is obtained. However, truncation errors may have an influence on the accuracy of the results. Outside this region the numerical scheme is unstable.

For the implicit Euler scheme, the spatial derivative is approximated by

$$\frac{du(t)}{dt} \cong \frac{u^n - u^{n-1}}{\Delta t}$$

and the model problem becomes

$$\begin{aligned} u^{n+1} &= (1 + A\Delta t)^{-1} u^{n-1} \\ &\vdots \\ &= \frac{1}{(1 + A\Delta t)^n} u_0 \\ &\cong \exp(-At^n). \end{aligned}$$

The solution vanishes when $j \to \infty$ if

$$|1 + A\Delta t| > 1$$

and the stability region is shown in Figure 81(a). Since $\Re(A) > 0$ and $\Delta x \geq 0$ then the scheme is always stable.

13 The Finite Volume Method

An alternative strategy to the use of operator splitting on a Cartesian grid when dealing with two-dimensional problems, is to use the *finite volume method*. Although it is considered computationally more expensive than the use of operator splitting on a Cartesian grid, its major advantage is in being able to readily handle complex geometry using arbitrary shaped polygons to discretize the computational domain. In general, triangular elements are usually employed. Its major advantages are; (*i*) the underlying principle is simple, (*ii*) flexible meshes, such as triangles or quadrilaterals which suit problems with complex geometries can be employed, and (*iii*) an integral conservation laws is used, therefore the solution may be either smooth or discontinuous.

The finite volume method was introduced first in the field of numerical fluid dynamics independently by McDonald[81] and MacCormack and Paullay[80] for the solution of two-dimensional, time-dependent Euler equations (Hirsch[51]). The finite volume method has been used to solve the two-dimensional shallow water wave equations by Anastasiou and Chan[3], Alcrudo and Garcia-Navarro[2], Zhao *et al.*[138] and by Zoppou and Roberts[139], Kim *et al.*[64], Lai *et al.*[68].

13.1 Basic Scheme

The finite volume method is based on the integration of conservation equation

$$\frac{\partial \mathbf{U}}{\partial t} + \frac{\partial \mathbf{E}(\mathbf{U})}{\partial x} + \frac{\partial \mathbf{G}(\mathbf{U})}{\partial y} = \mathbf{S}. \tag{193}$$

The discretization of the integral form of (193)) ensures that the basic quantities, mass and momentum will also be conserved across a discontinuity (see, for example Hirsch[51], p. 237) and the solution will admit shock waves and contact surfaces.

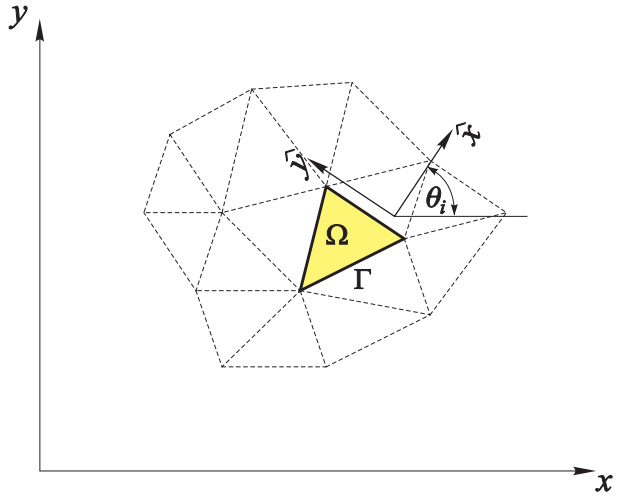


Figure 82: Descretization of the computational domain in two-dimensions using triangular elements in the finite volume method.

Integrating (193) over an arbitrary triangular cell, which has an area Ω and perimeter Γ shown in Figure 82 then

$$\int_\Omega \frac{\partial \mathbf{U}}{\partial t}\, d\Omega + \int_\Omega \frac{\partial \mathbf{E}(\mathbf{U})}{\partial x} + \frac{\partial \mathbf{G}(\mathbf{U})}{\partial y}\, d\Omega = \int_\Omega \mathbf{S}\; d\Omega. \tag{194}$$

Applying Gauss' divergence theorem

$$\int_\Omega \nabla \cdot \mathbf{F}\, d\Omega = \oint_\Gamma \mathbf{F} \cdot \mathbf{n}\, d\Gamma$$

to the second term on the left side of this equation on any region Ω enclosed by the contour Γ, then

$$\oint_\Gamma \mathbf{F} \cdot \mathbf{n}\, dL = \int_\Omega \frac{\partial \mathbf{E}(\mathbf{U})}{\partial x} + \frac{\partial \mathbf{G}(\mathbf{U})}{\partial y}\, d\Omega \tag{195}$$

and (194) becomes

$$\frac{\partial}{\partial t} \int_\Omega \mathbf{U}\, d\Omega + \oint_\Gamma \mathbf{F} \cdot \mathbf{n}\, d\Gamma = \oint_\Omega \mathbf{S}\, d\Omega \tag{196}$$

where $d\Gamma$ is the length of a boundary element with an arc $d\Omega$ and $\mathbf{n}$ is the outward unit vector normal, $\mathbf{n} \equiv [n_1, n_2] = [\cos(\theta), \sin(\theta)]$ to the boundary. This law states that the time rate of change in $\mathbf{U}$ inside a *control volume* Ω depends on the *total flux* through the surface Γ minus the sum of sources within the control volume.

Assuming the x-direction as the reference direction and using the *rotational invariance* property of the shallow water wave equation, the flux becomes

$$\mathbf{F} \cdot \mathbf{n} \equiv [\mathbf{E}(\mathbf{U}), \mathbf{G}(\mathbf{U})] \cdot \mathbf{n} \equiv \cos(\theta)\mathbf{E}(\mathbf{U}) + \sin(\theta)\mathbf{G}(\mathbf{U}] \equiv \mathbf{T}^{-1}\mathbf{E}(\mathbf{T}(\mathbf{U})) \tag{197}$$

where $\mathbf{T}$ is the transformation matrix

$$\mathbf{T} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos(\theta) & \sin(\theta) \\ 0 & -\sin(\theta) & \cos(\theta) \end{bmatrix}.$$

Equation (196) becomes, with the use of (197)

$$\frac{d\mathbf{U}}{dt} + \frac{1}{A} \int_\Gamma \mathbf{T}^{-1}\mathbf{E}(\hat{\mathbf{U}})\, d\Gamma = \mathbf{S} \tag{198}$$

where the area of the triangular element is A and its perimeter is Γ. The new conserved quantities

$$\hat{\mathbf{U}} = \mathbf{T}(\mathbf{U}) = \begin{bmatrix} h \\ \hat{u}h \\ \hat{v}h \end{bmatrix}$$

where $\hat{u} = u\cos(\theta) + v\sin(\theta)$ and $\hat{v} = -u\sin(\theta) + v\cos(\theta)$ are aligned with the rotated axis $\hat{x}$ and $\hat{y}$, see Figure 82 and

$$\mathbf{E}(\hat{\mathbf{U}}) = \begin{bmatrix} \hat{u}h \\ \hat{u}h + \frac{1}{2}gh^2 \\ \hat{u}\hat{v}h \end{bmatrix} \tag{199}$$

is the flux across the boundary separating two neighbouring elements.

Written in expanded form, the equations to be solved across each cell interface are

$$\frac{\partial h}{\partial t} + \frac{\partial(\hat{u}h)}{\partial x} = 0 \tag{200}$$

$$\frac{\partial(\hat{u}h)}{\partial t} + \frac{\partial\left(\hat{u}^2h + \frac{1}{2}gh^2\right)}{\partial x} + gh(S_{0_x} - S_f) = 0 \tag{201}$$

$$\frac{\partial(\hat{v}h)}{\partial t} + \frac{\partial(\hat{u}\hat{v}h)}{\partial x} = 0. \tag{202}$$

The first two equations are just the one-dimensional shallow water wave equations with source terms and the third equation defines the contact discontinuity. The one-dimensional shallow water wave equations are de-coupled from the third equation, which is simply an advection equation for $\hat{v}h$, which propagates with speed $\hat{u}$. It has the same form as (200) and could be considered as the conservation of the $\hat{y}$-directional momentum in the $\hat{x}$-direction and the can be treated numerical in the same way as the continuity equation.

A number of authors (for example, Causon *et al.*[13] and Molls and Molls[83]) either ignore or neglect the contact discontinuity. However, this equation is required for improved stability (Le Veque[75] and Toro *et al.*[120]) and to obtain full accuracy in a second- or higher-order scheme (Le Veque[75]).

Therefore, the Riemann problem for the two-dimensional problem is identical to the one-dimensional Riemann problem given by (112) and (113) and the same approximate one-dimensional Riemann solver can be used for both problems. The only additional consideration required for the two-dimensional problem is the treatment of the contact discontinuity.

The problem now involves the solution of a local one-dimensional problem in the direction normal to the element interface, where $\mathbf{U}_l$ and $\mathbf{U}_r$ are the states in neighboring elements with common boundaries. Recalling that within each cell $\mathbf{U}$ is constant, then this will usually result in a discontinuity across the surface $S_{i,j}$.

There are a number of schemes that can be used to define this flux. A wide range of numerical schemes have been examined for the solution of the one-dimensional shallow water wave equation applied to problems with discontinuities in the solution.

Therefore, it is only necessary to estimate the flux normal to the surface Γ which is the solution of a local one-dimensional Riemann problem using the $\hat{x}$-component of the flux of the transformed conservative quantities.

The two-dimensional problem can be dealt with as the solution of an augmented one-dimensional problems in the transformed $(\hat{x}, \hat{y})$ domain. The system is strictly *hyperbolic* with three real and distinct eigenvalues, $\lambda_1 = \hat{u} + \sqrt{gh}$, $\lambda_2 = \hat{u} - \sqrt{gh}$ and $\lambda_3 = \hat{v}$.

The wave structure for the two-dimensional Riemann problem normal to a cell interface is shown in Figure 83 The Riemann problem can be decomposed into three waves for the two-dimensional shallow water wave equations. The middle wave is always a contact discontinuity and the left and right waves can be either shocks or rarefaction fan. Classic approximate Riemann solvers assume that the solution of the one-dimensional SWWE consists of only two waves separating three constant state regions. For the two-dimensional shallow water wave equation however, there is a third wave, which is a contact discontinuity.

Toro *et al.*[120] showed how the contact discontinuity can be incorporated into the Harten, Lax and van Leer approximate Riemann solver.

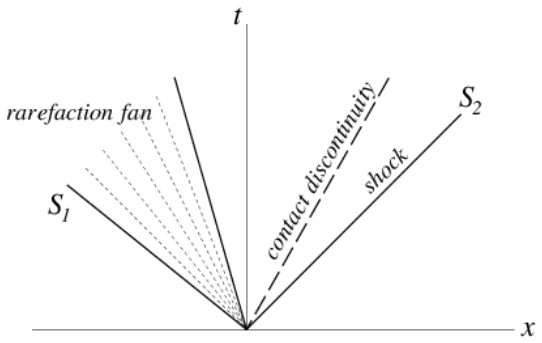


Figure 83: Shock, rarefaction fan and contact discontinuity for the augmented one-dimensional problem associated with the two-dimensional dam break problem with finite water depth everywhere.

Using operator splitting, the problem can be treated as two augmented problems involving the solution of an homogeneous operator and an ordinary differential equation operator sequentially.

The subsequent sections deal with the solution of the hyperbolic homogeneous operator only. Equation (202) can be solved using a suitable upwind scheme, for the two-dimensional problem. Equation (202) can be treated in the same way as the continuity equation, (200), since (202) is the continuity equation for the momentum in the y-direction.

If the ordinary differential problems contain processes with widely different time scales this will lead to a *stiff* ordinary differential equation. Stiffness is sometimes defined in terms of the ratio $\max|\lambda_i| / \min|\lambda_i|$ over all eigenvalues (LeVeque[76]). If this is large, then there is a large range of time scales in the problem. There are obvious problems with this expression, in particular at sonic points where one of the eigenvalues vanishes and for systems which only have one characteristic. However, a stiff problem can be resolved using an implicit solver. Fortunately, the shallow water wave equations, (112) is not considered as stiff.

The local initial value problem which solve the shallow water wave equations, involving dis-

continuous neighbouring states is known as the *Riemann* problem. Numerical schemes based on the solution of local Riemann problems are generally known as *Godunov-type* schemes. Their main advantages are that they are robust and accurately capture the location of discontinuities such as shocks and contact discontinuity[120]. Two-dimensional Riemann solvers do not appear to have matured enough to be used in the construction of multi-dimensional schemes. Even if such solvers were available, the resulting schemes are likely to be too complicated for common use[8]. Recently, a number of efficient one-dimensional *approximate Riemann* solvers have been proposed [101, 29, 93, 33, 8]. These can be used to solve the two-dimensional dam break problem on a Cartesian grid using *fractional step* or *splitting techniques* [113]. The *finite volume method* [81], can also be used to solve the two-dimensional problem, see for example Soulis[108], Zhao *et al.*[138], Zhao *et al.*[137], Erduran *et al.* [30] and Anastasiou and Chan[3]. Like finite elements, the major advantage of the finite volume method is that it can be applied to any unstructured grid. Generally, the finite volume method requires less computational effort than finite elements because it involves the solution of a local one-dimensional Riemann problem, which can be solved efficiently.

13.2 Side Channel Friction

Working with structured meshes, Molls *et al.*[84] modified the friction slope equation distributing the side-wall friction across the width of the channel. Every cell augments its wetted area in the form $S_i + (h_1 + h_N)/(N-1)$ where N is the number of computational nodes across the channel, h_1 and h_N are the water depths at the channel side-walls. Moreover, in the Manning equation a weighting factor is included that accounts for side-wall friction and can be viewed as a correction to the Manning coefficient η. On unstructured meshes, there is no row of cells numbered from 1 to n in cross-section and the above modification is not valid. The idea is to include the effect of the side-wall friction in the computation of the total friction term in each cell, developing on the number of edges that form part of a vertical solid wall and computing the real wetted area (bottom and wall), see Figure 84.

Two different type of cells are distinguished, internal cells, where the edges are adjacent to another cell, and boundary cells, where at least one of the edges is a vertical wall. For internal cells (**??**) are still appropriate, where η_b is the Manning resistance coefficient for the bottom of the cell. For boundary cells, the wetted area must be specified. First all, the total number of edges in a cell, which are part of a solid vertical wall is needed. The wetted area is $S_i + l_{w_k} h$, see Figure 84, where S_i is the area of the cell, l_{w_k} is the length of the solid wall edge involved with this cell, and h is the water depth. The Manning equation becomes, for the x-direction

$$S_{f_x} = \frac{\eta^2 u \sqrt{u^2 + v^2}}{R^{4/3}}$$

which gives the following expressions for the friction term (Brufau and Garcia-Navarro[11])

$$S_{f_x} = u\sqrt{u^2+v^2}\left(\frac{\eta_b^{2/3}}{h}+\sum_{w_k=1}^{\text{SWN}}\frac{\eta_w^{3/2}l_{w_k}}{S_i}\right)^{4/3} \quad \text{and}$$

$$S_{f_y} = v\sqrt{u^2+v^2}\left(\frac{\eta_b^{2/3}}{h}+\sum_{w_k=1}^{\text{SWN}}\frac{\eta_w^{3/2}l_{w_k}}{S_i}\right)^{4/3}$$

with η_w is the Manning resistance coefficient for the vertical walls. Numerical experiment were used by Brufau and Garcia-Navarro[11] to verify the above relationship.

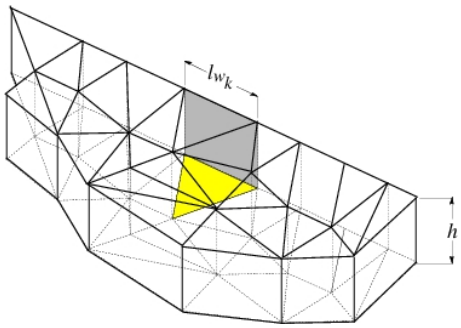


Figure 84: Treatment of friction at a wall in two-dimensional flow for an unstructured grid.

.

Contact discontinuity across which the density is again discontinuous but the velocity and water depth are constant

.

All of the state variables are continuous and there is a smooth transition. This is called a *rarefaction wave* since the water depth of the gas decreases as this wave passes through.

13.3 Wetting and Drying of Cells

Brufau and Garcia-Navarro[10] developed a simple procedure for treating the wetting and drying front. Begnudelli and Sanders[5] simple algebraic equations that relate fluid volume to the free surface elevation in partially submerged cells.

Bed elevation of the centroid is, $z_c = (z_1 + z_2 + z_3)/3$. $\eta = h + z_c$ is the stage. h is the water depth in the cell. It is assumed that $z_1 \leq z_2 \leq z_3$. If $\eta \leq z_1$ the cell is dry, so that $h = 0$. If $z_1 \leq \eta \leq z_2 \leq z_3$ then the ratio of the volume of fluid, V to the cell area, A denoted by $h = V/A$

$$h = \frac{(\eta - z_1)^3}{3(z_2 - z_1)(z_3 - z_1)}$$

or

$$\eta = z_1 + \sqrt[3]{3h(z_2 - z_1)(z_3 - z_1)}.$$

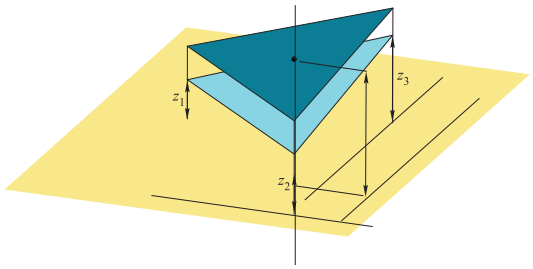
If $z_2 < \eta \leq z_3$ then

$$h = \frac{\eta^2 + \eta z_3 - 3\eta z_1 - z_3 z_2 + z_1 z_2 + z_1^2}{3(z_3 - z_1)}$$

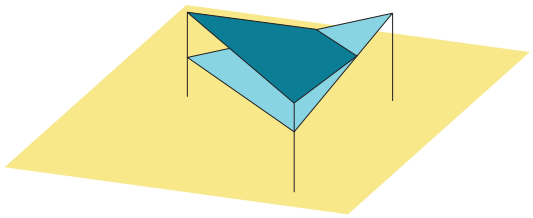
or

$$\eta = \frac{-\gamma_1 + \sqrt{\gamma_1^2 - 4\gamma_2}}{2}$$

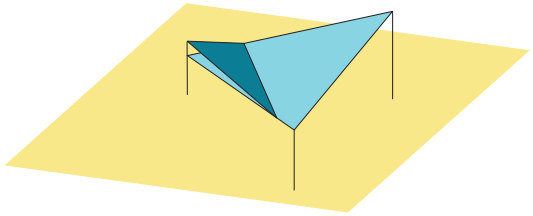
where $\gamma_1 = z_3 - 3a_1$ and $\gamma_2 = 3hz_1 - 3hz_3 - z_3 z_2 + z_1 z_2 + z_1^2$. If $z_3 < \eta$ then $\eta = h + z_c$.



(a)



(b)



(c)

Figure 85: Partial wetting and drying of a cell ?????.

References

[1] M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions*. Dover Publications, New York, 1970.

[2] F. Alcrudo and P. Garcia-Navarro. A high-resolution godunov-type scheme in finite volumes for the 2d shallow water equations. *International Journal for Numerical Methods in Fluids*, 16(6):489–505, 1993.

[3] K. Anastasiou and C. T. Chan. Solution of the 2d shallow water equations using the finite volume method on unstructured triangular meshes. *International Journal for Numerical Methods in Fluids*, 24(11):1225–1245, 1997.

[4] R. Ansorge. Survey of equivalence theorems in the theory of difference approximations for pdes. In J.J.H. Miller, editor, *Topics in Numerical Analysis III*, pages 1–16. Academic Press, London, 1977.

[5] L. Begnudelli and B. F. Sanders. Unstructured grid finite-volume algorithm for shallow-water flow and scalar transport with wetting and drying. *Journal of Hydraulic Engineering, American Society of Civil Engineering*, 132(4):371–384, 2006.

[6] A. Bermúdez and M. E. Vázquez. Flux-vector and flux-difference splitting methods for the shallow water equations in a domain with variable depth. In P.W Partridge, editor, *Computer Modelling of Seas and Coastal Regions*, pages 256–267. Computational Mechanics Publications, Elsevier Applied Science, London, 1992.

[7] A. Bermúdez and M. E. Vázquez. Upwind methods for hyperbolic conservation laws with source terms. *Computers and Fluids*, 23(8):1049–1071, 1994.

[8] S. J. Billett and E. F. Toro. On waf-type schemes for multidimensional hyperbolic conservation laws. *Journal of Computational Physics*, 130(1):1–24, 1997.

[9] J. P. Boris and D. L. Book. Flux-corrected transport i shasta, a fluid transport algorithm that works. *Journal of Computational Physics*, 11(1):38–69, 1973.

[10] F. Brufau and P. Garcia-Navarro. Unsteady free surface flow simulation over complex topography with multidimensional upwind technique. *Journal of Computational Physics*, 186(2):503–526, 2003.

[11] P. Brufau and P. Garcia-Navarro. Two-dimensional dam break flow simulation. *International Journal for Numerical Methods in Fluids*, 33(1):35–57, 2000.

[12] P. G. Buning and J. L. Steger. Solution of the two-dimensional euler equations with generalized coordinate transformations using flux vector splitting. Technical report, American Institute of Aeronautics and Astronautics, 1982. Report 82–0971.

[13] D. M. Causon, C. G. Mingham, and D. M. Ingram. Advances in calculation methods for supercritical flow in spillway channels. *Journal of Hydraulic Engineering, American Society of Civil Engineers*, 125(10):1039–1050, 1999.

[14] S. R. Chakravarthy and S. Osher. Computing with high resolution upwind schemes for hyperbolic equations. *Lectures in Applied Mathematics*, 22:57–86, 1985.

[15] S. C. Chang. The method of space-time conservation element and solution element: A new approach for solving the navier-stokes and euler equations. *Journal of Computational Physics*, 119(2):295–324, 1995.

[16] D. C. Chapman. Numerical treatment of cross-shelf open boundaries in a barotropic coastal ocean model. *Journal of Physical Oceanography*, 15(8):1060–1075, 1985.

[17] Y. Chen and R. A. Falconer. Advection-diffusion modelling using the modified quick scheme. *International Journal for Numerical Methods in Fluids*, 15(10):1171–1196, 1992.

[18] R. C. Y. Chin and G. W. Hedstrom. A dispersion analysis for difference schemes: Tables of generalized airy functions. *Mathematics of Computation*, 32(144):1163–1170, 1978.

[19] P. Collela and P. Woodward. The piecewise-parabolic method (ppm) for gas dynamical simulations. *Journal of Computational Physics*, 54(1):174–201, 1984.

[20] G. Comini, M. Manzan, and C. Nonino. Analysis of finite element schemes for convection-type problems. *International Journal for Numerical Methods in Fluids*, 20(6):443–458, 1995.

[21] R. Courant and K. O. Friedrichs. *Supersonic Flow and Shock Waves*. Interscience, New York, 1948.

[22] R. Courant, K. O. Friedrichs, and H. Lewy. Über die partiellen differenzen-gleichungen der mathematischen physik. *Mathematische Annalen*, 100:32–74, 1928.

[23] M. G. Crandall and A. Majda. The method of fractional steps for conservation laws. *Numerische Mathematik*, 34(3):285–314, 1980.

[24] S. F. Davis. A simplified tvd finite difference scheme via artificial viscosity, society for industrial and applied mathematics. *Journal on Scientific and Statistical Computing*, 8(1):1–18, 1987.

[25] A. I. Delis and C. P. Skeels. Tvd schemes for open channel flow. *International Journal for Numerical Methods in Fluids*, 26(7):791–809, 1998.

[26] A. I. Delis, C. P. Skeels, and S. Ryrie. Evaluation of some approximate riemann solvers for transient open channel flows. *Journal of Hydraulic Research*, 38(3):217–231, 2000.

[27] D. R. Durran. *Numerical Methods for Wave Equations in Geophysical Fluid Dynamics*. Text in Applied Mathematics. Springer-Verlag, New York, 1998.

[28] B. Engquist and A. Majda. Absorbing boundary conditions for the numerical simulation of waves. *Mathematics of Computation*, 31(139):629–651, 1977.

[29] B. Engquist and S. Osher. One sided difference approximations for nonlinear conservation laws. *Mathematics of Computation*, 36(154):321–351, 1981.

[30] K. S. Erduran, V. Kutija, and C. J. M. Hewett. Performance of finite volume solutions to the shallow water equations with shock-capturing schemes. *International Journal for Numerical Methods in Fluids*, 40(10):1237–1273, 2002.

[31] R. J. Fennema and M. H. Chaudhry. Explicit numerical schemes for unsteady free-surface flows with shocks. *Water Resources Research*, 22(13):1923–1930, 1986.

[32] R. J. Fennema and M. H. Chaudhry. Explicit methods for 2-d transient free-surface flows. *Journal of Hydraulic Engineering, American Society of Civil Engineers*, 116(8):1013–1034, 1990.

[33] L. Fraccarollo and E. F. Toro. Experimental and numerical assessment of the shallow water model for two-dimensional dam-break type. *Journal of Computational Physics*, 33(6):843–864, 1995.

[34] P. Garcia-Navarro and J. M. Saviron. 1-d open-channel flow simulation using tvd-mccormack scheme. *Journal of Hydraulic Engineering, American Society of Civil Engineers*, 118(10):1359–1372, 1992.

[35] P. Garcia-Navarro and M. E. Vazqez-Cendon. On numerical treatment of the source terms in the shallow water equations. *Computers and Fluids*, 29(8):951–979, 2000.

[36] A. Gharangik and M. H. Chaudhry. Numerical simulation of hydraulic jump. *Journal of Hydraulic Engineering, American Society of Civil Engineers*, 117(9):1195–1211, 1995.

[37] D. Givoli. Non-reflective boundary conditions. *Journal of Computational Physics*, 94(1):1–29, 1993.

[38] P. Glaister. Approximate riemann solutions of the shallow water equations. *Journal of Hydraulic Research*, 26(3):293–306, 1988.

[39] P. Glaister. Flux difference splitting for open-channel flows. *International Journal for Numerical Methods in Fluids*, 16(7):629–654, 1993.

[40] S. K. Godunov. A finite-difference method for the numerical computation of discontinuous solutions of the equations of fluid dynamics. *Matematicheskii Sbornik*, 47(89):271–306, 1959.

[41] C. E. Grosch and S. A. Orszag. Numerical solution of problems in unbounded regions: Coordinate transforms. *Journal of Computational Physics*, 25(1):273–296, 1977.

[42] A. Harten. High resolution schemes for hyperbolic conservation laws. *Journal of Computational Physics*, 49(3):357–393, 1983.

[43] A. Harten. Eno schemes with subcell resolution. *Journal of Computational Physics*, 83(1):148–184, 1989.

[44] A. Harten and J. M. Hyman. Self adjusting grid methods for one-dimensional hyperbolic conservation laws. *Journal of Computational Physics*, 50(2):235–269, 1983.

[45] A. Harten, P. D. Lax, and A. van Leer. On upstream differencing and godunov-type scheme for hyperbolic conservation laws. *Society for Industrial and Applied Mathematics, Review*, 25(1):235–261, 1983.

[46] A. Harten and S. Osher. Uniformly high-order accurate nonoscillatory schemes, i. *Society for Industrial and Applied Mathematics, Journal on Numerical Analysis*, 24(2):279–309, 1987.

[47] A. Harten, S. Osher, B. Engquist, and S. R. Chakravarthy. Some results on uniformly high-order accurate essentially nonoscillatory schemes. *Applied Numerical Mathematics*, 2(3):347–377, 1986.

[48] G. W. Hedstrom. Models of difference for $u_t + u_x = 0$ by partial differential equations. *Mathematics of Computation*, 29(132):969–977, 1975.

[49] G. W. Hedstrom. Nonreflecting boundary conditions for nonlinear hyperbolic systems. *Journal of Computational Physics*, 30(2):222–237, 1979.

[50] C. Hirsch. *Numerical Computation of Internal and External Flows, Volume 1, Fundamentals of Numerical Discretization.* John Wiley and Sons, New York, 1988.

[51] C. Hirsch. *Numerical Computation of Internal and External Flows, Volume 2: Computational Methods for Inviscid and Viscous Flows.* John Wiley and Sons, Chitchester, 1990.

[52] C. W. Hirt. Heuristic stability theory for finite-difference equations. *Journal of Computational Physics*, 2(4):339–355, 1968.

[53] F. M. Holly and A. Preissmann. Accurate calculation of transport in two dimensions. *Journal of the Hydraulics Division, American Society of Civil Engineers*, pages 1259–1277, 1977.

[54] John Jakeman. On numerical solutions of the shallow water wave equations. Honours thesis, ANU, October 2006.

[55] A. Jameson, W. Schmidt, and E. Turkel. Numerical solutions of the euler equations by finite volume methods using runge-kutta time-stepping schemes. In *, American Institute ofAeronautics and Astronautics, Fluid and Plasma Dynamics Conference, Palo Alto, California*, pages Paper 81–1259, 1981.

[56] A. K. Jha, J. Akiyama, and M. Ura. First- and second-order flux difference splitting schemes for dam-break problem. *Journal of Hydraulic Engineering, American Society of Civil Engineers*, 121(12):877–884, 1995.

[57] A. K. Jha, J. Akiyama, and M. Ura. Flux-splitting schemes for 2d flood flows. *Journal of Hydraulic Engineering, American Society of Civil Engineers*, 126(1):33–42, 2000.

[58] G. S. Jiang and E. Tadmor. Nonoscillatory central schemes for multidimensional hyperbolic conservation laws. *Society for Industrial and Applied Mathematics, Journal on Scientific Computing*, 19(6):1892–1917, 1998.

[59] M. Jin and D. L. Fread. Dynamic flood routing with explicit and implicit numerical solution schemes. *Journal of Hydraulic Engineering, American Society of Civil Engineers*, 123(3):166–173, 1997.

[60] S. K. Kar and R. P. Turco. Formulation of a lateral sponge layer for limited-area shallow-water models and an extension for the vertically stratified case. *Monthly Weather Review*, 123(5):1542–1559, 1995.

[61] S. Karni. Far-field filtering operators for suppression of reflections from artificial boundaries. *Society for Industrial and Applied Mathematics, Journal on Numerical Analysis*, 33(3):1014–1047, 1996.

[62] W. J. and Karplus. *Analog Simulation: Solution of Field Problems*. McGraw-Hill, New York, 1958.

[63] J. Kevorkian. *Partial Differential Equations: Analytical Solution Techniques*. Chapman-Hall, New York, 1989.

[64] D. H. Kim, Y. S. Cho, and W. G. Kim. Weighted averaged flux-type schemes for shallow-water equations with fractional step method. *Journal of Engineering Mechanics, American Society of Civil Engineers*, 130(2):152–160, 2004.

[65] R. Kosloff and D. Kosloff. Absorbing boundaries for wave propagation problems. *Journal of Computational Physics*, 63(2):363–376, 1986.

[66] A. Kurganov, S. Noelle, and G. Petrova. Semidiscrete central-upwind schemes for hyperbolic conservation laws and hamilton-jacobi equations. *Society for Industrial and Applied Mathematics, Journal on Scientific Computing*, 23(3):707–740, 2001.

[67] A. Lafon and H. C. Yee. Dynamical approach study of spurious steady-state numerical solutions of nonlinear differential equations part iv. stability vs methods of discretizing nonlinear source terms in reaction-convection equation. *Computational Fluid Dynamics*, 6(2):89–123, 1996.

[68] J. J. Lai, G. F. Lin, and W. D. Guo. An upstream flux-splitting finite-volume scheme for 2d shallow water equations. *International Journal for Numerical Methods in Fluids*, 48(10):1149–1174, 2005.

[69] P. D. Lax. Weak solutions of nonlinear hyperbolic equations and their numerical computation. *Communications on Pure and Applied Mathematics*, 7(1):159–193, 1954.

[70] P. D. Lax. Hyperbolic systems of conservation laws, ii. *Communications on Pure and Applied Mathematics*, 10(4):537–566, 1957.

[71] 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.

[72] P. D. Lax and B. Wendroff. System of conservation laws. *Communications on Pure and Applied Mathematics*, 13(2):217–237, 1960.

[73] P. D. Lax and B. Wendroff. Difference methods with high order of accuracy for solving hyperbolic equations. *Communications on Pure and Applied Mathematics*, 17(3):381–398, 1964.

[74] R. J. Le Veque. *Numerical Methods for Conservation Laws*. Lectures in Mathematics. ETH Zürich, Birkhäuser Verlag, Basel, 1992.

[75] R.J. Le Veque. Balancing source terms and flux gradients in high-resolution godunov methods: The quasi-steady wave-propagation algorithm. *Journal of Computational Physics*, 146(1):346–365, 1998.

[76] R.J. Le Veque. Finite difference methods for differential equations. Unpublished lecture notes, University of Washington, 1998.

[77] S. M. Liang and J. J. Chang. An imploved upwind scheme for the euler equations. *Journal of Computational Physics*, 84(2):461–473, 1989.

[78] K. A. Lie and S. Noelle. On the artificial compression method for second-order nonoscillatory central difference schemes for systems of conservation laws, society for industrial and applied mathematics. *SIAM Journal on Scientific Computation*, 24(4):1157–1174, 2003.

[79] R. W. MacCormack. The effect of viscosity in hypervelocity impact cratering. Technical report, American Institute of Aeronautics and Astronautics, 1969. Paper 69-354.

[80] R. W. MacCormack and A. J. Paullay. Computational efficiency achieved by time splitting of finite difference operators. Technical report, American Institute of Aeronautics and Astronautics, 1972. Paper 72-154.

[81] P. W. McDonald. The computation of transonic flow through two-dimensional gas turbine cascades. Technical Report Paper 71-GT-89, American Society of Mechanical Engineers, 1971.

[82] C. G. Minghan and D. M. Causon. High resolution finite-volume method for shallow water flows. *Journal of Hydraulic Engineering, American Society of Civil Engineers*, 124(6):605–614, 1998.

[83] T. Molls and F. Molls. Space-time conservation method applied to saint venant equations. *Journal of Hydraulic Engineering, American Society of Civil Engineers*, 124(5):501–508, 1998.

[84] T. Molls, G. Zhao, and F. Molls. Friction slope in depth-averaged flow. *Journal of Hydraulic Engineering, American Society of Civil Engineers*, 124(1):81–85, 1998.

[85] L. A. Monthé, F. Benkhaldoun, and I. Elmahi. Positivity preserving finite volume roe schemes for transport-diffusion equations. *Computer Methods in Applied Mechanics and Engineering*, 178(3-4):215–232, 1999.

[86] T. Myint-U and L. Debnath. *Patrial Differential Equations for Scientists and Engineers.* North-Holland, New York, 1987.

[87] T. Nakatani and S. Komura. A numerical simulation of flow with hydraulic jump using tvd-maccormack scheme. *Proceedings of XXV Congress of International Association for Hydraulic Research*, 1:9–13, 1993.

[88] B. J. Noye. Three-point two-level finite difference methods for the one-dimensional advection equation. In B. Noye and R. May, editors, *Computational Techniques and Applications,*, pages 159–192. Elsevier Science Publishers, North-Holland, 1986.

[89] B. J. Noye and K. Hayman. Accurate finite difference methods for solving the advective-diffusion equation. In B. Noye and R. Mays, editors, *Computational Techniques and Applications*, pages 137–157. Elsevier Science Publishers, North-Holland, 1986.

[90] B. J. Noye and P. Steinle. Improved 5-point implicit finite difference methods for solving the 1-dimensional advection equation. In B. Noye and R. Mays, editors, *Computational Techniques and Applications*, pages 193–204. Elsevier Science Publishers, North-Holland, 1986.

[91] M. Nujic. Efficient implementation of non-oscillatory schemes for the computation of free-surface flows. *Journal of Hydraulic Research*, 33(1):100–111, 1995.

[92] I. Orlanski. A simple boundary condition for unbounded hyperbolic flows. *Journal of Computational Physics*, 21(3):251–269, 1976.

[93] S. Osher and F. Solomon. Upwind difference schemes for hyperbolic conservation laws. *Mathematics of Computation*, 38(158):339–374, 1982.

[94] R. B. Pember. Numerical methods for hyperbolic conservation laws with stiff relaxation ii. higher-order godunov methods. *Society for Industrial and Applied Mathematics, Journal on Scientific Computing*, 14(4):824–859, 1993.

[95] S. Piperno and S. Depeyre. Criteria for the design of limiters yielding efficient high resolution tvd schemes. *Computers and Fluids*, 27(2):183–197, 1998.

[96] M. Rahman and M. H. Chaudhry. Simulation of dam-break flow with grid adaptation. *Advances in Water Resources*, 21(1):1–9, 1989.

[97] R. D. Richtmyer and K. W. Morton. *Difference Methods for Initial-Value Problems*. Wiley and Sons, New York, 1967.

[98] W. J. Rider. Methods for extending high-resolution schemes to non-linear systems of hyperbolic conservation laws. *International Journal for Numerical Methods in Fluids*, 17(10):861–885, 1993.

[99] A. Ritter. Die fortpflanzung de wasserwellen. *Zeitschrift Verein Deutscher Ingenieure*, 36(33):947–954, 1892.

[100] P. Roache. *Computational Fluid Dynamics*. Hermosa Publishers, Albuquerque, 1985.

[101] P. L. Roe. Approximate riemann solvers, parameter vectors, and difference schemes. *Journal of Computational Physics*, 43(2):357–372, 1981.

[102] P. L. Roe and J. Pike. Efficient construction and utilization of approximate riemann solutions. In R. Glowinski and J.L. Lions, editors, *Computing Methods in Applied Science and Engineering, VI, Eds.,*, pages 499–518. North-Holland, Amsterdam, 1984.

[103] L. J. Savic and F. M. Holly. Dambreak flood waves computed by modified godunov method. *Journal of Hydraulic Research*, 31(2):187–204, 1993.

[104] Y. I. Shokin. *The Method of Differential Approximation.* Springer, New York, 1983.

[105] C. W. Shu. Tvb uniformly high-order schemes for conservation laws. *Mathematics of Computation*, 49(179):105–121, 1987.

[106] C. W. Shu. Total-variational-diminishing time discretizations. *Society for Industrial and Applied Mathematics, Journal on Scientific and Industrial Computing*, 9(6):1073–1084, 1988.

[107] C. W. Shu and S. Osher. Efficient implementation of essentially non-oscillatory shock-capturing schemes, ii,. *Journal of Computational Physics*, 83(1):32–78, 1989.

[108] J. V. Soulis. Computation of two-dimensional dam-break flood flows. *International Journal for Numerical Methods in Fluids*, 14(6):631–664, 1992.

[109] J. L. Steger and R. F. Warming. Flux vector splitting of the inviscid gas dynamic equations with application to finite-difference methods. *Journal of Computational Physics*, 40(2):263–293, 1981.

[110] S. K. Stein. *Calculus and Analytical Geometry.* McGraw-Hill, New York, 1987.

[111] J. Stoer and R. Bulirsch. *Introduction to Numerical Analysis.* Springer-Verlag, New York, 1992.

[112] J. J. Stoker. *Water Waves, The Mathematical Theory with Applications.* Interscience, London, 1957.

[113] G. Strang. On the construction and comparison of finite difference schemes. *Society for Industrial and Applied Mathematics, Journal on Numerical Analysis*, 5(3):506–517, 1968.

[114] J. C. Strikwerda. *Finite Difference Schemes and Partial Differential Equations.* Wadsworth, Belmont, California, 1989.

[115] P. K. Sweby. High resolution schemes using flux limiters for hyperbolic conservation laws. *Society for Industrial and Applied Mathematics, Journal on Numerical Analysis*, 21(5):995–1011, 1984.

[116] P. Tamamidis and D. N. Assanis. Evaluation of various high-order-accuracy schemes with and without flux limiters. *International Journal for Numerical Methods in Fluids*, 16(10):931–948, 1993.

[117] W. K. Thompson. Time dependent boundary conditions for hyperbolic systems. *Journal of Computational Physics*, 68(1):1–24, 1987.

[118] E. F. Toro. Riemann problems and the waf method for solving the two-dimensional shallow water equations. *Philosophical Transactions of the Royal Society of London, Series A*, 338:43–68, 1992.

[119] E. F. Toro. *Riemann Solvers and Numerical Methods for Fluid Dynamics*. Springer-Verlag, Berlin, 1997.

[120] E. F. Toro, M. Spruce, and W. Speares. Restoration of the contact surface in the hll-riemann solver. *Shock Waves*, 4(1):25–34, 1994.

[121] L. N. Trefethen. Group velocity in finite difference schemes. *Society for Industrial and Applied Mathematics, Review*, 24(2):113–137, 1982.

[122] M. H. Tseng. Explicit finite volume non-oscillatory schemes for 2d transient free-surface flows. *International Journal for Numerical Methods in Fluids*, 30(7):831–843, 1999.

[123] GD van Albada, B. van Leer, and WW Roberts Jr. A comparative study of computational methods in cosmic gas dynamics. *Astronomy and Astrophysics*, 108(1):76–84, 1982.

[124] B. van Leer. Towards the ultimate conservative difference scheme, i the quest of monotonicity. *Lecture Notes in Physics*, 18:163–168, 1973.

[125] B. van Leer. Towards the ultimate conservative difference scheme, iv a new approach to numerical convection. *Journal of Computational Physics*, 23(3):276–298, 1977.

[126] B. van Leer. Towards the ultimate conservative difference scheme, v a second-order sequel to godunov's method. *Journal of Computational Physics*, 32(1):101–136, 1979.

[127] R. F. Warming and R. M. Beam. Upwind second-order schemes and applications in aerodynamic flows. *American Institute of Aeronautics and Astronautics, Journal*, 14(9):1241–1249, 1976.

[128] R. F. Warming and B. J. Hyett. The modified equation approach to the stability and accuracy of finite-difference methods. *Journal of Computational Physics*, 14(2):159–179, 1974.

[129] G. Watson, D. H. Peregrine, and E. F. Toro. Numerical solution of the shallow-water equations on a beach using the weighted average flux method. In Periaux Hirsch and Cordulla, editors, *Proceedings of the First European Computational Fluid Dynamics Conference '92*, pages 495–502. Elsevier Science Publishers, Amsterdam Hirsch, 1992.

[130] G. B. Whitham. *Linear and Nonlinear Waves*. John Wiley and Sons, New York, 1974.

[131] C. Wu, G. Huang, and Y. Zheng. Theoretical solution of dam-break shock wave. *Journal of Hydraulic Engineering, American Society of Civil Engineers*, 125(11):1210–1215, 1999.

[132] H. Q. Yang and A. J. Prezekwas. A comparative study of advanced shock-capturing schemes applied to burger's equation. *Journal of Computational Physics*, 102(1):139–159, 1992.

[133] J. Y. Yang, C. A. Hsu, and S. H. Chang. Computations of free surface flows part 1: One-dimensional dam-break flow. *Journal of Hydraulic Research*, 31(1):19–34, 1993.

[134] H. C. Yee. Construction of explicit and implicit symmetric tvd schemes and their applications. *Journal of Computational Physics*, 68(1):151–179, 1987.

[135] H. C. Yee. A class of high-resolution explicit and implicit shock-capturing methods. Technical Report Technical Memorandum 101088, National Aeronautics and Space Administration, Ames Research Center, Moffett Field, California, 222p, 1989.

[136] S. T. Zalesak. Fully multidimensional flux-corrected transport algorithms for fluids. *Journal of Computational Physics*, 31(3):335–362, 1979.

[137] D. H. Zhao, H. W. Shen, J. S. Lai, and G. Q. Tabios. Approximate riemann solvers in fvm for 2d hydraulic shock wave modeling. *Journal of Hydraulic Engineering, American Society of Civil Engineers*, 122(12):692–702, 1996.

[138] D. H. Zhao, H. W. Shen, G. Q. Tabios, J. S. Lai, and W. Y. Tan. Finite-volume two-dimensional unsteady-flow model for river basins. *Journal of Hydraulic Engineering, American Society of Civil Engineers*, 120(7):863–883, 1994.

[139] C. Zoppou and S. Roberts. Catastrophic collapse of water supply reservoirs in urban areas. *Journal of Hydraulic Engineering, American Society of Civil Engineers*, 125(7):686–695, 1999.

[140] C. Zoppou and S. Roberts. Numerical solution of the two-dimensional unsteady dam break. *Applied Mathematical Modelling*, 24(7):457–475, 2000.

Index

accuracy
- first-order, 46
- round-off error, 53
- second-order, 47, 84
- truncation error, 53

global, 52
local, **52**
Airy function, **40**
artificial diffusion, 93
artificial viscosity, 69, 84, 85

boundary condition, 144
- Dirichlet, 147
- absorbing, **145**
- actual, **145**
- artificial, **145**, 145
- free-space, **145**
- internal, **145**, 145
- non-reflective, **145**, 148
- non-reflective, Sommerfield, 150
- non-reflective, characteristic based, 151
- non-reflective, co-ordinate transformation, 153
- non-reflective, simple wave, 152
- non-reflective, sponge layer, 148
- non-reflective, weakly reflective, 151
- one-way, **145**
- open, **145**
- radiating, **145**
- silent, **145**
- transmissive, 144
- transmitting, **145**
- transparent, **145**

Burgers equation, 12

chain rule, 103
channel
- non-prismatic, **122**
- non-uniform, **122**

characteristic, 59
- constant state region, 25
- domain of dependence, 60
- simple wave region, 25, 27

complementary error function, **38**
conservation laws
- integral form, 33

consistency, 55
consistency condition, **92**, 94
contact discontinuity, 164, 166, 167
convergence, 55
Courant, R., **37**
Courant-Friedrichs-Lewy
- condition, **60**
- number, 37, 47, 53, 88, 133, 134

critical
- flow, 94

dam break, 86
- dry bed, 30

dimensional splitting
- operator, **141**

dissipation, 84, 88

eigenvalue, 98, 165
eigenvector
- left, 104
- right, 106

entropy
- fix, 93, 125
- violation, 92

equation
- Burgers
  - inviscid, 33, 91
- Euler, 90

- Korteweg-de Vries, linear, 40
- advection, 35, 46, 47, 164
- advective-diffusion, 38
- advective-dispersion, 42, 47
textbf, 40
- conservative, 33
- homogeneous, 143, 165
- degree one, 90
- degree two, 91
- order one, 91
- hyperbolic, 165
- modified equivalent partial differential equation, 36, **37**, 46
- non-homogeneous, 143
- shallow water wave, 91, 165
- Jacobian, 90, 98
- stiff, **159**

essentially non-oscillatory, **71**
Euler's formula, **43**
Euler's identity, **91**, 95

finite difference, 35
- accuracy, 49
- backward, 49
- backward Euler, 160
- centered, 49
- conservative form, 33
- consistency, 49, 53
- convergence, 49, 53
- efficiency, 55
- explicit, **56**
- explicit Euler, 160
- first-order, 46, 50, 65
- forward, 49
- forward Euler, 160
- forward time centered space, 57
- forward time upwind, 58
- implicit, **56**
- implicit Euler, 161
- local truncation error, **57**
- monotone, 64
- non-conservative form, 33
- second-order, 47, 50
- stability, 49, 54
- truncation error, 53

flux, 67, 68
flux splitting, 91
Fourier analysis, 35
Fourier series, 54
Fourier transform, **38**, 40
fractional step, 166
Friedrichs, K.O., **37**
Froude number, 30

Gibb's phenomenon, 36
Godunov scheme, 99
Godunov, S.K., **61**
grid
- unstructured, 166

Hugoniot locus, 110
hydraulic jump, 27

initial value problem, 165
intermediate region, 107
interpolation
- linear, 100
- piecewise constant, 100
- quartic, 136

Lax entropy condition, **108**
Lax's equivalence theorem, 50, **55**
Lewy, H., **37**
limiter, 84, 85
- MUSCL, **74**, 85, 89
- Piperno and Depeyre, 82
- SLIP, 76
- flux, 69, 84

- minmod, **74**, **76**, 76, 84, 85, 87, 89, 134
- monotonic, **74**, 85, 89
- slope, 76, 84
- superbee, **74**, 85, 89, 134
- total variation diminishing, 83
- universally non-oscillatory, 83
- van Albada, **74**, 76
- van Leer, **74**, 76

matrix
- Jacobian, 103
- commute, **142**
- diagonal, 159

mean value theorem, 64, **97**
method
- finite volume, **162**

modified equivalent partial differential equation, 36, **37**

norm
- L_1, 46, 47

numerical diffusion, 37, 45
numerical dispersion, 37, 45
numerical scheme
- Beam-Warming, 37, 42, 59, 67, **69**, 69
- Bermudez-Vazquez, 67, 90, **91**, 92
- Godunov, 99, 166
- Godunov scheme
  - averaging, 99
  - cell-averaging step, 99
  - evolution step, 99, 102
  - re-averaging, 99, 100
  - reconstruction step, 99
- Harten, 67, 90, **97**, 98
- Harten, Lax and van Leer, 68, 128, 165
- Harten, Lax and van Leer Contact**,** 128, 131
- Holly and Priessmann, 45, 47
- Lax-Friedrichs, 37, 50, **53**, 57, **65**, 65, 67, 70
- Lax-Wendroff, 37, 40, 47, 50, **53**, 58, **65**, 65, 67, 69, 71, 73, 84, 85, 87
- MacCormack, 50, 67, **69**, 69, 73, 85, 86, 88
- Osher's P-ordering, 102, **126**, 127
- Richtmyer, **65**, 65, 73
- Roe, 68, 102, 122
- Roe's, **121**
- Steger-Warming, 67, 85, 87, 89, 90, **94**
- TVD Lax-Wendroff, 67
- Yang *et al.*, 67, 90, **96**
- accuracy, 35, 36
  - near a discontinuity, 46
- approximate Riemann solver, 166
- biased average procedure, 67
- central, 68
- conservative, 33
- essentially non-oscillatory, 68
- exact Riemann solver, **107**
- finite element, 166
- finite volume, 166
- first-order, 73
- flux corrected transport, 85, **87**, 87
- flux vector splitting, 90
- forward time backward space, **53**
- forward time upwind, 39, 70
- kinetic, 68
- monotone, **61**
- piecewise parabolic method, **135**, 135
- second-order, 73
- total variation diminishing, **70, 71**, 85–89, 133
- unstable, 67, 73
- upwind, 37, 47, **53**, 73, 85, 90
- upwind central, 68
- weighted average flux, 68, **132**, 132

oscillations, 36

Rankine-Hugoniot condition, 98
rarefaction
    - fan, 25, 28, 30, 167
    - shock, 110
Riemann invariants, 110
Riemann problem, 30, 99, 107, 110, 113, 122, 166
rotational invariance, 163

Schwartz inequality, 63
shock, 25, 166
    - expansive, 94, 110
shock-wave, 12
similarity solution, 102
sonic point, 85, 88, 98
source term, 156
splitting, **141**, 166
    - Strang, 142
    - dimensional, **141**
    - fractional, **142**
    - operator, **141**, 165
stability, 55, 164
    - discrete perturbation analysis, **54**
    - heuristic analysis, **54**
    - matrix method, **54**
    - von Neumann, **54**
subcritical flow, 85, 88
supercritical flow, 88
Sweby diagram, **78**

Taylor series, 36, **50**, 50, 55, 62, 141, 143
total variation, **70**
total variation bounded, **70**

wave
    - front, **107**
wave amplitude, **121**
wave strength, **121**
waves
    - back, 107
    - front, 107