UNIVERSITY OF READING
DEPARTMENT OF MATHEMATICS

ANALYSIS AND COMPUTATION OF STEADY OPEN CHANNEL FLOW

by

IAN MACDONALD

This thesis is submitted for the degree of
Doctor of Philosophy

SEPTEMBER 1996

Abstract

The Saint-Venant equations provide a one-dimensional model of free surface water flow in a channel. This thesis is concerned with both analytical and numerical aspects of steady state solutions to this model, with particular emphasis on the subject of transcritical flows.

Under certain conditions it is shown that there is at most one physically allowable steady solution for given boundary conditions, and when a solution exists, we demonstrate the convergence of a certain family of numerical methods to the solution as the grid size vanishes.

The numerical schemes are obtained from applying a family of monotone shock capturing schemes to a scalar conservation law which has identical steady solutions to the Saint-Venant model. We generalise this "scalar approach" to include other scalar shock capturing schemes and compare the performance in terms of accuracy and efficiency against more established methods. Methods of further increasing the computational efficiency of the "scalar approach" are also considered.

To assess the accuracy of the different numerical methods we can compare the numerical solutions against the exact solutions for a series of test cases. However, other than for some idealised situations, there appear to be no such test cases in the literature. We describe a relatively simple method that allows the construction of a wide range of test problems with known exact solutions, including solutions having multiple transitions. Numerical results from the numerical schemes are compared with the exact solutions.

Acknowledgements

Firstly I would like to thank my academic supervisors, Prof. M.J. Baines and Prof. N.K. Nichols, and industrial supervisor Dr. P.G. Samuels (HR Wallingford Ltd.) for their help and guidance throughout this Ph.D.

I would like to thank my family for their love and support throughout my time at university.

I am grateful to Ed Dicks for proof reading this Thesis and for giving many useful comments.

Finally I acknowledge a CASE studentship from the EPSRC and HR Wallingford Ltd. who were the industrial CASE organisation.

Contents

Notation for the Saint-Venant Model

x	Distance along channel (m)
t	Time (s)
L	Length of channel (m)
z_b	Bed level (m)
η	Height relative to the bed level (m)
σ	Width of channel as a function of x and η (m)
g	Acceleration due to gravity (ms^{-2})
ρ	Density (kgm^{-3})
h	Depth (m)
Q	Discharge ($\mathrm{m}^3\mathrm{s}^{-1}$)
A	Wetted area (m^2)
T	Free surface width (m)
P	Wetted perimeter (m)
$F = \frac{Q^2}{A} + gI_1$	Momentum flux per unit density ($\mathrm{m}^4\mathrm{s}^{-2}$)
$I_1 = \int_0^h (h-\eta)\sigma d\eta$	Hydrostatic pressure term (m^3)
$D = gA(S_0 - S_f) + gI_2$	Source term ($\mathrm{m}^3\mathrm{s}^{-2}$)
$S_0 = -z_b'$	Bed slope
$S_f = \frac{\lvert Q\rvert Q}{K^2}$	Friction slope
K	Conveyance ($\mathrm{m}^3\mathrm{s}^{-1}$)
n	Friction coefficient
$I_2 = \int_0^h (h-\eta)\sigma_x d\eta$	Side reaction term for a non-prismatic channel (m^2)
$u = \frac{Q}{A}$	Component of fluid velocity in x direction (ms^{-1})
$c = \left(\frac{gA}{T}\right)^{\frac{1}{2}}$	Wave celerity (ms^{-1})
h_c, h_n	Critical and normal depths (m)
S_{0c}	Critical bed slope
$F_r = \left(\frac{Q^2 T}{gA^3}\right)^{\frac{1}{2}}$	Froude number
$E = \frac{Q^2}{2A^2} + gh$	($\mathrm{m}^2\mathrm{s}^{-2}$)
B, Z	Width (m) and side slope for a trapezoidal channel

$\boldsymbol{w} = (A, Q)^T$, $\boldsymbol{F} = (Q, F)^T$, $\boldsymbol{D} = (0, D)^T$

Chapter 1

Introduction

The study of free-surface water flow in channels has many important applications, one of the most significant being in the area of river modelling. With major river engineering projects, such as flood prevention measures, becoming ever more common and ambitious, there is an increasing need to be able to model and predict the far ranging consequences on the environment as a whole of any potential project. A major part of this process is to predict the new hydraulic characteristics of the system. For example constricting the river at some point may result in an increased risk of flooding at a point upstream. The basic equations expressing hydraulic principles were formulated in the 19th century by de St Venant and Boussinesq. Properties of these relationships were studied in the first half of this century, but application to real river engineering projects awaited the advent of electronic computers. The hydraulic equations are also of great importance in the modelling and design of networks of artificial channels, as for example may occur in industrial plants or sewage systems.

The original hydraulic model of de St Venant[11] is written in the form of a system of two partial differential equations, known as the Saint-Venant equations. These are derived under the hypothesis that the flow is one-dimensional. One-dimensional flows do not actually exist in nature, but the equations remain valid provided the flow is approximately one-dimensional. Until recently, two or three-dimensional models have been too computationally expensive to be practical. Even now it is often prohibitively expensive to obtain the amount of survey data for a river network necessary to make

use of the added realism of a higher dimensional model. For this reason the bulk of river modelling still makes use of a one-dimensional model, with key parts of the network perhaps modelled with a higher-dimensional model. Empirical correction factors are often included in the one-dimensional model to correct for deviations away from one-dimensionality. A real river model would also include many other effects not taken account of in the basic Saint-Venant model, for example flow to and from flood plains and sediment transport. Such effects are discussed in the important text on river modelling of Cunge, Holly and Verwey[9].

This thesis is concerned only with the basic Saint-Venant model where, for given initial data, the system of differential equations may at some point in time fail to have a solution. When solutions break down we are forced to turn to the more general integral formulation of the model. This form admits discontinuous solutions, where a discontinuity represents a region of flow where the flow variables change relatively rapidly, and is known as a hydraulic bore.

Very often in nature a flow will approach a steady state, that is where the flow is essentially unchanging in time. The study of steady flow is therefore an important subject in hydraulics. Under the assumption of steady flow the Saint-Venant equations reduce in complexity yielding a single nonlinear ordinary differential equation which describes the variation of the free surface. This equation has been much studied, for example in the classic hydraulics text of Chow[4]. As in the unsteady case the differential form does not describe all solutions and we need to refer to the more general formulation which admits discontinuous solutions.

The subject of this thesis is the steady flow problem, i.e. the problem of determining the steady state flow without regard for the transient behaviour. In order for this approach to be feasible there must be at most one steady solution for given boundary conditions, otherwise we could not determine which of the steady flows actually occurs. If one considers only smooth solutions given by the steady flow differential equation, then there can clearly be at most one solution for given boundary conditions. However if discontinuous solutions are also considered, then it is possible that there may be more than one solution; in fact we illustrate such a case. In Chapter 4 we present theory which shows that under certain conditions there is at

most one physically possible steady solution for any given boundary conditions. The proof relies on a novel formulation of the steady flow problem, with the solutions constructed as the vanishing viscosity limit of solutions to a singular perturbation problem. Properties of the smooth solutions of the singular perturbation problem give information about the not necessarily continuous solutions of the steady flow problem.

As well as theoretical results for the steady flow problem, this thesis is also concerned with numerical computation of solutions. The steady flow differential equation can be accurately and efficiently integrated in order to compute the free surface profile. This is in general only useful for computing smooth solutions, although Humpidge and Moss[26] present an algorithm for discontinuous solutions which works by computing smooth surface profiles and fitting discontinuities at the appropriate locations. A more conventional approach for computing steady solutions is to numerically model the transient flow in time until the numerical solution attains a steady state. Current commercial packages for modelling time dependent flows are often based on the Preissman box scheme or other similar schemes (see refs. [9], [60], [61] and [52]). Such methods are accurate and computationally efficient for the gradually varying flows which on the whole make up most of river flows, but break down with the formation of a hydraulic bore. For a scheme to be capable of computing discontinuous solutions we require that, as well as approximating the Saint-Venant differential equations, the scheme approximates the more general integral form of the model. The scheme must also be stable in the presence of discontinuities. A scheme which has these properties is known as a shock capturing scheme. The schemes used in commercial codes are not in general shock capturing schemes.

Chapter 3 discusses the subject of shock capturing schemes and in particular the scheme of Roe[55] which has been applied to the Saint-Venant equations by many authors, for example in [2] and [53]. This scheme is effective at modelling discontinuous solutions, although as we demonstrate it has some difficulties. In particular it can be computationally inefficient compared to the more traditional schemes because of a severe restriction on the allowed time step. Attempts to improve the efficiency of the scheme are made in [17].

In this thesis we attempt a new approach for improving efficiency of the computation of steady solutions. Instead of applying shock capturing method to the Saint-Venant system as it stands, we apply shock capturing methods to a suitable scalar partial differential equation which is chosen so as to have identical steady solutions to the Saint-Venant model. The first benefit of this approach (which we refer to as the "scalar approach") is that analysis for scalar methods is much simpler than for the case of systems, and in Chapter 5 we present theory for a particular family of schemes. Under identical conditions to the theory in Chapter 4 we show that at steady state the system of difference equations has a unique solution and we also demonstrate convergence to the unique physical solution of the steady flow problem (as the grid spacing vanishes).

In Chapter 6 we give a relatively simple technique for constructing test problems with known exact solutions. Although analytic solutions have previously been constructed for idealised problems, this appears to be the first time that solutions for problems with realistic features have been made available[38]. Such features include varying channel geometries and discontinuous solutions. Details are given for a wide selection of test cases so as to allow other research workers to test and compare their own numerical methods.

At the start of Chapter 7 we apply some of the numerical methods satisfying the theory in Chapter 5 to a selection of the test problems discussed above. We assess the usefulness of certain a-priori estimates arising from the theory. We next compare the methods in terms of accuracy against the scheme of Roe and then generalise the "scalar approach" in order to achieve higher-order accuracy schemes.

The "scalar approach" leads to a system of nonlinear difference equations and one way to compute a solution is through a time stepping iteration which effectively models to steady state the transient behaviour of a scalar partial differential equation. In Chapter 8 we investigate other possible methods for solving the difference equations, in terms of efficiency and robustness, including Newton's method.

The numerical methods discussed in the main part of this thesis are only applicable to prismatic channels where the cross-section of the channel does not vary throughout its length. In Chapter 9 we consider an extension of both the scalar

schemes and Roe's scheme to the case of non-prismatic channels and compare the accuracy of the various schemes.

Chapter 2

The Saint-Venant Equations

In this chapter the Saint-Venant equations are introduced and some of their properties discussed. Attention is then fixed on the steady state form of these equations which are the main subject of this thesis. Background information is given on the steady flow problem.

2.1 The Unsteady Model

The Saint-Venant equations model fluid flow (usually water flow) in a channel. The Saint-Venant model assumes that the flow is strictly one-dimensional, although in practice it is used to model flows which are only approximately one-dimensional. Correction factors are often introduced into the model to correct for any deviation from one-dimensionality, but such factors are regarded as outside the scope of this thesis.

There are two common approaches to deriving the Saint-Venant equations. They may be derived by averaging the Reynolds equations over a cross-section of the channel which is normal to the direction of the flow. This is the approach found in [73]. Alternatively the equations may be derived by applying mass conservation, momentum conservation (in the direction of the flow) and Newton's second law to a suitable control volume of the channel. This second technique is considered the more useful since it yields the Saint-Venant equations in an integral form. This form of the equations continues to hold even when the differential form breaks down. The

control volume technique can be found, for example, in [9].

Before we introduce the Saint-Venant equations, we introduce the notation used to describe the channel geometry. We let x, y, z denote a Cartesian coordinate system with z pointing vertically upwards, and we consider a channel of length L along the x direction. For simplicity the channel is assumed symmetric about the plane $y = 0$, but this is not a restriction on the Saint-Venant model. Each cross-section (cross-section will always refer to a cross-section of constant x) of the channel is given by:

$$-\frac{\sigma(x,\eta)}{2} \leq y \leq \frac{\sigma(x,\eta)}{2}, \qquad \eta = z - z_b(x) \geq 0. \tag{2.1}$$

Here the *bed level* z_b is the height of the lowest point of the cross-section, η is a coordinate which measures height relative to this level and σ gives the width of the channel as a function of η. It is assumed that σ and z_b are continuously differentiable functions and that σ is positive for positive η. Figure 2.1 illustrates a typical cross-section.

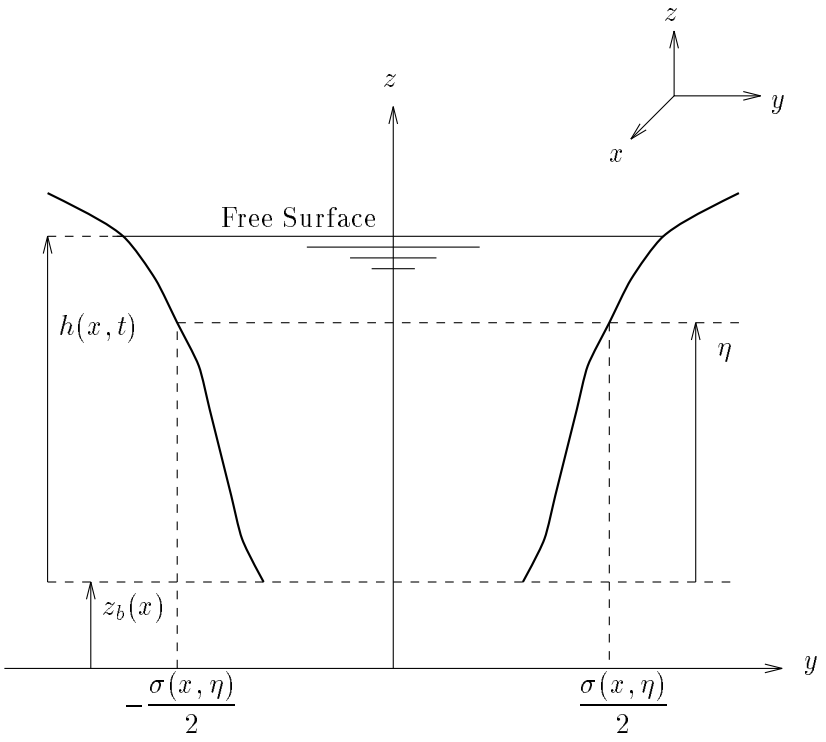


Figure 2.1: A typical channel cross-section

The set of assumptions under which the Saint-Venant equations are derived are

as follows.

(1) The fluid is incompressible, homogeneous and internal stresses are negligible.

(2) The flow is one-dimensional with the fluid velocity depending solely on x and time (t).

(3) At each cross-section the free surface is represented by a horizontal line.

(4) The streamline curvature is small and the vertical accelerations are negligible so the pressure can be taken as hydrostatic.

The *depth* $h(x,t)$ is the level of the free surface above the bed level and is illustrated in Figure 2.1. The *discharge* $Q(x,t)$ is defined to be the total volume flux through a given cross-section. If $u(x,t)$ is the x component of the fluid velocity then

$$Q = \int_0^h \int_{-\frac{\sigma}{2}}^{\frac{\sigma}{2}} u dy d\eta = Au, \tag{2.2}$$

where the *wetted area* $A(x,t)$ (the instantaneous area of the flow through any cross-section) is given by

$$A = \int_0^h \sigma d\eta. \tag{2.3}$$

Using the above assumptions the Saint-Venant equations can be derived by considering an arbitrary region of channel $x_1 \leq x \leq x_2$ over an arbitrary time interval $t_1 \leq t \leq t_2$. Applying conservation of mass yields the equation

$$\rho \int_{x_1}^{x_2} [A(x,t)]_{t_1}^{t_2} dx + \rho \int_{t_1}^{t_2} [Q(x,t)]_{x_1}^{x_2} dt = 0, \tag{2.4}$$

where ρ is the density,

$$[A(x,t)]_{t_1}^{t_2} = A(x,t_2) - A(x,t_1)$$

and

$$[Q(x,t)]_{x_1}^{x_2} = Q(x_2,t) - Q(x_1,t).$$

Equation (2.4) is referred to as the *integral mass equation*. Note we have assumed that there is no lateral inflow, i.e. that mass only enters the region through the cross-sections at x_1 and x_2.

Applying conservation of momentum (x component) to the same control region and same time interval yields the equation

$$\rho \int_{x_1}^{x_2} [Q(x,t)]_{t_1}^{t_2}\, dx + \rho \int_{t_1}^{t_2} [F(x,t)]_{x_1}^{x_2}\, dt = \rho \int_{t_1}^{t_2} \int_{x_1}^{x_2} D(x,t) dx dt. \tag{2.5}$$

Here $F(x,t)$ is given by

$$F = \frac{Q^2}{A} + gI_1, \tag{2.6}$$

where I_1 is given by

$$I_1 = \int_0^h (h-\eta)\sigma d\eta,$$

and g is the acceleration due to gravity. ρF represents the momentum flux through a cross-section and is composed of the advected momentum and a contribution from the hydrostatic pressure forces over the cross-section. $\rho D dx$ represents the instantaneous external forces acting on the fluid at a cross-section due to the channel boundary. It is composed of frictional forces and the reaction forces from hydrostatic pressure acting on the boundary. The function $D(x,t)$ is given by

$$D = gA(S_0 - S_f) + gI_2, \tag{2.7}$$

where I_2 is given by

$$I_2 = \int_0^h (h-\eta)\sigma_x d\eta,$$

S_0 is the *bed slope* given by

$$S_0 = -z_b', \tag{2.8}$$

and S_f is the *friction slope* which models frictional forces and is discussed later in this chapter. Equation (2.5) is referred to as the *integral momentum equation.*

Equations (2.4) and (2.5) constitute the Saint-Venant model. One possible choice of dependent variables is the depth h and discharge Q. Since the equations are derived for an arbitrary stretch of channel and an arbitrary time interval, a particular pair of functions h and Q are said to be a solution of these equations if the equations hold for all $0 \le x_1 \le x_2 \le L$ and $t_2 \ge t_1 \ge 0$.

Suppose that h and Q solve equations (2.4) and (2.5) and that, for some open region of the space-time domain, A, Q and F are continuously differentiable and D

is continuous, then it may be shown that the following differential equations hold on this region:

$$\frac{\partial A}{\partial t} + \frac{\partial Q}{\partial x} = 0, \tag{2.9}$$

$$\frac{\partial Q}{\partial t} + \frac{\partial F}{\partial x} = D. \tag{2.10}$$

These are the *differential form* of the Saint-Venant equations.

2.1.1 Discontinuous Solutions

Even though the differential form is not as general as the integral form, it is the form that is most often used in practice. The differential form breaks down with the formation of a shock (known in hydraulics as a *hydraulic bore* or simply *bore*). When this happens one must return to the integral form of the Saint-Venant equations to obtain conditions that describe the shock. These conditions are known as Rankine-Hugoniot conditions (see [5] [30] and [62]). Suppose the solution to the integral form contains a shock given by a simple jump discontinuity which moves smoothly through the space-time domain with path $x = x_S(t)$. Then at any point $(x,t) = (x_S(t),t)$ on the discontinuity, the following conditions must be satisfied:

$$Q_r - Q_l = s(A_r - A_l), \tag{2.11}$$

$$F_r - F_l = s(Q_r - Q_l), \tag{2.12}$$

where the l and r subscripts denote values of the quantities on the left and right of the discontinuity, respectively (e.g. $Q_l = Q(x-,t)$ and $Q_r = Q(x+,t)$) and $s = x_S'(t)$ is the *shock speed*. The conditions (2.11) and (2.12) were originally derived for the homogeneous system, but are unaffected by the inclusion of a source term.

It is a standard practice to refer to solutions of the integral form of a system conservation laws as *weak solutions* of the differential form even though they may or may not be classical solutions of the differential form. This extension of the concept of solution avoids having to explicitly refer to the integral form. The reference to the integral form of the underlying conservation law is, however, implicit in the term "weak solution".

If functions h and Q satisfy the differential form of the Saint-Venant equations except at discontinuities, where the Rankine-Hugoniot conditions (2.11) and (2.12) hold, then this is enough to ensure that they form a weak solution. However not all weak solutions are viable solutions to the physical problem. Some weak solutions may be so-called "entropy violating". This term arises from gas dynamics where certain solutions to the Euler equations result in a decrease in the entropy of the system, thus violating a fundamental law of thermodynamics. An extra condition on the shock which is called an *entropy condition* is required to prevent such solutions (see [5]). A similar situation occurs for the Saint-Venant equations where certain shocks create energy rather than dissipating energy. This is unreasonable because a hydraulic bore is clearly a dissipative phenomena with no mechanism to create energy. Following the argument in [64] (for the frictionless shallow water equations) the following equivalent condition can be obtained:

$$m\left(E_r - E_l - s(u_r - u_l)\right) \leq 0, \tag{2.13}$$

where $u = Q/A$ is the fluid velocity, E is given by

$$E = \frac{u^2}{2} + gh, \tag{2.14}$$

and

$$m = Q_r - sA_r = Q_l - sA_l.$$

Any physical hydraulic bore must satisfy the above condition.

2.1.2 Characteristic Speeds

The Saint-Venant equations are a hyperbolic system of partial differential equations. To see this we write the system in the vector form

$$\frac{\partial \boldsymbol{w}}{\partial t} + \frac{\partial \boldsymbol{F}}{\partial x} = \boldsymbol{D}, \tag{2.15}$$

where

$$\boldsymbol{w} = \begin{pmatrix} A \\ Q \end{pmatrix}, \quad \boldsymbol{F} = \begin{pmatrix} Q \\ F \end{pmatrix}, \quad \boldsymbol{D} = \begin{pmatrix} 0 \\ D \end{pmatrix}.$$

Now (2.15) can be written as

$$\frac{\partial \boldsymbol{w}}{\partial t} + J\frac{\partial \boldsymbol{w}}{\partial x} = \hat{\boldsymbol{D}}, \tag{2.16}$$

where J is the Jacobian given by

$$J = \frac{\partial \boldsymbol{F}}{\partial \boldsymbol{w}} = \begin{pmatrix} 0 & 1 \\ c^2 - u^2 & 2u \end{pmatrix},$$

c is the wave celerity given by

$$c = \sqrt{\frac{gA}{T}},$$

and $T = \sigma(x, h)$ is the free surface width. The modified source term is given by

$$\hat{\boldsymbol{D}} = \begin{pmatrix} 0 \\ gA(S_0 - S_f) + \frac{gA}{T}\int_0^h \sigma_x d\eta \end{pmatrix}.$$

The Jacobian J has real and distinct eigenvalues

$$\lambda_1 = u - c, \qquad \lambda_2 = u + c,$$

which give the characteristic speeds. The theory of characteristics can be found in [5] and [64]. The system of equations can be decomposed into two ordinary differential equations which hold along characteristic curves given by $dx/dt = \lambda_1$ and $dx/dt = \lambda_2$, respectively. Examples of this type of decomposition are given in [54] and [31]. It is important to know the directions of λ_1 and λ_2, since information is transmitted along these curves. The flow is classified according to the *Froude number*

$$F_r = \frac{|u|}{c} = \sqrt{\frac{Q^2 T}{gA^3}},$$

a dimensionless parameter which plays a role analogous to the Mach number in gas dynamics. For the case $F_r < 1$ which corresponds to $|u| < c$, one characteristic speed is negative and one is positive. Hence information is transmitted both upstream and downstream. This type of flow is known as *subcritical flow* and occurs when gravitational forces dominate over inertial forces

For the case $F_r > 1$ which corresponds to $|u| > c$, both characteristics are in the same direction as u. Hence information is only transmitted downstream. This type

of flow is known as *supercritical flow* and occurs when inertial forces dominate over gravitational forces.

The case $F_r = 1$ corresponds to $|u| = c$, and one characteristic is vertical and the other is in the same direction as u. This type of flow is known as *critical flow* and occurs when inertial forces and gravitational forces exactly balance.

The theory of characteristics also gives the data requirements for the boundaries of the space-time domain, in order for the problem to be well posed. *At any point on the boundary the number of independent flow variables specified must correspond exactly to the number of characteristic curves that enter the domain.* On the boundary $t = 0$, since both characteristics always enter the domain, two independent variables must always be specified. This information is called the initial data for the problem. Specifying the correct data on the boundaries $x = 0$ and $x = L$ is more complicated since the requirements depend on the actual solution at the particular point under consideration, and so in general cannot be determined in advance. A discussion of the boundary conditions for the Saint-Venant equations can be found in [9].

2.1.3 The Friction Slope

The friction slope S_f is intended to model effects due to boundary friction and turbulence. In this section explicit formulas are given for this term. These are empirical laws which were originally developed for use with steady state flow. The friction slope is usually written as

$$S_f = \frac{Q|Q|}{K^2},$$

where K is a quantity called the *conveyance*. In this thesis the following expression for the conveyance is used:

$$K = \frac{A^{k_1}}{n P^{k_2}}, \tag{2.17}$$

where $P(x,t)$ is the wetted perimeter (the instantaneous perimeter length of the flow in contact with channel at a given cross-section) given by

$$P = \sigma(x,0) + \int_0^h \sqrt{4 + \sigma_\eta^2}\, d\eta,$$

and n is a positive constant which determines the roughness of the channel. Equation (2.17) includes two of the most widely used forms of the conveyance, the Manning formula where $k_1 = 5/3$ and $k_2 = 2/3$ (for which the constant n is called the Manning friction coefficient) and the Chezy formula where $k_1 = 3/2$ and $k_2 = 1/2$ (for which the constant $C = 1/n$ is called the Chezy friction coefficient). More detailed information about these and other friction laws can be found in [4] and [9].

2.2 The Steady Problem

In this section we turn to the main topic of this thesis, the steady flow problem. Given an unsteady flow under steady boundary conditions, it is expected that the flow will eventually tend towards a steady state. Assuming that this happens, we ask the following question: Can the steady state be determined from the steady flow equations, under appropriate boundary conditions, without regard for the transient behaviour of the flow? The steady flow equations are obtained from the unsteady equations by assuming no time dependence. This leads to a system of two ordinary differential equations, for which the mass transport equation is trivial, effectively leaving only a single equation. The qualitative behaviour of solutions to the remaining equation are investigated for a special, but useful, class of channels. Because of shocks, the system of ordinary differential equations do not in general describe the entire steady flow. When they break down, the integral form must be used. We discuss the implications of the steady version of the Rankine-Hugoniot conditions.

The assumption that the flow will eventually reach a steady state is not always valid. In some circumstances the steady solutions of the Saint-Venant equations are not stable with respect to time, and thus steady state may never be attained. We discuss this topic further in section 2.2.5.

Suppose that $h = h(x)$ and $Q = Q(x)$ are a steady state solution of the Saint-Venant equations. Substituting these into the integral relationships (2.4) and (2.5) and using the fact that A, F and D are now independent of time gives that the equations

$$[Q]_{x_1}^{x_2} = 0, \tag{2.18}$$

$$[F]_{x_1}^{x_2} = \int_{x_1}^{x_2} D dx, \tag{2.19}$$

must be satisfied for all $0 \leq x_1 \leq x_2 \leq L$. These are the integral form of the steady flow equations. Equation (2.18) clearly implies that Q is constant throughout the entire reach. Without loss of generality the constant discharge Q is assumed positive, since if the discharge is negative then the x direction can be reversed to give a positive discharge. The case of zero discharge is a trivial case for which the solution is always given by a horizontal free surface.

At steady state the differential form of the Saint-Venant equations ((2.9) and (2.10)) reduce to

$$\frac{dQ}{dx} = 0, \tag{2.20}$$

$$\frac{dF}{dx} = D. \tag{2.21}$$

Equation (2.20) is clearly consistent with a constant discharge. At steady state any bore must be stationary, i.e. have zero velocity. A stationary bore is known as a *hydraulic jump*. Setting $s = 0$ in the Rankine-Hugoniot conditions (2.11) and (2.12) yields the jump conditions:

$$Q_r = Q_l, \tag{2.22}$$

$$F_r = F_l. \tag{2.23}$$

Again (2.22) is consistent with a constant discharge. Setting $s = 0$ in the "entropy" condition (2.13) yields the requirement that

$$E_r \leq E_l. \tag{2.24}$$

This inequality depends on the discharge being positive with the inequality being reversed for a negative discharge.

Expanding the derivative in equation (2.21) and using (2.20) gives the following differential equation for depth:

$$\frac{dh}{dx} = \frac{S_0 - \dfrac{Q^2}{K^2} + \dfrac{Q^2}{gA^3}\displaystyle\int_0^h \sigma_x d\eta}{1 - \dfrac{Q^2 T}{gA^3}}. \tag{2.25}$$

This is the form of the steady differential equation that is most amenable to both the analysis and numerical solution of smooth flows. However it breaks down when the

denominator of the right-hand side vanishes, i.e. when

$$F_r^2 = \frac{Q^2 T}{gA^3} = 1, \tag{2.26}$$

which corresponds to critical flow.

2.2.1 The Hydraulic Jump

In this section we consider the restrictions placed on a hydraulic jump by the conditions (2.23) and (2.24). In order to do this we require a relationship between the quantities F and E. We have from (2.6) and (2.14) that

$$\frac{\partial F}{\partial h} = gA\left(1 - \frac{TQ^2}{gA^3}\right) = gA(1 - F_r^2),$$

and

$$\frac{\partial E}{\partial h} = g\left(1 - \frac{TQ^2}{gA^3}\right) = g(1 - F_r^2) = \frac{1}{A}\frac{\partial F}{\partial h},$$

giving

$$\frac{\partial}{\partial h}\left(\frac{F - F_l}{A}\right) = -\frac{T}{A^2}(F - F_l) + \frac{\partial E}{\partial h}.$$

If we integrate this equation from h_l to h_r then we obtain the relationship

$$E_r - E_l = \int_{h_l}^{h_r} \frac{T}{A^2}(F - F_l)\, dh + \frac{F_r - F_l}{A_r}. \tag{2.27}$$

We can now consider the implications of (2.23) and (2.24) for a "well-behaved" channel cross-section. By "well-behaved" we mean that the width of the channel does not gradually approach zero as the depth becomes large and that there is a unique depth corresponding to critical flow. Mathematically these correspond to the following:

(1) $T \geq T_0 > 0$ as $h \to \infty$, for some constant T_0.

(2) There is only one depth h_c such that $h = h_c$ solves equation (2.26).

The depth h_c is called the *critical depth*. For depths below the critical depth the flow is supercritical and for depths above the critical depth the flow is subcritical.

Under the above assumptions the function F has the following properties:

(1) $F \to \infty$ as $h \downarrow 0$.

(2) $F \to \infty$ as $h \to \infty$.

(3) $\partial F/\partial h = 0$ at $h = h_c$.

(4) $\partial F/\partial h < 0$ for $h < h_c$.

(5) $\partial F/\partial h > 0$ for $h > h_c$.

To interpret the implications of (2.23) and (2.24) we ask what depths $h_r \neq h_l$ satisfy both of these conditions. There are three cases to consider

- If $h_l < h_c$, then there is exactly one $h_r \neq h_l$ satisfying (2.23). This depth $h_r > h_c$ is called the *sequent depth* of h_l and is denoted by h_l^*. Since (2.23) holds for this depth and we have $F - F_l < 0$ for $h_l < h < h_r$, then the relationship (2.27) clearly implies that (2.24) is satisfied. Hence there is exactly one allowable depth h_r.

- If $h_l > h_c$, then there is again exactly one $h_r \neq h_l$ satisfying (2.23). This depth $h_r < h_c$ is again called the sequent depth of h_l and is denoted by h_l^*. However since $F - F_l < 0$ for $h_r < h < h_l$, the relationship (2.27) shows that (2.24) is violated. Hence in this case there are no allowable values for h_r.

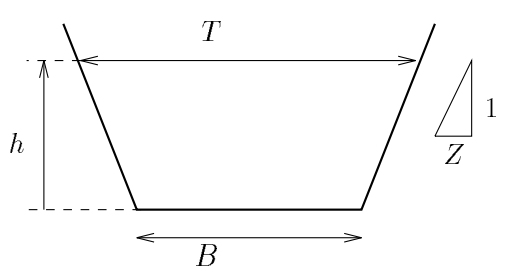
- If $h_l = h_c$, then (2.23) is satisfied if and only if $h_r = h_c$, so again there are no allowable values for $h_r \neq h_l$. For consistency we define $h_c^* = h_c$.

We conclude that a hydraulic jump satisfies both (2.23) and (2.24) if and only if

$$h_l < h_c < h_r = h_l^*. \tag{2.28}$$

2.2.2 Surface Profiles for a Prismatic Channel

In this section the solutions of (2.25) are examined for a prismatic channel. The elementary analysis follows that of many standard textbooks such as [4], [16] and [24]. A channel is *prismatic* if its cross-section is unchanging throughout its length, i.e. if the function σ is independent of x. For a prismatic channel, the quantities A,



$$T = B + 2hZ$$

$$A = h(B + hZ)$$

$$P = B + 2h\sqrt{1 + Z^2}$$

Figure 2.2: Trapezoidal channel cross-section

T, K, F and E are solely functions of h and hence equation (2.25) reduces to

$$\frac{dh}{dx} = \frac{S_0 - \dfrac{Q^2}{K^2}}{1 - \dfrac{Q^2 T}{gA^3}}. \tag{2.29}$$

We assume that the cross-section satisfies the following conditions:

$$\left.\begin{array}{ll} (1) & \dfrac{T}{A^3} \text{ is strictly decreasing in } h \text{ for } h > 0. \\ (2) & T \geq T_0 > 0 \text{ as } h \to \infty, \text{ for some constant } T_0. \end{array}\right\} \tag{2.30}$$

These conditions ensure that the shape of channel is "well-behaved" in the sense of section 2.2.1.

Conditions are also placed on the conveyance K. These are as follows:

$$\left.\begin{array}{ll} (1) & K > 0 \text{ is strictly increasing in } h \text{ for } h > 0. \\ (2) & K \to 0 \text{ as } h \downarrow 0. \\ (3) & K \to \infty \text{ as } h \to \infty. \end{array}\right\} \tag{2.31}$$

To show that the above conditions hold for practical problems, consider the channel given by

$$T = B + 2Zh. \tag{2.32}$$

This formula covers three different shapes of cross-section. These are *rectangular* ($B > 0, Z = 0$), *triangular* ($B = 0, Z > 0$) and *trapezoidal* ($B > 0, Z > 0$, see Figure 2.2). For these shapes

$$\frac{T}{A^3} = \frac{B + 2hZ}{h^3(B + hZ)^3},$$

so that

$$\frac{d}{dh}\left(\frac{T}{A^3}\right) = \frac{-3B^2 - 10BZh - 10Z^2h^2}{h^4(B+hZ)^4} < 0,$$

hence satisfying (2.30(1)). Condition (2.30(2)) is also clearly satisfied.

Suppose now that equation (2.17) is used for the conveyance with $k_1 > 0$ and $k_1 > k_2$, which includes both the widely used Manning and Chezy formulae. The conveyance is now given by

$$K = \frac{(Bh + h^2Z)^{k_1}}{n(B + 2h\sqrt{1+Z^2})^{k_2}},$$

giving

$$\frac{dK}{dh} = \frac{K}{AP}\left(Bk_1(B+2hZ) + 2h\sqrt{1+Z^2}\,(B(k_1-k_2) + hZ(2k_1-k_2))\right) > 0, \tag{2.33}$$

satisfying (2.31(1)). For the case of a rectangular channel ($Z = 0$, $B > 0$), we have

$$K = \frac{(Bh)^{k_1}}{n(B+2h)^{k_2}} \to 0 \text{ as } h \downarrow 0,$$

and

$$K = \frac{B^{k_1}}{n(B/h+2)^{k_2}} h^{k1-k2} \to \infty \text{ as } h \to \infty.$$

For the case of a triangular channel ($B = 0$, $Z > 0$) we have

$$K = \frac{Z^{k_1}}{n(2\sqrt{1+Z^2})^{k_2}} h^{2k_1-k_2},$$

which tends to zero as $h \downarrow 0$ and tends to infinity as h tends to infinity. Finally for the trapezoidal case ($B > 0$, $Z > 0$), K clearly tends to zero as $h \downarrow 0$ and

$$K = \frac{(B/h+Z)^{k_1}}{n(B/h+2\sqrt{1+Z^2})^{k_2}} h^{2k_1-k_2} \to \infty \text{ as } h \to \infty.$$

This demonstrates all the properties (2.30) and (2.31) for this family of channel cross-sections with the given friction formulae. We can now use these properties to investigate the solutions of the differential equation (2.29).

The Normal Depth

If at a given cross-section $S_0(x) > 0$, the *normal depth* $h_n(x)$ is defined to be the unique depth satisfying

$$S_0(x) = \frac{Q^2}{K^2}. \tag{2.34}$$

The set of conditions (2.31) ensure that such a depth exists. For the case $S_0(x) \leq 0$ equation (2.34) has no solution, but for convenience we define $h_n(x)$ to have a value of infinity.

Conditions (2.30) and (2.31) mean that given any position x and any depth h, the corresponding sign of dh/dx can be determined solely from the position of h relative to h_c and $h_n(x)$. The depth range is divided into three zones and we summarise the information in Table 2.1.

Zone	**Depth range**	$\boldsymbol{dh/dx}$
1†	$h > \max\{h_c, h_n(x)\}$	Positive
2	$\min\{h_c, h_n(x)\} < h < \max\{h_c, h_n(x)\}$	Negative
3	$0 < h < \min\{h_c, h_n(x)\}$	Positive

Table 2.1: Sign of depth gradient as a function of depth

To determine the relative positions of the h_c and $h_n(x)$, we define the *critical bed slope* S_{0c} by

$$S_{0c} = \left.\frac{Q^2}{K^2}\right|_{h=h_c}. \tag{2.35}$$

At any particular cross-section the bed slope of the channel is classified in the following manner. If $S_0(x) < 0$ then the slope is called *adverse*. If $S_0(x) = 0$ then the slope is horizontal. If $0 < S_0(x) < S_{0c}$ then the slope is said to be *mild* and $h_n(x) > h_c$. If $S_0(x) = S_{0c}$ then the slope is said to be *critical* and $h_n(x) = h_c$. If $S_0(x) > S_{0c}$ then the slope is said to be *steep* and $h_n(x) < h_c$.

Constant Bed Slope

The situation is most straightforward when the bed slope is constant throughout the length of the reach. In this case the normal depth is also constant throughout the channel length and we can use the properties (2.30) and (2.31) to obtain the following facts: If $S_0 \neq S_{0c}$ then

$$\left|\frac{dh}{dx}\right| \longrightarrow \infty \text{ as } h \longrightarrow h_c. \tag{2.36}$$

†Zone 1 does not exist for $S_0(x) \leq 0$.

If $S_0 > 0$ and $S_0 \neq S_{0c}$ then

$$\frac{dh}{dx} \longrightarrow 0 \text{ as } h \longrightarrow h_n. \tag{2.37}$$

Also we have

$$\frac{dh}{dx} \longrightarrow S_0 \text{ as } h \longrightarrow \infty, \tag{2.38}$$

so the free surface tends to the horizontal at large depths. The limit of dh/dx as $h \downarrow 0$ cannot be determined without more specific information about the friction law and the channel shape. Likewise for the limit as $h \to h_c = h_n$ for a critical bed slope. Putting all the above information together allows us to determine the behaviour of the free surface at any particular depth for any type of bed slope. For example consider a mild bed slope ($0 < S_0 < S_{0c}$, $h_n > h_c$).

In zone 1 (see Table 2.1) the depth increases with x, the free surface tending to the horizontal downstream as depth becomes large. The depth approaches the normal depth in the upstream direction.

In zone 2 the depth decreases with x. Downstream the free surface becomes vertical as the depth tends to the critical depth, at which point the differential equation breaks down. The depth tends towards the normal depth in the upstream direction.

In zone 3 the depth increases with x. Downstream the free surface becomes vertical as the depth tends to the critical depth and the differential equation breaks down. Upstream the differential equation again breaks down as the depth approaches zero.

The different surface profiles are illustrated in Figure 2.3 and are labeled M1, M2 and M3. ("M" denotes mild bed slope. The same convention will be used for the remaining types of bed slope, with "S" denoting steep bed slope, "C" denoting critical bed slope, "H" denoting horizontal bed slope and "A" denoting adverse bed slope.) The behaviour of the surface profiles for the remaining types of bed slope are illustrated in Figures 2.4, 2.5, 2.6 and 2.7.

We now give an example to illustrate how the previous material can be used to describe the behaviour of a simple problem with a hydraulic jump. Similar examples can be found in [4]. Consider a channel with constant mild bed slope and assume that at the inflow boundary ($x = 0$) the depth is $h_0 < h_c$ and that at the outflow

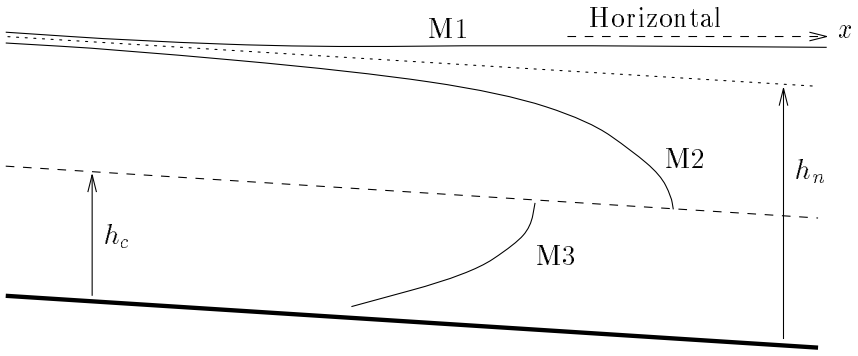


Figure 2.3: Behaviour of free surface for a channel with constant mild bed slope

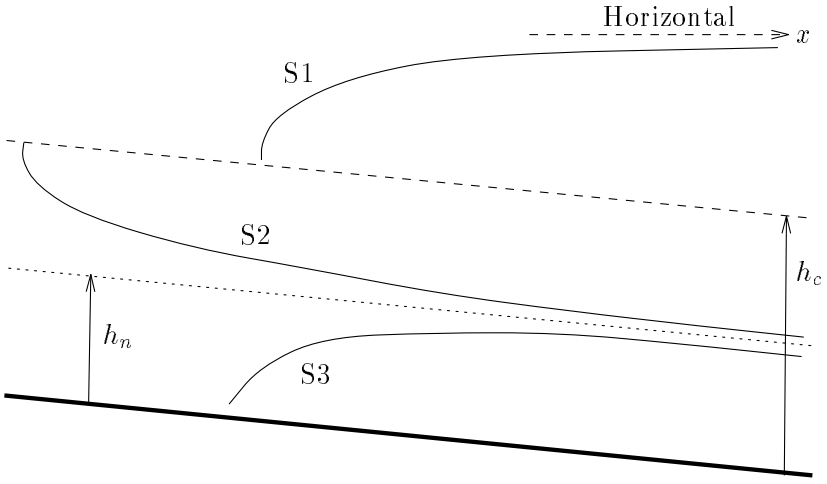


Figure 2.4: Behaviour of free surface for a channel with constant steep bed slope

boundary ($x = L$) the depth is $h_L > h_n$. Proceeding downstream from the inflow we have an M3 profile, and it is assumed that the channel is long enough so that this profile terminates before it reaches the outflow. Moving upstream from the outflow there is an M1 profile. This profile will eventually approach the normal depth. In [59] it is shown that close to the normal depth, the depth behaves as

$$h = h_n + C_0 \exp(\lambda x), \tag{2.39}$$

where C_0 and $\lambda > 0$ are constants. λ can be calculated and gives a measure of how fast the depth tends to the normal depth. Figure 2.8 shows the M1 and M3 profiles. The M3* profile shows the allowable jumps from the M3 profile, i.e. any jump from the M3 to the M3* curve satisfies both (2.23) and (2.24). A hydraulic jump occurs

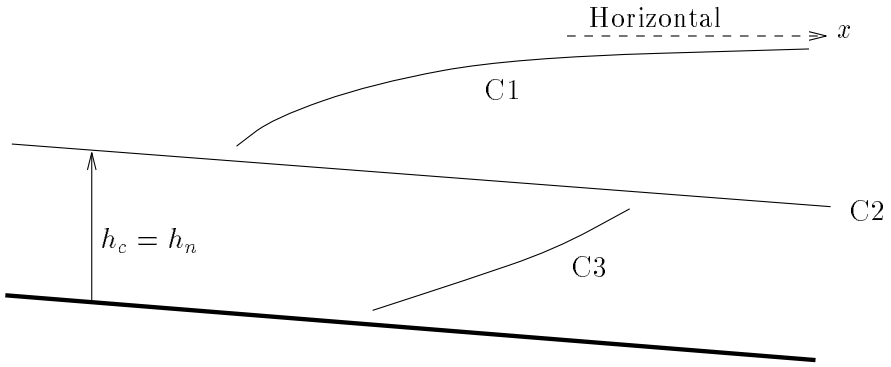


Figure 2.5: Behaviour of free surface for a channel with critical bed slope

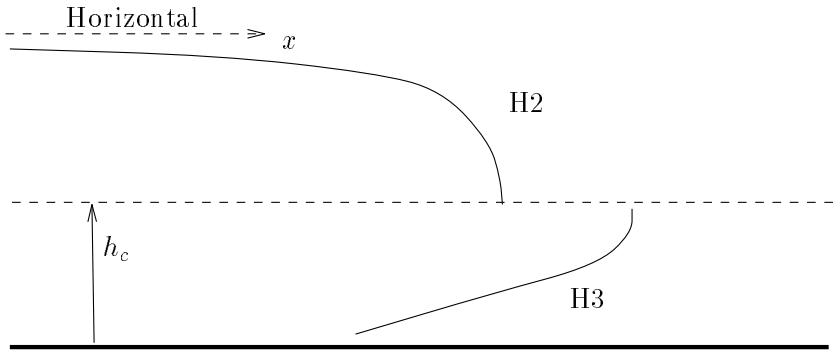


Figure 2.6: Behaviour of free surface for a horizontal channel

at a point where the M3* curve intersects the M1 curve. Since the downstream side of any jump is likely to be close to the normal depth, it is likely that a good estimate of the height of the jump can be obtained (since $h(x+) \approx h_n$ so that $h(x-) \approx h_n^*$) and the position of the jump will depend almost entirely on the M3 profile. If F_0 and F_n denote the values of the quantity F for depths h_0 and h_n, respectively, then on the M1 curve $F > F_n$ and on the M3 curve $F \leq F_0$. If $F_0 \leq F_n$ then there can be no jump so that there is no solution satisfying both boundary conditions. In this case the M3 flow is completely drowned out and the flow will be the subcritical M1 flow for the entire reach.

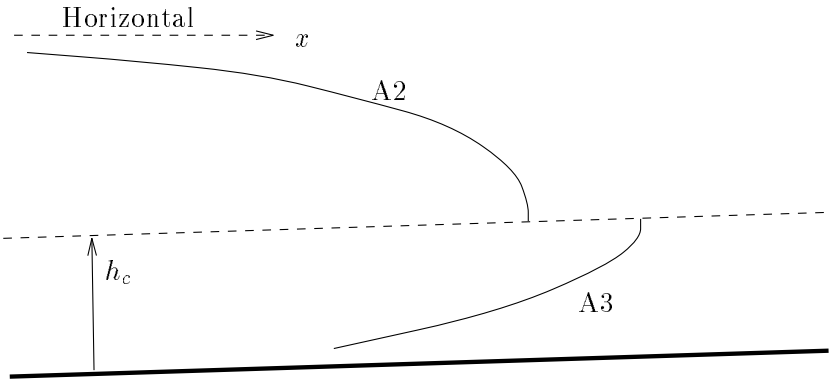


Figure 2.7: Behaviour of free surface for a channel with constant adverse bed slope

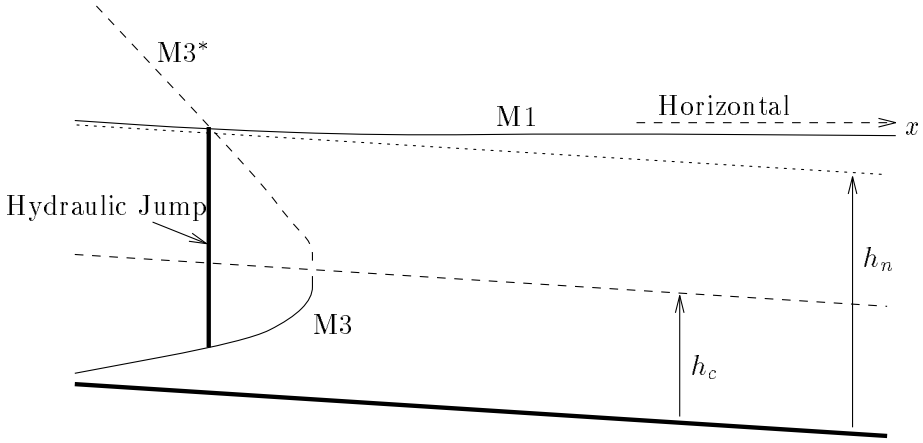


Figure 2.8: Example of problem with hydraulic jump

Varying Bed Slope

We now consider the situation where the bed slope is allowed to vary along the reach. The normal depth then varies with x and solutions will not in general asymptote to the normal depth curve. However, the sign of dh/dx can still be predicted from Table (2.1) and $|dh/dx|$ still becomes infinite as the critical depth is approached.

There is now the possibility of *singular points*. A singular point is where the numerator and the denominator of equation (2.29) both simultaneously vanish. This happens at a point x_c when the bed slope passes through its critical value, i.e.

$$S_0(x_c) = S_{0c}. \tag{2.40}$$

At such a point it is possible for equation (2.29) to have a solution which passes smoothly through the critical depth $h_c = h_n(x_c)$. Equation (2.29) cannot immediately be used to calculate the gradient when passing through the critical depth, but we can apply L'Hôpitals rule [63], [65] to the right hand side of (2.29) to obtain

$$\left.\frac{dh}{dx}\right|_{x_c} = \lim_{x\to x_c} \frac{S_0 - \dfrac{Q^2}{K^2}}{1 - \dfrac{Q^2 T}{gA^3}} = \lim_{x\to x_c} \frac{\dfrac{dS_0}{dx} + 2\dfrac{Q^2}{K^3}\dfrac{dK}{dh}\dfrac{dh}{dx}}{-\dfrac{Q^2}{g}\dfrac{d}{dh}\left(\dfrac{T}{A^3}\right)\dfrac{dh}{dx}} = \frac{\left.\dfrac{dS_0}{dx}\right|_{x_c} + a_1 \left.\dfrac{dh}{dx}\right|_{x_c}}{a_2 \left.\dfrac{dh}{dx}\right|_{x_c}}, \tag{2.41}$$

where

$$a_1 = 2\frac{Q^2}{K^3} \left.\frac{dK}{dh}\right|_{h_c} > 0, \tag{2.42}$$

and

$$a_2 = -\frac{Q^2}{g}\frac{d}{dh}\left.\left(\frac{T}{A^3}\right)\right|_{h_c} > 0. \tag{2.43}$$

This relationship can be solved for the depth gradient at the critical point, yielding

$$\left.\frac{dh}{dx}\right|_{x_c} = \frac{a_1}{2a_2}\left(1 \pm \sqrt{1 + \frac{4a_2}{a_1^2}\left.\frac{dS_0}{dx}\right|_{x_c}}\right). \tag{2.44}$$

In the case $dS_0/dx > 0$ at $x = x_c$, i.e. the slope changes from mild to steep, then there are always two possible values for dh/dx at the critical point. One value is negative and one value is positive. Figure 2.9 illustrates the two solutions AOB and A*OB* passing through the singular point O. The solution AOB is extremely important in the theory of steady flow since it is the only mechanism for which the flow can accelerate from subcritical to supercritical flow. This is because hydraulic jumps from subcritical to supercritical flow are prohibited by the requirement (2.28). For a given channel, discharge and friction law, the possible locations of a transition from subcritical to supercritical can be found in advance from (2.40). These are referred to as *critical sections*. Under the correct conditions the solution A*OB* can also occur, but this is only one of the many ways that the flow may change from supercritical to subcritical. It is also possible that solutions such as A*OB and AOB* can happen.

Next consider the case where $dS_0/dx < 0$ at $x = x_c$, i.e. the slope changes from mild to steep. If the value under the square root in (2.44) is positive, then there are again two possible values for dh/dx at the critical point, and these are both positive,

so in this case the flow cannot change from subcritical to supercritical at such a point. If the value under the square root in (2.44) is zero there is only one value, and this is again positive. Finally, if the value under the square root is negative there are no allowable values for dh/dx at the critical point.

We conclude that the flow may only change from subcritical to supercritical at a point where the bed slope changes from mild to steep. The change in the flow type is in the form of a smooth transition through the critical depth.

Using the theory of nonlinear systems of ordinary equations (see [13], [28]) more advanced analysis of the flow near singular points may be carried out. This is discussed in [4]. The singular points can be classified as either saddle, nodal, spiral or vortex types.

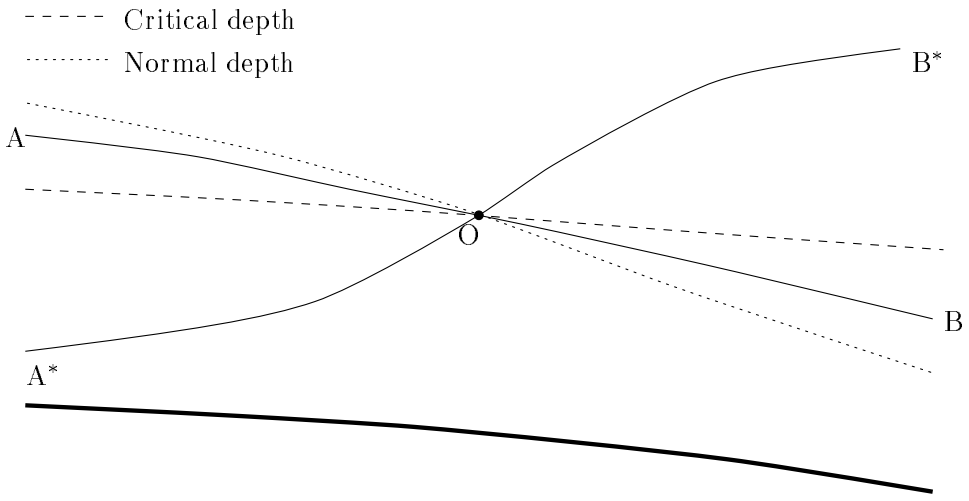


Figure 2.9: Flow profiles through a singular point for an increasing bed slope

2.2.3 Non-Prismatic Channels

We now discuss very briefly the theory which carries over from the prismatic case to the non-prismatic case where the function σ now depends on x. The current definition of normal depth is no longer useful. This is because there are now three terms on the numerator of the right-hand side of (2.25). There is no longer a straightforward competition between gravity and friction. Terms due to gravity, friction and forces

due to contraction or expansion of the channel cross-section must compete with each other. A new more complex definition of normal depth requires taking into account the variation of the channel cross-section with x. The theory of singular points also becomes more complicated. The theory for non-prismatic channels is beyond the scope of this thesis, but it is discussed in for example [4].

2.2.4 Steady Boundary Conditions

A steady solution of the Saint-Venant equations is only a special example of an unsteady solution, so it must satisfy the same boundary data requirements as any unsteady solution, but clearly any boundary data must be constant in time. The number of independent variables specified must be exactly equal to the number of characteristics entering the domain. Consider the characteristics speeds given by $\lambda_1 = u - c$ and $\lambda_2 = u + c$. Since Q and hence u is taken as positive, λ_2 is always positive.

Since $\lambda_2 > 0$ at the inflow boundary then at least one variable must always be specified there. This will be taken to be the value of the constant discharge. If a value of the depth is to be specified then it must be such that $\lambda_1 > 0$, i.e. a depth corresponding to supercritical flow.

At the outflow boundary one variable is the maximum that may be specified. This will always be taken to be the depth, and should be such that $\lambda_1 < 0$, i.e. a depth corresponding to subcritical flow.

Even if boundary conditions are supplied which obey the above requirements, this does not guarantee that there exists a steady solution satisfying these boundary conditions, and this will be demonstrated in section 4.9.

2.2.5 Stability of the Steady State

In this chapter steady solutions of the steady Saint-Venant equations are investigated. For such a steady solution to be of any practical use, the solution must be stable in time. This is very often overlooked in practice. Suppose that at some time t the flow is at a steady state and that it is perturbed slightly away from the steady state. The

steady state is said to be *stable* if for any such small perturbation, the flow eventually tends back to the original steady state. If for some small perturbation the flow does not tend back to the original state, then it is said to be *unstable*. In any practical situation small perturbations are always present, so any unstable steady state will not be maintainable indefinitely. Such steady states often degenerate into what are called *roll waves* (for example see [4], [64], [31]); the perturbations in the flow grow until an otherwise smooth flow breaks up into surges, entirely changing the character of the flow. This is an important subject in the theory of channel design since a channel that is designed to carry an expected steady flow may be overtopped by the formation of roll waves.

There is a shortage of analysis of the stability of steady state solutions to the Saint-Venant equations. In [31] there is analysis of the stability of normal flow for general shape prismatic channels. Perturbations are taken in the flow variables about normal flow. The characteristic form of the Saint-Venant equations are used to decide whether the perturbations grow or decay along each characteristic. An important quantity is the *Verdernikov* number given by

$$V_e = \frac{Q}{K}\frac{d}{dh}\left(\frac{K}{A}\right)\sqrt{\frac{A}{gT}}. \tag{2.45}$$

It is shown that the perturbations will grow unless

$$|V_e| \leq 1. \tag{2.46}$$

This condition gives information about the stability of the mathematical model rather than the actual physical flow, although such instabilities are observed in reality. The absolute value of the Verdernikov number can be written as

$$|V_e| = \frac{A^2}{K}\left|\frac{d}{dh}\left(\frac{K}{A}\right)\right| F_r. \tag{2.47}$$

For many channels (for example rectangular channels) condition (2.46) becomes violated as the Froude number becomes large when using the Manning or Chezy friction laws. More details can be found in [31].

The above theory simply uses linear stability analysis. More advanced techniques could be used to analyse stability and be applied for more cases.

Chapter 3

Shock Capturing Methods

The Saint-Venant equations form a system of conservation laws of hyperbolic type. Such systems of equations occur frequently in applied mathematics and so much effort has been put into developing numerical methods for their solution. This thesis is primarily interested in computing steady solutions for which a conventional approach is to numerically model the transient behaviour of the system until steady state is attained. We discuss the subject of time dependent schemes here in order to allow comparison with "more direct" approaches. However a more important reason for discussing these schemes is that when applied to a particular scalar partial differential equation, they can themselves lead to more direct methods.

For a numerical scheme to be of practical use it must be possible to write it in what is known as conservation form and such schemes are known as conservative. We start this chapter by defining this property which leads on to the approach of Godunov. We then discuss conservative methods for the scalar problem and then show how these can be generalised to systems of equations via the use of approximate Riemann solvers. In addition to requiring that a scheme be conservative, we require that a scheme be stable in the presence of discontinuities, and such schemes are known as *shock capturing* schemes. We discuss the subject of stability and the extension of shock capturing schemes to higher order accuracy by the use of limiter functions.

In this chapter we also discuss more efficient computation of steady states using implicit schemes and methods of discretisation of source terms. The most commonly used approximate Riemann solver is that due to Roe[55]. We end the chapter by

giving details of how this method can be implemented for the Saint-Venant equations.

Consider the system of equations

$$\frac{\partial \boldsymbol{u}}{\partial t} + \frac{\partial}{\partial x}\boldsymbol{f}(\boldsymbol{u}) = 0. \tag{3.1}$$

We shall consider numerical approximations on a uniform grid in x-t space, with Δx and Δt denoting the grid spacing in space and time respectively. The nodes are labelled by the indices j and n with positions given by $(x_j, t_n) = (j\Delta x, n\Delta t)$. It is normal in the theory of numerical methods to let $\boldsymbol{u}_j^n$ denote the value of the approximation to the exact solution at the particular grid point, i.e. $\boldsymbol{u}_j^n \approx \boldsymbol{u}(x_j, t_n)$. However for modelling systems of conservation laws it is more appropriate for $\boldsymbol{u}_j^n$ to denote the value of the approximation to the cell average of the exact solution, where a cell is given by the spatial interval $[x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}]$ with $x_{j+\frac{1}{2}} = (j + \frac{1}{2})\Delta x$. We therefore have that

$$\boldsymbol{u}_j^n \approx \frac{1}{\Delta x}\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} \boldsymbol{u}(x, t_n)dx. \tag{3.2}$$

3.1 The Conservation Form

The system (3.1) is of *hyperbolic* type if the Jacobian of the function $\boldsymbol{f}$ has all real eigenvalues and has a full set of linearly independent eigenvectors. The early attempts at numerically modelling this type of equation were for the linear case and simply involved replacing the derivatives by finite difference formulae. This led to many classical schemes such as Lax-Wendroff[29]. Such schemes were successful for solving problems with smooth solutions but failed miserably for discontinuous solutions. Certain schemes such as the one-sided second order scheme could compute discontinuous solutions successfully, but new difficulties were encountered when generalising these to the nonlinear problem. Although convincing discontinuous solutions were obtained, more often than not at further inspection the discontinuities were found to move at the wrong speed. The fact that a numerical solution does not satisfy the appropriate Rankine-Hugoniot conditions indicates that the scheme does not approximate consistently the underlying conservations law. Examples of this behaviour can be found in [30] Chapter 11.

Fortunately a simple requirement exists that if satisfied ensures that a scheme does approximate the correct conservation law. This requirement is that the scheme can be written in conservation form. A scheme is in *conservation form* if it is written in the form

$$\frac{\boldsymbol{u}_j^{n+1} - \boldsymbol{u}_j^n}{\Delta t} + \frac{\boldsymbol{g}_{j+\frac{1}{2}}^n - \boldsymbol{g}_{j-\frac{1}{2}}^n}{\Delta t} = 0, \tag{3.3}$$

where

$$\boldsymbol{g}_{j+\frac{1}{2}}^n = \boldsymbol{g}(\boldsymbol{u}_{j+k}^n, \ldots, \boldsymbol{u}_{j+1}^n, \boldsymbol{u}_j^n, \ldots . \boldsymbol{u}_{j-k+1}^n)$$

for some k. For this scheme to be consistent with the differential equation (3.1) the function $\boldsymbol{g}$ is required to satisfy the consistency condition

$$\boldsymbol{g}(\boldsymbol{u}, \ldots, \boldsymbol{u}, \boldsymbol{u}, \ldots, \boldsymbol{u}) = \boldsymbol{f}(\boldsymbol{u}), \tag{3.4}$$

for all $\boldsymbol{u}$. We also require $\boldsymbol{g}$ to be Lipschitz continuous. Under these conditions Lax and Wendroff [29] demonstrated that any convergent sequence of solutions (as $\Delta x \downarrow 0$, $\Delta t \downarrow 0$ with $\Delta t/\Delta x$ fixed) must converge to a weak solution of the conservation law.

The function $\boldsymbol{g}$ is called the numerical flux function since it approximates the time average flux across the cell interfaces, i.e.

$$\boldsymbol{g}_{j+\frac{1}{2}}^n \approx \frac{1}{\Delta t} \int_{t_n}^{t_{n+1}} \boldsymbol{f}(\boldsymbol{u}(x_{j+\frac{1}{2}}, t))dt. \tag{3.5}$$

To see why this is true substitute

$$\boldsymbol{u}_j^n = \frac{1}{\Delta x} \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} \boldsymbol{u}(x, t_n)dx, \tag{3.6}$$

$$\boldsymbol{g}_{j+\frac{1}{2}}^n = \frac{1}{\Delta t} \int_{t_n}^{t_{n+1}} \boldsymbol{f}(\boldsymbol{u}(x_{j+\frac{1}{2}}, t))dt, \tag{3.7}$$

into the scheme (3.3) and multiply by a factor $\Delta x \Delta t$ to obtain:

$$\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} [\boldsymbol{u}]_{t_n}^{t_{n+1}} \, dx + \int_{t_n}^{t_{n+1}} [\boldsymbol{f}(\boldsymbol{u})]_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} \, dt = 0.$$

This is the integral form of the conservation law over the rectangle $[x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}] \times [t_n, t_{n+1}]$.

3.2 The Godunov Method

The method of Godunov[19] comes from the observation that the numerical solution satisfies the integral form of the conservation law exactly if (3.6) and (3.7) hold. The method proceeds in the following manner:

(1) Replace the initial data $\boldsymbol{u}_0$ by a piecewise constant function where the constant value in each cell is given by the cell average.

(2) Use the piecewise constant function and the formula (3.7) to compute the numerical fluxes across each cell interface.

(3) Use the numerical scheme to compute the cell averages at the next time level, hence defining a new piecewise constant function.

(4) Repeat from step 2.

This appears to be a relatively straightforward strategy for stepping the numerical solution forward in time. The crucial step is the third step which involves computing the time average flux at each cell interface. At each cell interface we have a Riemann problem. A *Riemann problem* consists of the the system (3.1) with piecewise constant initial data of the form

$$\boldsymbol{u}_0(x) = \begin{cases} \boldsymbol{u}_l & x < 0 \\ \boldsymbol{u}_r & x > 0. \end{cases}$$

For many systems of the conservation laws the exact solution of the Riemann problem can be found. It can be shown that all solutions are similarity solutions in x/t, i.e. are of the form $\boldsymbol{u}(x,t) = \boldsymbol{w}(x/t)$ (see [30]). For this reason the solution to the Riemann problem has the constant value $\boldsymbol{w}(0)$ on the line $x = 0$ for all positive t. Thus if $\boldsymbol{w}(0)$ is known then it is a trivial matter to integrate the flux along this line in time. The simplest case to solve for is the case of the scalar problem and in particular when the scalar flux function f is convex. In this case one solution of the Riemann problem is of the form

$$u(x,t) = \begin{cases} u_l & x < st \\ u_r & x > st, \end{cases} \tag{3.8}$$

where s is the shock speed given by

$$s = \frac{f(u_r) - f(u_l)}{u_r - u_l}.$$

This last equation is simply the Rankine-Hugoniot condition for the conservation law. For this solution we have

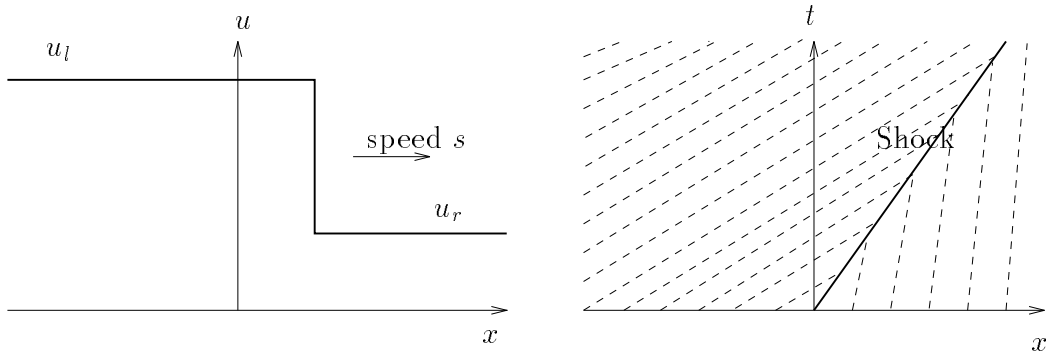


Figure 3.1: Illustration of solution and characteristic diagram for a shock

$$w(0) = \begin{cases} u_l & s \geq 0 \\ u_r & s < 0. \end{cases} \tag{3.9}$$

For given initial data a conservation law may have many different weak solutions, but has only one physical solution. This unique physical solution is referred to as the *entropy satisfying solution* or just as the entropy solution, since it is identified from all the possible weak solutions by the fact that it satisfies conditions known as entropy conditions. The subject of entropy conditions is discussed in the next chapter. For a convex f the entropy condition requires that the characteristics lines must go into a shock (see [30], [62]). This requires that

$$f'(u_r) > s > f'(u_l).$$

From the convexity of f this only holds in the case $u_l > u_r$. In the case $u_r > u_l$ the entropy satisfying solution to the Riemann problem is an *expansion shock* (also known as an *rarefaction wave*). This smooth solution is given by

$$u(x,t) = \begin{cases} u_l & x < f'(u_l)t \\ \hat{w}(x/t) & f'(u_l)t < x < f'(u_r)t \\ u_r & x > f'(u_r)t, \end{cases}$$

where $\hat{w}$ satisfies

$$f'(\hat{w}(\xi)) = \xi.$$

For this solution

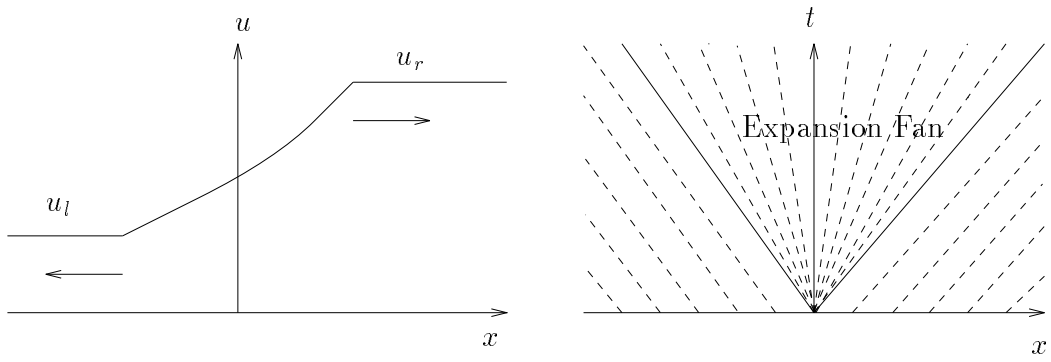


Figure 3.2: Illustration of solution and characteristic diagram for an expansion wave

$$w(0) = \begin{cases} u_l & f'(u_l) \geq 0 \\ u_c & f'(u_l) < 0 < f'(u_r) \\ u_r & f'(u_r) \leq 0, \end{cases} \tag{3.10}$$

where u_c is unique sonic point satisfying $f'(u_c) = 0$. Combining (3.9) and (3.10) we obtain the entropy satisfying value of $w(0)$:

$$w(0) = \begin{cases} u_l & f'(u_l), f'(u_r) \geq 0 \\ u_r & f'(u_l), f'(u_r) \leq 0 \\ u_l & f'(u_l) \geq 0 \geq f'(u_r),\ s \geq 0 \\ u_r & f'(u_l) \geq 0 \geq f'(u_r),\ s < 0 \\ u_c & f'(u_l) < 0 < f'(u_r). \end{cases} \tag{3.11}$$

We can now write the time average flux across each cell interface as

$$g^n_{j+\frac{1}{2}} = \frac{1}{\Delta t}\int_{t_n}^{t_{n+1}} f(u(x_{j+\frac{1}{2}}, t))dt = \frac{1}{\Delta t}\int_{t_n}^{t_{n+1}} f(w(0))dt = f(w(0)),$$

where here $w(0)$ denotes (3.11) with $u_l = u^n_j$ and $u_r = u^n_{j+1}$. The same arguments allow the calculation of $w(0)$ for a concave f. If f is neither concave or convex then the situation is considerably more complex. The solution of the Riemann problem can now also be a combination of a shock and an expansion wave. A general form

for $w(0)$ can still be found (see [48]), giving the general form for the numerical flux

$$g^{\text{God}}_{j+\frac{1}{2}} = \begin{cases} \max\{f(s) : u_{j+1} \le s \le u_j\} & \text{for} \quad u_{j+1} \le u_j \\ \min\{f(s) : u_j \le s \le u_{j+1}\} & \text{for} \quad u_{j+1} > u_j. \end{cases} \tag{3.12}$$

The notation we use is to omit the n superscripts in the definition of quantities when it is clear that the definition at a particular time level n is simply obtained by introducing n superscripts to all the time dependent variables. The solution of a single Riemann problem is constant at the position of the initial discontinuity for all time. In the case of the numerical scheme where in effect we solve a sequence of neighbouring Riemann problems, this will not be the case for large times since waves will arrive from neighbouring Riemann problems. To prevent this occurring one must limit the size of the time step Δt to prevent neighbouring Riemann problems from interacting. The wave speeds are bounded by the eigenvalues of the Jacobian of the system and the neighbouring Riemann problems are distance Δx away. For the scalar problem it is therefore sufficient to require that

$$\left| f'(u_j) \frac{\Delta t}{\Delta x} \right| \le 1, \tag{3.13}$$

for all j at each time level. For a system of conservation laws, f' must be replaced by the eigenvalue of largest magnitude of the system. This condition can also be derived from a domain of dependence argument and is a fundamental requirement. The need for such a condition was first recognized by Courant, Friedrichs and Lewy in their famous paper of 1928[6] (translation in [7]) and is called a CFL condition after these authors. In order to satisfy the CFL condition in practice one would allow a variable time step and choose the time step at each time level so as to satisfy the CFL condition at that current time level.

For a system of conservation laws the task of computing a solution to the Riemann problem is considerably more taxing. Typically the solution is composed of m waves, where m is the size of the system. In well behaved cases (analogous to the situation for convex f in the scalar case) each of these waves is either a shock or a rarefaction wave. The Riemann solution has been found for certain well known systems, for example the references [62] and [30] work through the solution for the Euler equations. Of particular interest is the work in [68] which constructs the Riemann solution for

the Saint-Venant equations and uses this as part of a Godunov approach. Even when the structure of the Riemann solution is known, it tends to be computationally expensive to compute the actual solution, since intersections of the Hugoniot and integral curve must be found (usually numerically). For this reason it is often not practical to base a numerical scheme on the exact solution of the Riemann problem. Instead an attempt is made to compute an approximation to the Riemann solution and use this to compute the numerical flux. This is the topic of the next section.

3.3 Approximate Riemann Solvers

Even in the scalar case where the simple formula (3.12) gives the Godunov flux, it may still be more efficient to use only an approximation to the Riemann solution in order to compute the numerical flux. For example one can take the solution of the Riemann problem to be the moving shock given by (3.8), regardless of whether this is the entropy satisfying solution. From (3.9) the numerical flux is then given by

$$g^{\text{FOU}}_{j+\frac{1}{2}} = \begin{cases} f(u_j) & s_{j+\frac{1}{2}} \geq 0 \\ f(u_{j+1}) & s_{j+\frac{1}{2}} < 0, \end{cases}$$

where

$$s_{j+\frac{1}{2}} = \begin{cases} \dfrac{f(u_{j+1}) - f(u_j)}{u_{j+1} - u_j} & u_{j+1} \neq u_j \\ f'(u_j) & u_{j+1} = u_j. \end{cases} \tag{3.14}$$

We can also write

$$g^{\text{FOU}}_{j+\frac{1}{2}} = \frac{1}{2}\left(f(u_j) + f(u_{j+1}) - |s_{j+\frac{1}{2}}|(u_{j+1} - u_j)\right). \tag{3.15}$$

This scheme is known as the *first-order upwind scheme* or simply the *upwind scheme* and is attributed to Roe[55], although Murman and Cole[42][41] came up with a similar scheme much earlier. The first-order upwind scheme forms the basis of many more advanced schemes.

Another important approximate Riemann solver was proposed by Engquist and Osher[15][14]. The numerical flux for this case is given by

$$g^{\text{E-O}}_{j+\frac{1}{2}} = f_-(u_{j+1}) + f_+(u_j) + f(c), \tag{3.16}$$

where the functions $f_\pm$ are given by

$$\begin{aligned} f_-(u) &= \int_c^u \min\{f'(s), 0\} ds, \\ f_+(u) &= \int_c^u \max\{f'(s), 0\} ds \end{aligned}$$

and c is arbitrary. For a convex or concave flux function this form is equivalent to the Godunov flux for a rarefaction wave and only differs in the case of a shock.

In the case of systems of conservations laws the need for an approximate Riemann solver is more pressing. By far the most used approximate Riemann solver is that due to Roe[55]. This works by linearising the system of equations at each cell interface and then calculating the flux at the interface by exactly solving the resulting linear Riemann problem. Solving a linear Riemann problem is straightforward and is described in [30]. At the interface at $x_{j+\frac{1}{2}}$ and at time t_n the linearised system is given by

$$\frac{\partial \boldsymbol{u}}{\partial t} + \tilde{J}^n_{j+\frac{1}{2}} \frac{\partial \boldsymbol{u}}{\partial x} = 0, \tag{3.17}$$

where $\tilde{J}_{j+\frac{1}{2}} = \tilde{J}(\boldsymbol{u}_{j+1}, \boldsymbol{u}_j)$ is a constant matrix which approximates the Jacobian $J(\boldsymbol{u}) = \partial \boldsymbol{f}/\partial \boldsymbol{u}$ at the interface. Roe gives the properties that the matrix $\tilde{J}$ should satisfy. These are:

(1) $\boldsymbol{f}(\boldsymbol{u}_r) - \boldsymbol{f}(\boldsymbol{u}_l) = \tilde{J}(\boldsymbol{u}_r, \boldsymbol{u}_l)(\boldsymbol{u}_r - \boldsymbol{u}_l)$ for all $\boldsymbol{u}_l$, $\boldsymbol{u}_r$.

(2) $\tilde{J}(\boldsymbol{u}_r, \boldsymbol{u}_l)$ is diagonalisable for each $\boldsymbol{u}_l$, $\boldsymbol{u}_r$

(3) For each $\boldsymbol{u}$, $\tilde{J}(\boldsymbol{u}_r, \boldsymbol{u}_l) \to J(\boldsymbol{u})$ as $\boldsymbol{u}_l$, $\boldsymbol{u}_r \to \boldsymbol{u}$.

Condition (1) is necessary to ensure the resulting scheme is conservative. Condition (2) ensures that the linearised system is hyperbolic and hence solvable. Condition (3) ensures that the the system (3.17) is a true linearisation of the nonlinear system so that the scheme is valid for smooth solutions. A matrix satisfying the above three conditions is often called a *Roe matrix*. There is not in general a unique choice for the Roe matrix for a particular problem. In his paper Roe demonstrates how to calculate a Roe matrix via an intermediate variable called a parameter vector. We specify a Roe matrix for the Saint-Venant system at the end of this chapter. Before

we state the numerical flux function we must define some notation. For the scalar quantity s, we define

$$s^+ = \frac{s+|s|}{2} = \begin{cases} 0 & s \leq 0 \\ s & s > 0, \end{cases}$$

$$s^- = \frac{s-|s|}{2} = \begin{cases} s & s \leq 0 \\ 0 & s > 0, \end{cases}$$

$$\Gamma(s) = \begin{cases} 0 & s \leq 0 \\ 1 & s > 0. \end{cases}$$

To generalise these quantities to the matrix $\tilde{J}$ we diagonalise this matrix. Suppose $\tilde{\lambda}_1, \ldots, \tilde{\lambda}_m$ denote the eigenvalues of $\tilde{J}$ and $\tilde{\boldsymbol{e}}_1, \ldots, \tilde{\boldsymbol{e}}_m$ are the corresponding eigenvectors. Define the matrices

$$\begin{aligned} \tilde{X} &= (\tilde{\boldsymbol{e}}_1, \ldots, \tilde{\boldsymbol{e}}_m), \\ \tilde{\Lambda} &= \mathrm{diag}\{\tilde{\lambda}_1, \ldots, \tilde{\lambda}_m\}, \end{aligned}$$

so that by virtue of (2) above

$$\tilde{J} = \tilde{X}\tilde{\Lambda}\tilde{X}^{-1}.$$

For the diagonal matrix $\tilde{\Lambda}$ we define

$$\begin{aligned} |\tilde{\Lambda}| &= \mathrm{diag}\{|\tilde{\lambda}_1|, \ldots, |\tilde{\lambda}_m|\}, \\ \tilde{\Lambda}^{\pm} &= \mathrm{diag}\{\tilde{\lambda}_1^{\pm}, \ldots, \tilde{\lambda}_m^{\pm}\}, \\ \Gamma(\tilde{\Lambda}) &= \mathrm{diag}\{\Gamma(\tilde{\lambda}_1), \ldots, \Gamma(\tilde{\lambda}_m)\}. \end{aligned}$$

We can now define

$$\begin{aligned} |\tilde{J}| &= \tilde{X}|\tilde{\Lambda}|\tilde{X}^{-1}, \\ \tilde{J}^{\pm} &= \tilde{X}\tilde{\Lambda}^{\pm}\tilde{X}^{-1}, \\ \Gamma(\tilde{J}) &= \tilde{X}\Gamma(\tilde{\Lambda})\tilde{X}^{-1}. \end{aligned}$$

The numerical flux function using Roe's approximate Riemann solver is now given by

$$\boldsymbol{g}^{\mathrm{Roe}}_{j+1\frac{1}{2}} = \frac{1}{2}\left(\boldsymbol{f}(\boldsymbol{u}_j) + \boldsymbol{f}(\boldsymbol{u}_{j+1}) - |\tilde{J}_{j+\frac{1}{2}}|(\boldsymbol{u}_{j+1} - \boldsymbol{u}_j)\right). \tag{3.18}$$

By comparing the numerical fluxes (3.18) and (3.15), we can see the numerical flux of Roe's approximate Riemann solver is clearly the generalisation of the first-order upwind numerical flux to the case of a system of equations.

The Engquist-Osher numerical flux is extended to systems of equations in [49]. However this approximate Riemann solution is almost as difficult to find as the exact Riemann solution.

Not all conservative difference methods are derived from approximate solutions of the Riemann problem. Take for example the Lax-Wendroff scheme[29] given by

$$g^{\text{L-W}}_{j+\frac{1}{2}} = \frac{1}{2}\left(f(u_j) + f(u_{j+1}) - \frac{\Delta t}{\Delta x}\left(s_{j+\frac{1}{2}}\right)^2 (u_{j+1} - u_j)\right),$$

and the Lax-Friedrichs scheme[8] given by

$$g^{\text{L-F}}_{j+\frac{1}{2}} = \frac{1}{2}\left(f(u_j) + f(u_{j+1}) - \frac{\Delta x}{\Delta t}(u_{j+1} - u_j)\right),$$

which are both scalar schemes constructed using Taylor's series. There tends to be a fundamental difference between those schemes constructed via Riemann solutions and those which are not. Consider a scalar problem with convex flux function and consider a smooth region of the solution where the wave speed f' is everywhere positive. The theory of characteristics yields that the solution at each point (x_j, t_{n+1}) depends only on values of the solution at points (x, t) with $x \leq x_j$ and $t \leq t_{n+1}$. In other words the solution depends only on the solution at previous times in the upwind direction. In this situation the Godunov, first-order upwind and Engquist-Osher schemes all reduce to

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + \frac{f(u_j^n) - f(u_{j-1}^n)}{\Delta x} = 0.$$

Since the value of u_j^{n+1} depends only on u_j^n and u_{j-1}^n, the schemes are mimicking the behaviour of the exact solution since the value of u_j^{n+1} depends only on the values at upwind points. For a smooth region of solution where the wave speed f' is everywhere negative, the value of the exact solution at any point (x_j, t_{n+1}) again depends only on the previous solution in the upwind direction. In this case the upwind direction is $x \geq x_j$ and the three schemes reduce to:

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + \frac{f(u_{j+1}^n) - f(u_j^n)}{\Delta x} = 0.$$

The value of u_j^{n+1} depends only on the values of u_j^n and u_{j+1}^n and so again the discrete solution only depends on the solution at upwind points. We observe that the schemes switch their behaviour depending on the on the local wave direction. Schemes which exhibit this behaviour are known as *upwind schemes*. In the case of a system of equations, upwind schemes, such as Roe's approximate Riemann solver, essentially decompose the solution into its component waves and apply a scalar upwind scheme to each individual wave. As opposed to upwind schemes, schemes such as Lax-Wendroff and Lax-Friedrichs are known as *symmetric schemes*. These schemes have a constant stencil regardless of the wave direction of the solution. Upwind schemes are in general found to be far superior for computing discontinuous solutions.

3.4 Nonlinear Stability

The Lax-Wendroff Theorem[29] mentioned in section 3.1 shows that any convergent sequence of solutions to a conservative difference method must converge to a weak solution of the conservation law. However it does not guarantee that a sequence of solutions with $\Delta x, \Delta t \downarrow 0$ will converge. To achieve this guarantee a scheme requires some level of stability. In this section we consider this problem for scalar schemes where there are many different forms of stability.

One of the weakest forms of stability is that of monotonicity preservation. A scheme is *monotonicity preserving* if given any monotone initial data, the solution remains monotone for all time. This property prevents oscillations from occuring at discontinuities.

A stronger form of stability and one of the most important is that of total variation stability. The total variation of a function is a measure of the oscillatory nature of the function. A function u on the interval $[c, d]$ has bounded total variation if

$$V_c^d u = \sup \left\{ \sum_{i=1}^{N} |u(x_i) - u(x_{i-1})| : c = x_0 < x_1 < \ldots < x_{N-1} < x_N = d \right\} < \infty,$$

in which case $V_c^d u$ is the total variation of u on $[c, d]$. If u is continuously differentiable then

$$V_c^d u = \|u'\|_1 = \int_c^d |u'(x)| dx.$$

It can be shown that for a solution of a scalar conservation law that the total variation decreases in time. It is therefore natural to require a numerical solution to also have this property. A numerical scheme for which the total variation of the solution always decreases in time is known as *Total Variation Diminishing* or TVD. Examples of TVD schemes are the Godunov, first-order upwind, Engquist-Osher and Lax-Friedrichs schemes. Note that a requirement for these schemes to be TVD is that the appropriate CFL condition must hold. We can therefore think of the CFL condition as the sole condition required so that these schemes are TVD. The TVD property again prevents spurious oscillations from developing in the solution, as happens for example in the Lax-Wendroff scheme which is not TVD.

It can be shown that any conservative, TVD method is convergent (see [30], Chapter 15). However it is not guaranteed that the limit is the unique entropy solution of the conservation law. For example, recall that the first-order upwind scheme can be derived by taking the solution of the Riemann problem to be a shock, regardless of whether this is the entropy satisfying solution or not. It is therefore not surprising that this scheme can converge to a shock solution when the appropriate solution is a rarefaction wave.

An even stronger form of stability than the TVD property is to require that a scheme be *monotone*. The entropy satisfying solution of a scalar conservation law has the following property. Suppose u_0 and v_0 are two sets of initial data such that

$$v_0(x) \geq u_0(x) \qquad \text{for all } x,$$

and that v and u denote the corresponding entropy solutions. It follows that

$$v(x,t) \geq u(x,t) \qquad \text{for all } x \text{ and } t.$$

A scheme is called monotone if it has the analogous behaviour. That is for any numerical solutions $\{u_j^n\}$ and $\{v_j^n\}$ with

$$v_j^n \geq u_j^n \qquad \text{for all } j,$$

we have

$$v_j^{n+1} \geq u_j^{n+1} \qquad \text{for all } j.$$

If we write the scheme in the form

$$u_j^{n+1} = G(u_{j+k}^n, \ldots, u_{j+1}^n, u_j^n, \ldots, u_{j-k+1}^n),$$

then the scheme is monotone if the function G is an increasing function of all of its arguments. It is shown in [8] and [23] that any conservative, monotone scheme converges to the unique entropy satisfying solution of the conservation law. Examples of monotone schemes include the Engquist-Osher, Godunov and Lax-Friedrichs schemes. Note again that this form of stability also requires the appropriate CFL condition to hold. We conclude that monotone schemes have some very nice properties. Not only are they convergent to the unique entropy solution of the conservation law, but since they must also be TVD they are non-oscillatory. The one big drawback is that they can at most be first order accurate, as proved in [23].

3.5 High Order TVD Schemes

Nearly all the numerical methods encountered so far in this chapter have been first order accurate. Such schemes give poor accuracy in smooth parts of the solution and tend to heavily smear discontinuities. This is explained by the large amount of numerical dissipation these schemes possess. Classical second order schemes such as Lax-Wendroff give good accuracy for smooth solutions but develop oscillations in the vicinity of discontinuities. These oscillations are due to the fact that such schemes have too little numerical dissipation. In this section we describe an approach which uses the TVD criteria to obtain scalar schemes which have are both non-oscillatory and give high order accuracy where the solution is smooth. This is not by any means the only approach to obtaining such schemes, for example there is the ENO (Essentially Non-Oscillatory) approach of [21].

The approach described here is to start with a second order scheme, such as the Lax-Wendroff scheme, and add to it a term which increases the numerical dissipation only in the locality of discontinuities. Equivalently, one can take a first order scheme, such as the first-order upwind scheme, and add a term to limit the numerical dissipation away from discontinuities. This can be done by constructing a new numerical

flux

$$\tilde{g}_{j+\frac{1}{2}} = g^{\mathrm{L}}_{j+\frac{1}{2}} + \delta\left(g^{\mathrm{H}}_{j+\frac{1}{2}} - g^{\mathrm{L}}_{j+\frac{1}{2}}\right),$$

where $g^{\mathrm{L}}_{j+\frac{1}{2}}$ is a numerical flux for a first order scheme and $g^{\mathrm{H}}_{j+\frac{1}{2}}$ is a numerical flux for a second order scheme. If we require that $\delta \sim 1$ in smooth regions of the solution then the scheme will be second order there. If we require $\delta \sim 0$ near discontinuities then the scheme will be expected to be non-oscillatory since it will have the same amount of numerical dissipation as for the first order scheme. To achieve the aim of being second order accurate almost everywhere as well as oscillation free the function δ, which is called a *limiter* because it limits the numerical dissipation, must be a non-linear function of the current solution. Requirements are placed on this function which ensure that the resulting scheme is TVD. Harten[20] and Sweby[67] took the first order scheme to be the first-order upwind and the second order scheme to be Lax-Wendroff to obtain the numerical flux

$$\begin{aligned}\tilde{g}_{j+\frac{1}{2}} &= \frac{1}{2}\left(f(u_j) + f(u_{j+1}) - \mathrm{sgn}(s_{j+\frac{1}{2}})(f(u_{j+1}) - f(u_j))\right) \\ &\quad + \frac{\delta}{2}\left(\mathrm{sgn}(s_{j+\frac{1}{2}}) - \frac{\Delta t}{\Delta x}(f(u_{j+1}) - f(u_j))\right).\end{aligned}$$

The particular form of the limiter function is taken to be $\delta = \delta(r_{j+\frac{1}{2}})$, where

$$r_{j+\frac{1}{2}} = \frac{u_{j+1+i} - u_{j+i}}{u_{j+1} - u_j}, \qquad i = \mathrm{sgn}(s_{j+\frac{1}{2}}).$$

This leads to a five point scheme since $\tilde{g}_{j+\frac{1}{2}} = g(u_{j+2}, u_{j+1}, u_j, u_{j-1})$. Sweby[67] also located the bounds on the limiter functions such that the scheme is both second order (in space and time) and TVD. Some suggested limiter functions which satisfy these bounds are

$$\begin{aligned}\delta(r) &= \mathrm{minmod}(1, r) = \max\{0, \min\{1, r\}\}, \\ \delta(r) &= \frac{r + |r|}{1 + r}, \\ \delta(r) &= \max\{0, \min\{2r, 1\}, \min\{2, r\}\}.\end{aligned}$$

There are other variations on constructing high order TVD methods using limiter functions, for example see [20], [71], [70] and [72]. To implement high order TVD schemes for systems of equations requires ensuring that each of the component waves is in effect updated by a high order scalar TVD scheme.

3.6 Implicit Schemes

The methods discussed so far in this chapter are all subject to the a CFL condition of the type (3.13) in order that the scheme be stable. This condition restricts the size of the time step that may be used. This may not be too much of a restriction for transient computations where the time step must also be kept small to achieve the required accuracy in time. However for steady state computations, where the accuracy of the transient solution is of no importance, we wish to take as large a time step as possible. The larger the time step we can take, the fewer times steps it takes to reach the steady state and the more economical the method is on computer CPU time.

To relax or even remove the time step restriction one can consider implicit methods. Consider the family of schemes

$$\frac{\boldsymbol{u}_j^{n+1} - \boldsymbol{u}_j^n}{\Delta t} + \theta \left(\frac{\boldsymbol{g}_{j+\frac{1}{2}}^{n+1} - \boldsymbol{g}_{j-\frac{1}{2}}^{n+1}}{\Delta x} \right) + (1-\theta) \left(\frac{\boldsymbol{g}_{j+\frac{1}{2}}^{n} - \boldsymbol{g}_{j-\frac{1}{2}}^{n}}{\Delta x} \right) = 0, \tag{3.19}$$

where $0 \leq \theta \leq 1$. This can be written in the conservative form (3.3) with the numerical flux

$$\tilde{\boldsymbol{g}}_{j+\frac{1}{2}}^n = \theta \boldsymbol{g}_{j+\frac{1}{2}}^{n+1} + (1-\theta)\boldsymbol{g}_{j+\frac{1}{2}}^n.$$

We can re-write the scheme (3.19) as

$$L_j \boldsymbol{u}^{n+1} = R_j \boldsymbol{u}^n,$$

where

$$\begin{aligned} L_j \boldsymbol{u} &= \boldsymbol{u}_j + \theta \frac{\Delta t}{\Delta x} \left(\boldsymbol{g}_{j+\frac{1}{2}} - \boldsymbol{g}_{j-\frac{1}{2}} \right) \\ R_j \boldsymbol{u} &= \boldsymbol{u}_j - (1-\theta) \frac{\Delta t}{\Delta x} \left(\boldsymbol{g}_{j+\frac{1}{2}} - \boldsymbol{g}_{j-\frac{1}{2}} \right). \end{aligned}$$

For the case $\theta = 0$, the scheme reduces to (3.3). In this case L_j is a linear operator, in fact $L_j \boldsymbol{u} = \boldsymbol{u}_j$ so that

$$\boldsymbol{u}^{n+1} = R_j \boldsymbol{u}^n,$$

and the numerical solution at the next time level is given explicitly as a function of the solution at the current time level. The scheme is hence called *explicit*. In the case $\theta \neq 0$, L_j is now a nonlinear operator (except for a linear problem where

it is linear) and the solution at the next time level is only given as an implicit function of the solution at the previous time level. The scheme is now called *implicit*. There are two important special cases of θ. For $\theta = \frac{1}{2}$ the time differencing of the scheme corresponds to the trapezium rule. This case gives second order accuracy in time even for first order numerical fluxes. The most important case here is that of $\theta = 1$. This is often called the fully implicit case because the numerical fluxes are only evaluated at the next time level. It can be shown (for example see [20]) that for many numerical fluxes the scheme is unconditionally TVD, that is TVD for all positive time steps. The need for a CFL condition can no longer be argued from a domain of dependence argument, since each solution value at the $(n+1)^{\text{th}}$ time level will in general depend on every solution value at the n^{th} time level.

The case $\theta = 1$ appears ideal for steady computations, since there is no limit to the time step allowable. The situation is not as good as it may first appear. To compute the solution at each time level requires the solution of a system of nonlinear equations. This in itself is a complex numerical problem. Even when this can be done reliably it is still likely to counterbalance the efficiency gain from using a large time step. There is however an approach which avoids the need to solve a nonlinear system, and still in general allows much greater time steps than for the explicit case. This proceeds as follows: At each time level the nonlinear operator L_j is linearised to give a linear scheme

$$\tilde{L}_j \boldsymbol{u}^{n+1} = R_j \boldsymbol{u}^n,$$

where $\tilde{L}_j$ is a linear operator. At each time level we now only have the much simpler task of solving a linear system, which more often than not will be tridiagonal. The linearisation can often be carried out so that the resulting scheme is still conservative. Of course we cannot expect the new scheme, which is called the *linearised implicit scheme* or just linearised scheme, to inherit all the properties of the original scheme. Nevertheless we hope that although the linearised scheme may not be unconditionally TVD, it will allow a much greater time step than for the explicit scheme. A review of implicit/linearised methods for high order TVD schemes can be found in [72].

3.7 Inhomogeneous Conservation Laws

Many systems of conservation laws which arise in practice include a source term. In this section we consider the modifications required to model the inhomogeneous system

$$\frac{\partial \boldsymbol{u}}{\partial t} + \frac{\partial}{\partial x}\boldsymbol{f}(\boldsymbol{u}) = \boldsymbol{b}(x, \boldsymbol{u}).$$

More often than not the left hand-side is discretised exactly as for a homogeneous system and terms are then added to the right hand side of the scheme to take account of the source term $\boldsymbol{b}$. The pointwise discretisation of the source term shown is

$$\frac{\boldsymbol{u}_j^{n+1} - \boldsymbol{u}_j^n}{\Delta t} + \frac{\boldsymbol{g}_{j+\frac{1}{2}}^n - \boldsymbol{g}_{j-\frac{1}{2}}^n}{\Delta x} = \boldsymbol{b}(x_j, \boldsymbol{u}_j^n).$$

This is the simplest possible discretisation of the source term and is often sufficient to yield satisfactory results. More advanced discretisations take into account the wave nature of the problem. Consider the first-order upwind scheme for a scalar problem. This can be derived by linearising the differential equation over each interval $[x_j, x_{j+1}]$ and using the theory of characteristics to construct a solution. Carrying this out for the inhomogeneous problem we obtain the linearised differential equation:

$$\frac{\partial u}{\partial t} + s_{j+\frac{1}{2}}\frac{\partial u}{\partial x} = \tilde{b}_{j+\frac{1}{2}},$$

where $\tilde{b}_{j+\frac{1}{2}}$ denotes some average of the source term on the left and right of the cell interface. The theory of characteristics then motivates the following scheme:

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + \frac{\left(g_{j+\frac{1}{2}}^{\mathrm{FOU}}\right)^n - \left(g_{j-\frac{1}{2}}^{\mathrm{FOU}}\right)^n}{\Delta x} = \Gamma\left(s_{j-\frac{1}{2}}^n\right)\tilde{b}_{j-\frac{1}{2}}^n + \left(1 - \Gamma\left(s_{j+\frac{1}{2}}^n\right)\right)\tilde{b}_{j+\frac{1}{2}}^n,$$

where we have used the notation of section 3.3. Combined with Roe's approximate Riemann solver this scheme can be generalised to a system of equations as follows,

$$\frac{\boldsymbol{u}_j^{n+1} - \boldsymbol{u}_j^n}{\Delta t} + \frac{\left(\boldsymbol{g}_{j+\frac{1}{2}}^{\mathrm{FOU}}\right)^n - \left(\boldsymbol{g}_{j-\frac{1}{2}}^{\mathrm{FOU}}\right)^n}{\Delta x} = \Gamma\left(\tilde{J}_{j-\frac{1}{2}}^n\right)\tilde{\boldsymbol{b}}_{j-\frac{1}{2}}^n + \left(1 - \Gamma\left(\tilde{J}_{j+\frac{1}{2}}^n\right)\right)\tilde{\boldsymbol{b}}_{j+\frac{1}{2}}^n,$$

where $\tilde{J}$ is the Roe matrix. The above is a rather crudely devised upwind discretisation of the source term. More systematic discretisations are discussed in [57] and [18].

This section has described all the source term discretisations used throughout the rest of this thesis. We will see that the given upwind discretisation with the choice

$$\tilde{\boldsymbol{b}}_{j+\frac{1}{2}} = \frac{\boldsymbol{b}(x_j, \boldsymbol{u}_j) + \boldsymbol{b}(x_{j+1}, \boldsymbol{u}_{j+1})}{2}, \tag{3.20}$$

has a very nice property, namely, at steady state and in smooth regions of the solution, the approximate solution has second order accuracy.

3.8 Roe's Approximate Riemann Solver for the Saint-Venant Equations

Roe's approximate Riemann solver has been applied to the Saint-Venant equations by many authors, for example [2], [53] and [17]. In [17] the approximate Riemann solver is used to develop a linearised implicit, high order TVD scheme. The implementation of Roe's scheme for a prismatic channel is relatively well known. Priestley[53] implements the scheme for a varying rectangular channel. In this thesis we suggest how the scheme can be extended to a general shape non-prismatic channel, and show how this is valid at least for steady state computations. For a prismatic channel the Saint-Venant equations can be written as

$$\frac{\partial \boldsymbol{w}}{\partial t} + \frac{\partial}{\partial x}\boldsymbol{F}(\boldsymbol{w}) = \boldsymbol{D}(x, \boldsymbol{w}),$$

where

$$\begin{aligned} \boldsymbol{w} &= (A, Q)^T, \\ \boldsymbol{F} &= \left(Q, \frac{Q^2}{A} + I_1\right)^T. \end{aligned}$$

We ignore the source term here as this can be treated as described in the previous section. The Roe matrix is derived using the parameter vector method of Roe[55] and is given by

$$\tilde{J}(\boldsymbol{w}_r, \boldsymbol{w}_l) = \begin{pmatrix} 0 & 1 \\ \tilde{c}^2 - \tilde{u}^2 & 2\tilde{u} \end{pmatrix}, \tag{3.21}$$

where

$$\tilde{c}^2 = \begin{cases} g\left(\dfrac{I_1(h_r) - I_1(h_l)}{A_r - A_l}\right) & A_l \neq A_r \\ \\ \dfrac{gA_l}{T(h_l)} & A_l = A_r, \end{cases}$$

and

$$\tilde{u} = \frac{\dfrac{Q_r}{\sqrt{A_r}} + \dfrac{Q_l}{\sqrt{A_l}}}{\sqrt{A_r} + \sqrt{A_l}}.$$

The eigenvalues of this matrix are given by

$$\tilde{\lambda}_1 = \tilde{u} - \tilde{c}, \qquad \tilde{\lambda}_2 = \tilde{u} + \tilde{c},$$

with corresponding eigenvectors

$$\tilde{\boldsymbol{r}}_1 = \left(1, \tilde{\lambda}_1\right)^T, \qquad \tilde{\boldsymbol{r}}_2 = \left(1, \tilde{\lambda}_2\right)^T.$$

The numerical flux for the scheme is given by

$$\boldsymbol{g}^{\text{Roe}}_{j+\frac{1}{2}} = \frac{1}{2}\left(\boldsymbol{F}(\boldsymbol{w}_j) + \boldsymbol{F}(\boldsymbol{w}_{j+1}) - |\tilde{J}_{j+\frac{1}{2}}|(\boldsymbol{w}_{j+1} - \boldsymbol{w}_j)\right).$$

In practice the scheme is not usually implemented in terms of this numerical flux. Using the property of the Roe matrix that

$$\boldsymbol{F}(\boldsymbol{w}_{j+1}) - \boldsymbol{F}(\boldsymbol{w}_j) = \tilde{J}_{j+\frac{1}{2}}(\boldsymbol{w}_{j+1} - \boldsymbol{w}_j),$$

we can write

$$\boldsymbol{g}^{\text{Roe}}_{j+\frac{1}{2}} = F(\boldsymbol{w}_j) + \tilde{J}^{-}_{j+\frac{1}{2}}(\boldsymbol{w}_{j+1} - \boldsymbol{w}_j).$$

Similarly we can write

$$\boldsymbol{g}^{\text{Roe}}_{j-\frac{1}{2}} = F(\boldsymbol{w}_j) - \tilde{J}^{+}_{j-\frac{1}{2}}(\boldsymbol{w}_j - \boldsymbol{w}_{j-1}).$$

The scheme can now be written as

$$\boldsymbol{w}_j^{n+1} = \boldsymbol{w}_j^n + \left(\boldsymbol{\Phi}^{+}_{j-\frac{1}{2}}\right)^n + \left(\boldsymbol{\Phi}^{-}_{j+\frac{1}{2}}\right)^n,$$

where

$$\boldsymbol{\Phi}^{\pm}_{j+\frac{1}{2}} = -\frac{\Delta t}{\Delta x}\tilde{J}^{\pm}_{j+\frac{1}{2}}(\boldsymbol{w}_{j+1} - \boldsymbol{w}_j).$$

In order to update the solution from time level n to $n+1$ we can use the following algorithm:

(1) For each j set $\hat{\boldsymbol{w}}_j = \boldsymbol{w}_j^n$.

(2) Compute and store $\left(\boldsymbol{\Phi}^{\pm}_{j+\frac{1}{2}}\right)^n$ for each cell interface.

(3) At each cell interface $x_{j+\frac{1}{2}}$ carry out

$$\begin{aligned} \hat{\boldsymbol{w}}_j &= \hat{\boldsymbol{w}}_j + \left(\boldsymbol{\Phi}^{-}_{j+\frac{1}{2}}\right)^n \\ \hat{\boldsymbol{w}}_{j+1} &= \hat{\boldsymbol{w}}_{j+1} + \left(\boldsymbol{\Phi}^{+}_{j+\frac{1}{2}}\right)^n . \end{aligned}$$

(4) For each j set $\boldsymbol{w}_j^{n+1} = \hat{\boldsymbol{w}}_j$.

In practice we overwrite the solution at the current time level with the solution at the next time level, without need for the intermediate solution vector $\{\hat{\boldsymbol{w}}_j\}$. The increments to the cell values are illustrated graphically in Figure 3.3. All that remains

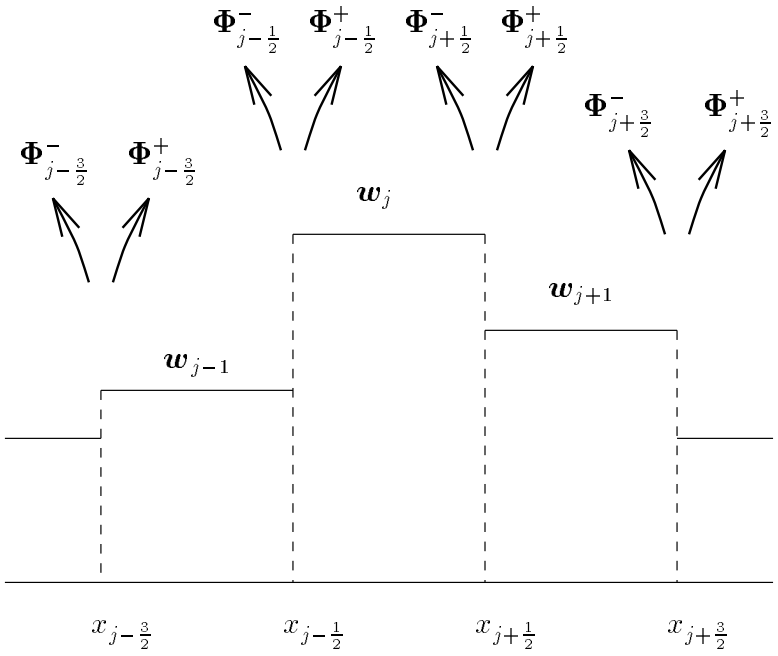


Figure 3.3: Cell updates for Roe's approximate Riemann solver

now is to see how to compute $\boldsymbol{\Phi}^{\pm}_{j+\frac{1}{2}}$ Some straightforward manipulation shows that

$$\boldsymbol{\Phi}^{\pm}_{j+\frac{1}{2}} = -\frac{\Delta t}{\Delta x}\tilde{\lambda}^{\pm}_{1,j+\frac{1}{2}}\tilde{\alpha}_{1,j+\frac{1}{2}}\tilde{\boldsymbol{r}}_{1,j+\frac{1}{2}} - \frac{\Delta t}{\Delta x}\tilde{\lambda}^{\pm}_{2,j+\frac{1}{2}}\tilde{\alpha}_{2,j+\frac{1}{2}}\tilde{\boldsymbol{r}}_{2,j+\frac{1}{2}}, \qquad (3.22)$$

where

$$\boldsymbol{w}_{j+1} - \boldsymbol{w}_j = \tilde{\alpha}_{1,j+\frac{1}{2}}\tilde{\boldsymbol{r}}_{1,j+\frac{1}{2}} + \tilde{\alpha}_{2,j+\frac{1}{2}}\tilde{\boldsymbol{r}}_{2,j+\frac{1}{2}}.$$

Solving this equation yields

$$\begin{aligned}\tilde{\alpha}_{1,j+\frac{1}{2}} &= \frac{\tilde{\lambda}_{2,j+\frac{1}{2}}(A_{j+1}-A_j)-(Q_{j+1}-Q_j)}{\tilde{\lambda}_{2,j+\frac{1}{2}}-\tilde{\lambda}_{1,j+\frac{1}{2}}},\\ \tilde{\alpha}_{2,j+\frac{1}{2}} &= \frac{-\tilde{\lambda}_{1,j+\frac{1}{2}}(A_{j+1}-A_j)+(Q_{j+1}-Q_j)}{\tilde{\lambda}_{2,j+\frac{1}{2}}-\tilde{\lambda}_{1,j+\frac{1}{2}}}.\end{aligned}$$

The algorithm can now be written as

(1) For each j set $\hat{\boldsymbol{w}}_j = \boldsymbol{w}_j^n$.

(2) For each cell interface $x_{j+\frac{1}{2}}$, compute $\tilde{\lambda}^n_{i,j+\frac{1}{2}}$, $\tilde{\alpha}^n_{i,j+\frac{1}{2}}$ for $i = 1, 2$.

(3) For each interface $x_{j+\frac{1}{2}}$ and $i = 1, 2$ carry out

$$\begin{aligned}\hat{\boldsymbol{w}}_j &= \hat{\boldsymbol{w}}_j - \frac{\Delta t}{\Delta x}\left(\tilde{\lambda}^-_{i,j+\frac{1}{2}}\right)^n \tilde{\alpha}^n_{i,j+\frac{1}{2}}\tilde{\boldsymbol{r}}^n_{i,j+\frac{1}{2}}\\ \hat{\boldsymbol{w}}_{j+1} &= \hat{\boldsymbol{w}}_{j+1} - \frac{\Delta t}{\Delta x}\left(\tilde{\lambda}^+_{i,j+\frac{1}{2}}\right)^n \tilde{\alpha}^n_{i,j+\frac{1}{2}}\tilde{\boldsymbol{r}}^n_{i,j+\frac{1}{2}}.\end{aligned}$$

(4) For each j set $\boldsymbol{w}_j^{n+1} = \hat{\boldsymbol{w}}_j$.

We have now given the implementation of the basic Roe scheme, but there are other points to consider. Since in practice we solve a problem for finite length of channel, we are required to enforce boundary conditions at the ends of the reach. Here we describe a simple approach to enforcing the boundary conditions.

3.8.1 Boundary Conditions

We describe the procedure for the boundary at $x = 0$, with the method for the remaining boundary following from an analogous argument. Consider the update of the solution from time level n to $n+1$. We assume that for each cell interface $x_{j+\frac{1}{2}}$ interior to the domain that the increments $\left(\boldsymbol{\Phi}^{\pm}_{j+\frac{1}{2}}\right)^n$ have been added to the appropriate cells. If we consider the element on the boundary at $x = 0$, then unlike interior elements which have received two increments, the boundary element has only received one increment $\left(\boldsymbol{\Phi}^{-}_{\frac{1}{2}}\right)^n$. To maintain the accuracy of the scheme at the boundary we may need to add an increment $\left(\boldsymbol{\Phi}^{+}_{-\frac{1}{2}}\right)^n$. This increment is chosen to enforce any boundary conditions. To calculate this increment requires values for

$\tilde{\lambda}^n_{i,-\frac{1}{2}}$ $(i = 1, 2)$. The simplest way to obtain values for these wave speeds is to extrapolate from inside the domain, and in particular to take

$$\tilde{\lambda}^n_{i,-\frac{1}{2}} = \tilde{\lambda}^n_{i,\frac{1}{2}}, \quad i = 1, 2.$$

There are three possible situations depending on the signs of these wave speeds. For the case $\tilde{\lambda}^n_{i,-\frac{1}{2}} \leq 0$ $(i = 1, 2)$, neither characteristic enters the domain and so no boundary conditions may be specified. The situation is very straightforward since now from (3.22) we have

$$\left(\mathbf{\Phi}^+_{-\frac{1}{2}}\right)^n = 0.$$

If $\tilde{\lambda}^n_{i,-\frac{1}{2}} > 0$ $(i = 1, 2)$, then both characteristics enter the domain and so both flow variables must be specified at the boundary. In this case we simply overwrite both flow variables at the boundary with the appropriate values. In the case where only one wave speed, say $\tilde{\lambda}^n_{i,-\frac{1}{2}}$ is positive, only one characteristic enters the domain so that only one flow variable must be specified on the boundary. Equation (3.22) gives

$$\left(\mathbf{\Phi}^+_{-\frac{1}{2}}\right)^n = -\frac{\Delta t}{\Delta x}\tilde{\lambda}^n_{i,-\frac{1}{2}}\tilde{\alpha}^n_{i,-\frac{1}{2}}\tilde{\boldsymbol{r}}^n_{i,-\frac{1}{2}}.$$

If the boundary condition is $A = A^0(t)$, then this is satisfied at time level $n+1$ if we choose $\tilde{\alpha}^n_{i,-\frac{1}{2}}$ to satisfy

$$A^0(t_{n+1}) = (1, 0)\left(\hat{\boldsymbol{w}}_0 - \frac{\Delta t}{\Delta x}\tilde{\lambda}^n_{i,-\frac{1}{2}}\tilde{\alpha}^n_{i,-\frac{1}{2}}\tilde{\boldsymbol{r}}^n_{i,-\frac{1}{2}}\right).$$

If the boundary condition is $Q = Q^0(t)$, then this is satisfied at time level $n+1$ if we choose $\tilde{\alpha}^n_{i,-\frac{1}{2}}$ to satisfy

$$Q^0(t_{n+1}) = (0, 1)\left(\hat{\boldsymbol{w}}_0 - \frac{\Delta t}{\Delta x}\tilde{\lambda}^n_{i,-\frac{1}{2}}\tilde{\alpha}^n_{i,-\frac{1}{2}}\tilde{\boldsymbol{r}}^n_{i,-\frac{1}{2}}\right).$$

3.8.2 Modifications to Roe's Scheme

One problem with Roe's linearisation is that the approximate Riemann solution consists of only shocks, and no rarefaction waves. For this reason the method can sometimes converge to weak solutions other than the entropy solution. One approach to remove this problem is given by [22] and is also discussed in [30](Chapter 18).

Roe's scheme can also be modified to give second order accuracy. The increment at each cell interface is then of the form

$$\begin{aligned} \hat{\boldsymbol{w}}_j &= \hat{\boldsymbol{w}}_j + \left(\boldsymbol{\Phi}^-_{j+\frac{1}{2}}\right)^n - \boldsymbol{B}^n_{j+\frac{1}{2}} \\ \hat{\boldsymbol{w}}_{j+1} &= \hat{\boldsymbol{w}}_{j+1} + \left(\boldsymbol{\Phi}^+_{j+\frac{1}{2}}\right)^n + \boldsymbol{B}^n_{j+\frac{1}{2}}, \end{aligned}$$

where $\boldsymbol{B}^n_{j+\frac{1}{2}}$ is a nonlinear function of $\left(\boldsymbol{\Phi}^{\pm}_{j-\frac{1}{2}}\right)^n$, $\left(\boldsymbol{\Phi}^{\pm}_{j+\frac{1}{2}}\right)^n$ and $\left(\boldsymbol{\Phi}^{\pm}_{j+\frac{3}{2}}\right)^n$. The role of the extra term is to limit the numerical dissipation of the scheme in smooth regions of the solution and so increase the order of accuracy there. Details of this approach can be found in [66]. Neither of the above modifications are used in this thesis.

Chapter 4

Theory for the Steady Flow Problem using Vanishing Viscosity

In this chapter we present some theory for the steady state Saint-Venant problem. The theory arises from a novel formulation of the problem and is applicable to a large number of cases.

4.1 Vanishing Viscosity

The Saint-Venant equations are a hyperbolic system of conservation laws. These suffer from two main difficulties, namely solutions may be discontinuous and secondly not all of these so-called weak solutions are physically possible.

Hyperbolic systems of conservation laws often arise from models of physical processes which ignore effects due to viscous or dispersive mechanisms. The next level of accuracy for any such model is to include these effects. The differential equations are modified by the addition of higher order derivatives which are multiplied by small coefficients called *viscosity coefficients*. For the original model to be consistent with the more complete model which includes the viscous or diffusive effects, it is required that the solutions of the two models are "close" in some sense. In particular any solution of the first order system must be the limit of the corresponding solution of the higher order system as the viscosity coefficients vanish. Solutions of the first order system which do have this property are known as *vanishing viscosity solutions*.

Unfortunately the two models are not generally consistent in the above sense, in that not all weak solutions of the first order system will be vanishing viscosity solutions. It is clearly only the vanishing viscosity solutions which have physical relevance.

In general the higher order system is parabolic and so always has smooth solutions. The apparent discontinuities (which form actual discontinuities in the vanishing viscosity limit) are actually narrow regions where the solution changes extremely rapidly. These regions are called *shock layers*.

The above concept is illustrated by the Euler model of gas dynamics. The Euler equations arise from neglecting terms which model the effects of fluid viscosity from the Navier-Stokes equations, the general model of fluid flow. This is done when the effects of viscosity are thought to be of only secondary importance relative to the effects of inertia. Solutions of the Euler equations, which include discontinuous solutions, are hoped to model the vanishing viscosity limit of solutions to the Navier-Stokes equations. However neglecting the viscous terms introduces solutions which are not vanishing viscosity solutions. Even though the effects of viscosity are small throughout almost all of the flow, they are sometimes still important. In particular their effects are always strong in shock layers. Viscosity prevents the solutions from becoming discontinuous and is also the mechanism for discriminating against unphysical discontinuities. There is a parallel here between the Euler equations and the Saint-Venant equations, since both systems can be derived from the Navier-Stokes equations and both models ignore viscous and diffusive effects. Extensions of the Saint-Venant system which include some of the effects of fluid viscosity are discussed in [58].

By considering the limit of solutions of "some" system of parabolic equations as the viscosity coefficients vanish, we may obtain results concerning the existence and uniqueness of physical solutions to a hyperbolic system. This approach is called the *vanishing viscosity method*. The parabolic problem will have only smooth solutions so that these may be easier to construct. The more difficult step is to then obtain estimates which are independent of the viscosity coefficients and allow passage to the limit.

Consider the scalar Cauchy problem

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x} f(u) = 0, \tag{4.1}$$

$$t > 0, \qquad -\infty < x < \infty, \qquad u(x,0) = U_0(x).$$

Equation (4.1) arises from the conservation of a quantity u transported with flux $f(u)$ and can be written in the integral form

$$\int_{x_1}^{x_2} [u]_{t_1}^{t_2}\, dx + \int_{t_1}^{t_2} [f(u)]_{x_1}^{x_2}\, dt = 0.$$

For given initial data U_0 there can be more than one weak solution to this problem, and hence if the conservation law is to model the real world, then clearly there can only be one physically relevant solution. Motivated by the above argument, the physically correct solution is defined as the vanishing viscosity solution, i.e. the limiting solution as $\epsilon \downarrow 0$ of the parabolic equation

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x} f(u) = \epsilon \frac{\partial^2 u}{\partial x^2}.$$

Oleinik[44] demonstrates the existence of a unique vanishing viscosity solution for any given initial data U_0. Furthermore, this vanishing viscosity solution is identified as the weak solution which, across all discontinuities and for all u between u_l and u_r, satisfies the entropy condition

$$\frac{f(u) - f(u_l)}{u - u_l} \geq s \geq \frac{f(u) - f(u_r)}{u - u_r}, \tag{4.2}$$

where s is the shock speed given by

$$s = \frac{f(u_r) - f(u_l)}{u_r - u_l}.$$

The entropy condition is the condition which identifies the physically allowable discontinuities, and weak solutions which satisfy the entropy condition are referred to as *entropy satisfying* or simply entropy solutions. The particular form (4.2) of the entropy condition is known as Oleinik's condition.

Ideally we would like to obtain analogous theory for the Saint-Venant system which as in section 2.1.2 can be written as

$$\frac{\partial \boldsymbol{w}}{\partial t} + \frac{\partial \boldsymbol{F}}{\partial x} = \boldsymbol{D}.$$

The analogous approach would be to consider limiting solutions as $\epsilon \downarrow 0$ of the system

$$\frac{\partial \boldsymbol{w}}{\partial t} + \frac{\partial \boldsymbol{F}}{\partial x} = \boldsymbol{D} + \epsilon M \frac{\partial^2 \boldsymbol{w}}{\partial x^2},$$

where M is a matrix such that the system is parabolic (see [62] p.254). We should note that from a physical point of view, it only makes sense to add diffusive terms to the momentum equation (the second component of the system). The theory for systems of equations is unfortunately much harder than for the scalar case and has only been achieved for some special systems (see [62], Chapters 16 and 20). The inclusion of the source term makes the situation beyond hope at this particular time, so in order to make progress, we fix attention solely on the steady state problem and this makes it possible to utilise the theory from the scalar case.

4.2 The Steady Problem

Consider the scalar problem

$$\frac{\partial h}{\partial t} + m \frac{\partial}{\partial x} F(x,h) = mD(x,h), \qquad 0 \le x \le L, \qquad t > 0, \tag{4.3}$$

$$h(x,0) = H_0(x), \qquad Q(x,t) \equiv \text{constant} > 0,$$

with appropriate boundary conditions, and $m = \pm 1$. This differential equation arises from the integral conservation law

$$\int_{x_1}^{x_2} [h]_{t_1}^{t_2} \, dx + m \int_{t_1}^{t_2} [F(x,h)]_{x_1}^{x_2} \, dt = m \int_{t_1}^{t_2} \int_{x_1}^{x_2} D(x,h) dx dt,$$

where $t_2 \ge t_1 \ge 0$ and $0 \le x_1 \le x_2 \le L$ are arbitrary. At steady state this reduces to

$$[F(x,h)]_{x_1}^{x_2} = \int_{x_1}^{x_2} D(x,h) dx.$$

This relationship is clearly identical to (2.19) so that at steady state the Saint-Venant equations and equation (4.3) have identical weak solutions, even though their transient behaviour is totally unrelated. If we assume that Oleinik's condition (4.2) continues to identify the correct discontinuities for problem (4.3), because source terms generally have no influence at discontinuities, then at steady state (s=0) this condition reduces to the requirement that

$$m \left(\frac{F(x,h) - F(x,h_l)}{h - h_l} \right) \ge 0, \tag{4.4}$$

for all h between h_l and h_r. If we take $m = -1$ then it is not difficult to see that this condition implies

$$E(x, h_r) \leq E(x, h_l), \tag{4.5}$$

because of the relationship (2.27). Thus we conclude that at steady state, any entropy satisfying solution of (4.3) (with $m = -1$) must also be a physical solution of the Saint-Venant equations. The converse is not necessarily true, however we show later that it is true for a certain class of channel geometries. The above observation means that we can compute steady solutions to the Saint-Venant system via computing steady solutions of the scalar differential equation

$$\frac{\partial h}{\partial t} - \frac{\partial}{\partial x} F(x, h) = -D(x, h). \tag{4.6}$$

To obtain results concerning the existence and uniqueness of entropy solutions for this problem we could study the viscous problem

$$\frac{\partial h}{\partial t} - \frac{\partial}{\partial x} F(x, h) = -D(x, h) + \epsilon \frac{\partial^2 h}{\partial x^2},$$

where $\epsilon > 0$. However since we are only interested in the solutions at steady state, we need only consider the steady form

$$\epsilon h'' + F(x, h)' = D(x, h). \tag{4.7}$$

We use the notation:

$$F(x, h)' = \frac{d}{dx} F(x, h) = \frac{\partial F}{\partial x} + h' \frac{\partial F}{\partial h}.$$

The differential equation (4.7) is the topic of this chapter and we present results concerning the existence and uniqueness of solutions both for positive ϵ and in the vanishing viscosity limit. One final point to note is that if we take $m = +1$ then, although the steady solutions of both (4.3) and the steady Saint-Venant system are again identical, in this case the entropy satisfying solutions of (4.3) violate (4.5).

4.3 The "Viscous" Problem

The differential form of the steady flow problem is given by

$$F(x, h)' = D(x, h).$$

Shocks may occur along a particular reach of channel, $0 \leq x \leq L$, and hence this equation will not in general hold everywhere. Motivated by the previous section we choose to study the following problem:

$$\begin{aligned} \epsilon h_\epsilon'' + F(x, h_\epsilon)' = D(x, h_\epsilon), \qquad h_\epsilon > 0, \qquad 0 \leq x \leq L, \\ h_\epsilon(0) = \gamma_0, \qquad h_\epsilon(L) = \gamma_1, \end{aligned} \tag{4.8}$$

where ϵ, γ_0, $\gamma_1 > 0$. We have added a viscous term to the differential equation, so that the differential equation is now second order and requires two boundary conditions. The simplest choice is Dirichlet boundary conditions. Initially the need for two boundary conditions appears as though it could cause trouble because the steady flow problem may have the depth specified at both ends, either end or neither end of the channel. This problem is resolved by the nonuniform nature of the limiting process, since solutions of the viscous problem are only used to define a solution in the vanishing viscosity limit, and the limit solution will not necessarily satisfy either of the boundary conditions given to the viscous problem. For the method to be useful we must be able to control the behaviour of the limiting solution by the boundary values γ_0 and γ_1.

4.4 Singular Perturbation Problems

There is a large amount of literature on the subject of two-point boundary value problems which depend on a small parameter ϵ, for example see [50], [45] and [25]. These problems can be separated into two categories. If the solution converges uniformly in x as $\epsilon \downarrow 0$, then the problem is called a *regular perturbation problem.* When the solution does not have a uniform limit in x then the problem is called a *singular perturbation problem.* Problem (4.8) is in the latter class and this stems from the fact that the order of the differential equation reduces from second order to first order as ϵ vanishes. Therefore in general there must be nonuniform behaviour, since we cannot expect a solution of the reduced problem ($\epsilon = 0$) to always satisfy both boundary values. The nonuniformities are classified depending on their type and where they occur. We now demonstrate some of these.

Consider the problem

$$\epsilon u_\epsilon'' + a u_\epsilon' = 0, \quad 0 \le x \le 1, \quad u_\epsilon(0) = 1, \quad u_\epsilon(1) = 0, \tag{4.9}$$

where $\epsilon > 0$ and $a \neq 0$ is a constant. The solution to this problem is given by

$$u_\epsilon(x) = 1 - \frac{e^{-\frac{ax}{\epsilon}} - 1}{e^{-\frac{a}{\epsilon}} - 1}.$$

First consider the case $a > 0$, where for small ϵ the solution decreases rapidly from one to zero near $x = 0$. In fact as ϵ tends to zero we have

$$u_\epsilon(x) \to \begin{cases} 1 & x = 0 \\ 0 & x \neq 0. \end{cases}$$

The nonuniform behaviour at $x = 0$ is known as a boundary layer and is characterised by the property:

$$1 = \lim_{\epsilon \downarrow 0} \lim_{x \downarrow 0} u_\epsilon(x) \neq \lim_{x \downarrow 0} \lim_{\epsilon \downarrow 0} u_\epsilon(x) = 0.$$

For the case $a < 0$ we have

$$u_\epsilon(x) \to \begin{cases} 1 & x \neq 1 \\ 0 & x = 1, \end{cases}$$

as ϵ tends to zero. This corresponds to a boundary layer at $x = 1$.

Next consider the problem:

$$\epsilon u_\epsilon'' + x u_\epsilon' = 0, \quad -1 \le x \le 1, \quad u_\epsilon(-1) = 1, \quad u_\epsilon(1) = 2, \tag{4.10}$$

where $\epsilon > 0$. The solution to this problem is given by

$$u_\epsilon(x) = 1 + \frac{\mathrm{erf}(\frac{x}{\sqrt{2\epsilon}}) + \mathrm{erf}(\frac{1}{\sqrt{2\epsilon}})}{2\mathrm{erf}(\frac{1}{\sqrt{2\epsilon}})},$$

where

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-s^2} ds.$$

As ϵ tends to zero we have

$$u_\epsilon(x) \to \begin{cases} 1 & x < 0 \\ 3/2 & x = 0 \\ 2 & x > 0. \end{cases}$$

In this case the nonuniformity is in the interior of the domain and as ϵ vanishes the solution tends to a discontinuity at $x = 0$. For this reason, this type of nonuniformity is known as a shock layer. There are other types of nonuniformity that are possible, for example corner layers where the limit is continuous but has a discontinuous first derivative.

The examples given above are all linear problems and the theory for such problems is well understood (for example see [50]). It is usually possible to predict in advance from the differential equation, the type and the position of the nonuniformities. For nonlinear problems this is not the case and the situation is considerably more complicated. Analysis of simple nonlinear problems can be found in [45] and [25]. These make use of asymptotic techniques and usually rely on being able to integrate the reduced differential equation.

Integration of the reduced differential equation is not possible for problem (4.8) and so another approach is required. It happens that theory exists for a class of problems which are very closely related to problem (4.8). This theory comes from a functional analysis approach (as opposed to an asymptotic approach) and applies to a general class of problems. The theory requires some adaptation before it can be applied to (4.8).

4.5 Functions of Bounded Variation

The theory in this chapter will make use of the class of functions which have bounded total variation. The term bounded total variation was defined in section 3.4, and we define $BV[c, d]$ to be the set of real functions on $[c, d]$ which have bounded total variation. A function $u \in BV[c, d]$ has the following properties:

(1) The function is bounded.

(2) All points of discontinuity are simple ($u(x-)$ and $u(x+)$ exist) and the set of discontinuities is countable. Also $u(c+)$ and $u(d-)$ exist.

We consider functions in $BV[0, L]$ which satisfy the integral relationship (2.19). A more common method of defining weak solutions is through the use of test functions

(see [30], [62]). A function h is then a weak solution of the steady flow problem if

$$\int_0^L \phi'(x)F(x,h(x)) + \phi(x)D(x,h(x))dx = 0, \tag{4.11}$$

for all $\phi \in C_0^1(0,L)$. Here $C_0^1(0,L)$ denotes the set of all continuously differentiable functions with compact support on the interval $(0,L)$. Like the integral relationship (2.19), this form requires solutions to have very little smoothness. For any given solution h we can modify the value at countably many points in an arbitrary manner, and still obtain a weak solution. Hence equation (4.11) does not define a unique value for the solution at every point, in fact it does not define a unique value anywhere. The form (2.19) is more prescriptive in that a unique value for $F(x,h(x))$ is defined at each point, however this does not define a unique value for $h(x)$. To obtain results concerning uniqueness of weak solutions, we group all the elements in the space $BV[0,L]$ into equivalence classes of almost everywhere equal functions. We then construct a normalised space $NBV[0,L]$ to contain a single representative function from each equivalence class, as follows:

$$NBV[c,d] = \{u \in NBV[c,d] : u(d) = u(d-), u(x) = u(x+) \text{ for } x \in [c,d)\}.$$

For each $u \in BV[c,d]$ there is a unique $\hat{u} \in NBV[c,d]$ such that $u = \hat{u}$ almost everywhere. Not only is this normalisation consistent with the relationship (2.19), but it is also sensible from a physical point of view. We expect any physical solution to have only a finite number of discontinuities, and such a solution in $NBV[0,L]$ will be a piecewise continuous function. Other possible normalisations may allow solutions in $NBV[0,L]$ to have points of isolated value, and such functions would not represent a realistic depth profile.

Only positive depths make sense for solutions to the steady flow problem. Furthermore we only consider solutions which are bounded below away from zero. If we allowed the depth to become arbitrarily close to zero, then this would mean that certain physical quantities, such as the energy, become unbounded. Hence we only consider solutions in the set $NBV_+[0,L]$, where we define

$$NBV_+[c,d] = \{u \in NBV[c,d] : u(x) \geq C > 0 \text{ for } c \leq x \leq d, \text{ for a constant } C\}.$$

4.6 The Theory of Lorenz

In this section we adapt theory from the literature so that it can be applied to problem (4.8). The argument is based on work by Lorenz in [32] and can be summarised in the following theorem.

Theorem 1 ([32]) *Consider the two point boundary value problem*

$$\begin{aligned} \epsilon u_\epsilon'' - f(u_\epsilon)' = b(x, u_\epsilon), \qquad 0 \le x \le 1, \\ u_\epsilon(0) = \gamma_0, \quad u_\epsilon(1) = \gamma_1, \end{aligned} \tag{4.12}$$

where $\epsilon > 0$. Suppose that $f \in C^2(-\infty, \infty)$, $b \in C^1([0,1] \times (-\infty, \infty))$ and that for some constant δ

$$b_u \ge \delta > 0, \tag{4.13}$$

for all u and all $x \in [0,1]$, then under these conditions the following hold:

(1) The problem has a unique solution $u_\epsilon \in C^2[0,1]$ for all $\epsilon > 0$.

(2) The solution is uniformly bounded in ϵ, i.e. $\|u_\epsilon\|_\infty \le K_0$ for all $\epsilon > 0$, where K_0 is independent of ϵ.

(3) The solution has total variation bounded in ϵ, i.e. $\|u_\epsilon'\|_1 \le K_1$ for all $\epsilon > 0$, where K_1 is independent of ϵ.

(4) There is a unique function $U \in NBV[0,1]$ such that $u_\epsilon \to U$ in L_1 as $\epsilon \downarrow 0$.

(5) $u = U$ is the only function in $NBV[0,1]$ which satisfies the following:

(i) If I is an interval where u is continuous, then $f(u(x))$ is differentiable on I, one-sided at end points, and the differential equation

$$-f(u)' = b(x, u),$$

holds on I. (4.14)

(ii) If u is discontinuous at $x \in (0,1)$, then

$$\begin{aligned} f(u_l) = f(u_r) \ge f(k) \quad \text{if} \quad u_l > u_r, \\ f(u_l) = f(u_r) \le f(k) \quad \text{if} \quad u_l < u_r, \end{aligned}$$

for all k between $u_l = u(x-)$ and $u_r = u(x+)$. (4.15)

$$\left.\begin{array}{l} \textit{(iii)} \quad \textit{For } j = 0,1 \textit{ and } k \textit{ between } u(j) \textit{ and } \gamma_j \\ \qquad\qquad (-1)^{j+1} sgn(u(j) - \gamma_j)(f(u(j)) - f(k)) \geq 0, \\ \qquad \textit{where } sgn(x) = -1,0,1 \textit{ for } x < 0, = 0, > 0, \textit{ respectively.} \end{array}\right\} \quad (4.16)$$

The above theory relates to a problem closely resembling problem (4.8). This is made clearer by a transformation onto the unit interval given by

$$\begin{aligned} u_\epsilon(x) &\equiv h_\epsilon(xL), \\ f(x,u) &\equiv -LF(xL,u), \\ b(x,u) &\equiv L^2 D(xL,u). \end{aligned} \quad (4.17)$$

Theorem 1 will be adapted to apply to this problem under certain conditions. To do this requires some understanding of Theorem 1 and how it is constructed.

Part 1 of the theorem gives the existence and uniqueness of the solution to the singular perturbation problem for each positive ϵ. The existence proof relies on Nagumo's Lemma ([43],[27]), which uses the fact that the problem has both upper and lower solutions. The functions $\overline{u}(x)$, $\underline{u}(x)$ are upper and lower solutions, respectively, if the following hold for all x in the interval $[0,1]$:

(1) $\underline{u} \leq \overline{u}$

(2) $\epsilon \underline{u}'' - f(\underline{u})' - b(x,\underline{u}) \leq 0$

(3) $\underline{u}(0) \leq \gamma_0, \quad \underline{u}(1) \leq \gamma_1$

(4) $\epsilon \overline{u}'' - f(\overline{u})' - b(x,\overline{u}_\epsilon) \geq 0$

(5) $\overline{u}(0) \geq \gamma_0, \quad \overline{u}(1) \geq \gamma_1$

The condition $b_u \geq \delta > 0$ ensures that the constant functions

$$\begin{aligned} \underline{u} &\equiv \min_{0 \leq x \leq 1} \left\{ 0, \frac{b(x,0)}{\delta}, \gamma_0, \gamma_1 \right\} \leq 0, \\ \overline{u} &\equiv \max_{0 \leq x \leq 1} \left\{ 0, \frac{b(x,0)}{\delta}, \gamma_0, \gamma_1 \right\} \geq 0, \end{aligned}$$

are upper and lower solutions. Nagumos's Lemma gives the existence of a solution satisfying $\underline{u} \leq u_\epsilon \leq \overline{u}$ $(0 \leq x \leq 1)$. The uniqueness of the solution is given by an inverse monotonicity argument and relies on the fact that $b_u > 0$.

The uniform bound of part 2 of the theorem comes directly from the existence proof, since the upper and lower solutions are independent of ϵ. This bound and the uniform bound on the total variation, from part 3 of the theorem, gives that the set $\{u_\epsilon\}_{\epsilon>0}$ is precompact in $L_1[0,1]$. Thus for any positive null sequence $S = \{\epsilon_n\}$, there is a subsequence $S' = \{\epsilon_{n_k}\}$ and a function $U \in NBV[0,1]$ such that

$$u_\epsilon \to U \text{ in } L_1 \text{ as } \epsilon \downarrow 0, \epsilon \in S'.$$

Part 5 of the theorem gives the properties of the limit function and states that there is exactly one function in $NBV[0,1]$ with these properties.

4.7 The Modified Theory

The main difference between problem (4.8) and the problem in Theorem 1 is that h is restricted to being positive for problem (4.8). With such a restriction, the condition $b_u \geq \delta > 0$ is no longer sufficient to guarantee the existence of upper and lower solutions, and in any case this condition does not hold for our particular function b. This can be remedied by the addition of conditions (4.20) which explicitly ensure the existence of upper and lower solutions. The condition $b_u > 0$ is, however, still required. The adapted theorem is as follows.

Theorem 2 *Consider the problem P_ϵ given by*

$$\begin{gathered} \epsilon u_\epsilon'' - f(u_\epsilon)' = b(x, u_\epsilon), \qquad u_\epsilon > 0, \qquad 0 \leq x \leq 1, \\ u_\epsilon(0) = \gamma_0, \qquad u_\epsilon(1) = \gamma_1, \end{gathered} \tag{4.18}$$

where ϵ, γ_0, $\gamma_1 > 0$, $f \in C^2(0,\infty)$, $b_x, b_u, b_{ux} \in C([0,1] \times (0,\infty))$ and

$$b_u > 0, \tag{4.19}$$

for all $u > 0$ and $x \in [0,1]$. Suppose also that there are positive constants m, M such that

$$b(x,m) \leq 0 \quad \textit{and} \quad b(x,M) \geq 0 \quad \textit{for all } x \in [0,1], \tag{4.20}$$

then under these conditions the following hold:

(1) *Problem P_ϵ has a unique solution $u_\epsilon \in C^2[0,1]$ for all $\epsilon > 0$ and this satisfies the bounds*

$$0 < \underline{u} \le u_\epsilon \le \bar{u} \qquad (0 \le x \le 1), \tag{4.21}$$

where $\underline{u} = \min\{\gamma_0, \gamma_1, m\}$ and $\overline{u} = \max\{\gamma_0, \gamma_1, M\}$.

(2) *$\|u_\epsilon'\|_1 \le K_1$ for all $\epsilon > 0$, where K_1 is independent of ϵ.*

(3) *There is a unique function $U \in NBV_+[0,1]$ such that $u_\epsilon \to U$ in L_1 as $\epsilon \downarrow 0$. The function U satisfies the bounds*

$$0 < \underline{u} \le U \le \bar{u} \qquad (0 \le x \le 1). \tag{4.22}$$

(4) *$u = U$ is the only function in $\in NBV_+[0,1]$ that has properties* (4.14), (4.15) *and* (4.16).

Proof of Theorem 2

Suppose that $0 < \alpha \le \underline{u}$, $\beta \ge \overline{u}$ and consider the problem $P_\epsilon^{\alpha,\beta}$ given by

$$\epsilon u_\epsilon'' - f^{\alpha,\beta}(u_\epsilon)' = b^{\alpha,\beta}(x, u_\epsilon), \quad 0 \le x \le 1, \quad \epsilon > 0$$
$$u_\epsilon(0) = \gamma_0, \qquad u_\epsilon(1) = \gamma_1,$$

where

$$f^{\alpha,\beta}(u) = \begin{cases} \Big(\frac{1}{2}(f''(\beta) + 2f'(\beta) + f(\beta))(u-\beta)^2 \\ +(f'(\beta) + f(\beta))(u-\beta) + f(\beta)\Big)e^{\beta - u} & u > \beta \\ \\ f(u) & \alpha \le u \le \beta \\ \\ \Big(\frac{1}{2}(f''(\alpha) - 2f'(\alpha) + f(\alpha))(u-\alpha)^2 \\ +(f'(\alpha) - f(\alpha))(u-\alpha) + f(\alpha)\Big)e^{u - \alpha} & u < \alpha, \end{cases}$$

and

$$b^{\alpha,\beta}(x,u) = \begin{cases} b(x,\beta) + (u-\beta)b_u(x,\beta) & u > \beta \\ b(x,u) & \alpha \le u \le \beta \\ b(x,\alpha) + (u-\alpha)b_u(x,\alpha) & u < \alpha. \end{cases}$$

Using this problem we prove the theorem by a sequence of results.

(1) *Problem $P_\epsilon^{\alpha,\beta}$ satisfies the conditions of Theorem 1.*

The function $f^{\alpha,\beta}$ is constructed to be continuous and have continuous first and second derivatives at both $u = \alpha$ and $u = \beta$, hence $f^{\alpha,\beta} \in C^2(-\infty, \infty)$. Also

$$b_u^{\alpha,\beta}(x,u) = \begin{cases} b_u(x,\beta) & u > \beta \\ b_u(x,u) & \alpha \le u \le \beta \\ b_u(x,\alpha) & u < \alpha, \end{cases}$$

and

$$b_x^{\alpha,\beta}(x,u) = \begin{cases} b_x(x,\beta) + (u-\beta)b_{ux}(x,\beta) & u > \beta \\ b_x(x,u) & \alpha \le u \le \beta \\ b_x(x,\alpha) + (u-\alpha)b_{ux}(x,\alpha) & u < \alpha, \end{cases}$$

so since b_x, b_u, b_{ux} are are continuous on $[0,1] \times (0,\infty)$, it follows that $b^{\alpha,\beta} \in C^1([0,1] \times (-\infty,\infty))$. Finally

$$b_u^{\alpha,\beta} \ge \delta > 0,$$

where

$$\delta = \min_{\substack{0 \le x \le 1 \\ \alpha \le u \le \beta}} \{b_u(x,u)\}.$$

(2) *Solutions u_ϵ of problem P_ϵ satisfy the bounds* (4.21).

Suppose that u_ϵ satisfies P_ϵ and that $u_\epsilon > \overline{u}$ for some $x \in [0,1]$. It follows that there must be a point $x^\star \in (0,1)$ such that

$$u_\epsilon(x^\star) > \overline{u}, \quad u_\epsilon'(x^\star) = 0, \quad u_\epsilon''(x^\star) \le 0.$$

The differential equation at $x = x^\star$ then reduces to

$$b(x^\star, u_\epsilon(x^\star)) = \epsilon u_\epsilon''(x^\star) \le 0.$$

This is a contradiction since, because $u_\epsilon(x^\star) > \overline{u} \ge M$ and $b_u > 0$, we must have

$$b(x^\star, u_\epsilon(x^\star)) > b(x^\star, M) \ge 0.$$

We conclude that $u_\epsilon \le \overline{u}$ for $0 \le x \le 1$.

Next suppose that $u_\epsilon < \underline{u}$ for some $x \in [0,1]$. It follows that there must be a point $x^\star \in (0,1)$ such that

$$u_\epsilon(x^\star) < \overline{u}, \quad u_\epsilon'(x^\star) = 0, \quad u_\epsilon''(x^\star) \geq 0.$$

The differential equation at $x = x^\star$ then reduces to

$$b(x^\star, u_\epsilon(x^\star)) = \epsilon u_\epsilon''(x^\star) \geq 0,$$

but this is again a contradiction since, because $u_\epsilon(x^\star) < \underline{u} \leq m$, we must have

$$b(x^\star, u_\epsilon(x^\star)) < b(x^\star, m) \leq 0.$$

We conclude that $u_\epsilon \geq \underline{u}$ for $0 \leq x \leq 1$.

(3) *Solutions u_ϵ of problem $P_\epsilon^{\alpha,\beta}$ satisfy the bounds* (4.21).

This follows from an identical argument to (2) above.

(4) *The function u_ϵ is a solution of problem P_ϵ if and only if it is a solution of problem $P_\epsilon^{\alpha,\beta}$.*

Suppose that u_ϵ is a solution of $P_\epsilon^{\alpha,\beta}$, then the bounds (4.21) imply that

$$\epsilon u_\epsilon'' - f(u_\epsilon)' - b(x, u_\epsilon) = \epsilon u_\epsilon'' - f^{\alpha,\beta}(u_\epsilon)' - b^{\alpha,\beta}(x, u_\epsilon) = 0,$$

thus u_ϵ also satisfies P_ϵ. Conversely, suppose that u_ϵ is a solution of P_ϵ, then the bounds (4.21) imply that

$$\epsilon u_\epsilon'' - f^{\alpha,\beta}(u_\epsilon)' - b^{\alpha,\beta}(x, u_\epsilon) = \epsilon u_\epsilon'' - f(u_\epsilon)' - b(x, u_\epsilon) = 0,$$

so that u_ϵ also satisfies $P_\epsilon^{\alpha,\beta}$.

(5) *Parts 1 and 2 of the theorem hold.*

Since Theorem 1 holds for problem $P_\epsilon^{\alpha,\beta}$, this problem has a unique solution for each $\epsilon > 0$. Now since the solutions of P_ϵ are exactly those of $P_\epsilon^{\alpha,\beta}$, then clearly problem P_ϵ has a unique solution for each $\epsilon > 0$. This gives part 1 of Theorem 2. Part 2 follows immediately from part 3 of Theorem 1.

(6) *There is a unique function* $U \in NBV_+[0,1]$ *such that* $u_\epsilon \to U$ *in* L_1 *as* $\epsilon \downarrow 0$. *The function* U *satisfies the bounds* (4.22).

Applying part 4 of Theorem 1 to $P_\epsilon^{\alpha,\beta}$ gives that there is a unique function $U \in NBV[0,1]$ such that $u_\epsilon \to U$ in L_1 as $\epsilon \downarrow 0$. We show that U satisfies the bounds (4.22) and hence is in $NBV_+[0,1]$.

We can choose a positive sequence $S = \{\epsilon_n\}$ such that $\epsilon_n \to 0$ as $n \to \infty$ and

$$u_\epsilon \to U \text{ a.e. as } \epsilon \downarrow 0, \epsilon \in S.$$

Define the set

$$X = \{x \in [0,1] : u_\epsilon(x) \to U(x) \text{ as } \epsilon \downarrow 0, \epsilon \in S\}. \tag{4.23}$$

The set $[0,1]\backslash X$ has zero measure and the bounds (4.22) clearly hold for all $x \in X$. Now for arbitrary x in $[0,1)$ by definition of the set NBV we have

$$\begin{aligned} U(x) &= \lim_{\substack{s \downarrow x \\ s \in X}} U(s), \\ U(1) &= \lim_{\substack{s \uparrow 1 \\ s \in X}} U(s), \end{aligned}$$

and thus the bounds (4.22) hold at all points, giving part 3 of Theorem 2.

(7) *Properties* (4.14), (4.15) *and* (4.16) *hold for the function* $u = U$.

This is simply a matter of writing down part 5 of Theorem 1 as applied to problem $P_\epsilon^{\alpha,\beta}$. The bounds on U are then used to replace $f^{\alpha,\beta}$ by f and $b^{\alpha,\beta}$ by b.

(8) *The function* $u = U$ *is the only function in* $NBV_+[0,1]$ *which satisfies properties* (4.14), (4.15) *and* (4.16).

Suppose that $u = \hat{u} \in NBV_+[0,1]$ satisfies properties (4.14), (4.15) and (4.16), then taking

$$\begin{aligned} \alpha &= \min_{0 \le x \le 1} \{\hat{u}(x), \underline{u}\}, \\ \beta &= \max_{0 \le x \le 1} \{\hat{u}(x), \overline{u}\}, \end{aligned}$$

and writing down properties (4.14), (4.15) and (4.16) it is clear that f can be replaced by $f^{\alpha,\beta}$ and b replaced by $b^{\alpha,\beta}$. Part 5 of Theorem 1 applied to

$P_\epsilon^{\alpha,\beta}$ then gives that $u = U$ is the only function in $NBV[0,1]$ which has these properties. Hence we must have $\hat{u} = U$, showing that part 4 of Theorem 2 holds.

This completes the proof of Theorem 2.

4.8 Application to the Steady Flow Problem

In this section we apply the theory derived in the previous section to the steady flow problem for a prismatic channel, to yield the following theorem.

Theorem 3 *Consider a prismatic channel and suppose that the following hold:*

(1) *The channel width T is continuously differentiable and positive for $h > 0$.*

(2) *The discharge Q is positive.*

(3) *The bed slope S_0 is positive and continuously differentiable for $0 \le x \le L$.*

(4) *The conveyance $K(h)$ satisfies:*

$$\left.\begin{array}{ll} \textit{(i)} & K > 0 \textit{ is continuously differentiable for } h > 0. \\ \textit{(ii)} & \dfrac{K}{\sqrt{A}} \textit{ is increasing for } h > 0. \\ \textit{(iii)} & K \to 0 \textit{ as } h \downarrow 0. \\ \textit{(iv)} & K \to \infty \textit{ as } h \to \infty. \end{array}\right\} \qquad (4.24)$$

Under the above conditions the following hold:

(1) *The problem (4.8) has a unique solution $h_\epsilon \in C^2[0,L]$ for all $\epsilon > 0$, and this satisfies the bounds*

$$0 < \underline{h} \le h_\epsilon \le \bar{h} \qquad (0 \le x \le L),$$

where

$$\underline{h} = \min_{0 \le x \le L} \{h_n(x), \gamma_0, \gamma_1\}, \quad \bar{h} = \max_{0 \le x \le L} \{h_n(x), \gamma_0, \gamma_1\},$$

and $h_n(x)$ is the normal depth which satisfies

$$K(h_n(x)) = \frac{Q}{\sqrt{S_0(x)}}, \tag{4.25}$$

for $0 \le x \le L$.

(2) $\|h'_\epsilon\|_1 \le K_1$ *for all $\epsilon > 0$, where K_1 is independent of ϵ.*

(3) *There is a unique function $H \in NBV_+[0,L]$ such that $h_\epsilon \to H$ in L_1 as $\epsilon \downarrow 0$, and this satisfies the bounds*

$$0 < \underline{h} \le H \le \bar{h} \qquad (0 \le x \le L). \tag{4.26}$$

(4) *$h = H$ is the only function in $NBV_+[0,L]$ which satisfies the following:*

(i) *If I is an interval where u is continuous, then $F(h(x))$ is differentiable on I, one-sided at end points, and the differential equation*

$$F(h)' = D(x,h),$$

holds on I. (4.27)

(ii) *If h is discontinuous at $x \in (0,L)$, then*

$$F(h_l) = F(h_r) \le F(k) \quad \text{if} \quad h_l > h_r,$$
$$F(h_l) = F(h_r) \ge F(k) \quad \text{if} \quad h_l < h_r,$$

for all k between $h_l = h(x-)$ and $h_r = h(x+)$. (4.28)

(iii) *For $j = 0, 1$ and k between $h(jL)$ and γ_j*

$$(-1)^{j+1} sgn(h(jL) - \gamma_j)(F(h(jL)) - F(k)) \le 0,$$

where $sgn(x) = -1, 0, 1$ for $x < 0, = 0, > 0$, respectively. (4.29)

Proof of Theorem 3 This is simply a matter of showing that Theorem 2 is satisfied for

$$\begin{aligned} f(u) &= -LF(u), \\ b(x,u) &= L^2 D(xL,u), \\ m &= \min_{0 \le x \le L} \{h_n(x)\}, \\ M &= \max_{0 \le x \le L} \{h_n(x)\}. \end{aligned} \tag{4.30}$$

Firstly we have that

$$F(h) = \frac{Q^2}{A(h)} + gI_1(h),$$

so that

$$F'(h) = -\frac{Q^2 T(h)}{A(h)^2} + gA(h)$$

and

$$F''(h) = \frac{2Q^2 T(h)^2}{A(h)^3} - \frac{Q^2 T'(h)}{A(h)^2} + gT(h).$$

The condition that T is positive and continuously differentiable is sufficient to ensure that F and hence f is in $C^2(0,\infty)$.

Next we have

$$D(x,h) = gA(h)\left(S_0(x) - \frac{Q^2}{K(h)^2}\right),$$

so that

$$D_h(x,h) = gT(h)S_0(x) - gQ^2\frac{d}{dh}\left(\frac{A(h)}{K(h)^2}\right),$$

$$D_x(x,h) = gA(h)S_0'(x)$$

and

$$D_{hx}(x,h) = gT(h)S_0'(x).$$

The assumptions on T, K and S_0 ensure that $D_h, D_x, D_{hx} \in C([0,L]\times(0,\infty))$ and thus $b_u, b_x, b_{ux} \in C([0,1]\times(0,\infty))$. Next we observe that

$$\frac{d}{dh}\left(\frac{A(h)}{K(h)^2}\right) = \frac{d}{dh}\left(\frac{K(h)}{\sqrt{A(h)}}\right)^{-2} = -2\left(\frac{K(h)}{\sqrt{A(h)}}\right)^{-3}\frac{d}{dh}\left(\frac{K(h)}{\sqrt{A(h)}}\right) \le 0,$$

because $K/\sqrt{A}$ is increasing in h. The assumption that $S_0 > 0$ implies that D_h and hence b_u is positive. Lastly we observe that

$$K'(h) = \frac{d}{dh}\left(\sqrt{A(h)}\frac{K(h)}{\sqrt{A(h)}}\right) = \sqrt{A(h)}\frac{d}{dh}\left(\frac{K(h)}{\sqrt{A(h)}}\right) + \frac{T(h)K(h)}{2A(h)} > 0,$$

so since $K(h) \to 0$ as $h \downarrow 0$ and $K(h) \to \infty$ as $h \to \infty$, it follows that for each $x \in [0,L]$ there is a unique $h_n(x) > 0$ satisfying (4.25). Because $K(h_n(x)) = Q/\sqrt{S_0(x)}$ is continuous on $[0,L]$, it is bounded on this interval and attains its bounds. Therefore there exist x_0, $x_1 \in [0,L]$ such that $K(h_n(x_0)) \le K(h_n(x)) \le K(h_n(x_1))$ for all $x \in [0,L]$. Since K is strictly increasing in h it follows that

$$0 < h_n(x_0) \le h_n(x) \le h_n(x_1) \qquad (0 \le x \le L).$$

Hence h_n is bounded and in particular bounded below above zero. Now taking m and M as in (4.30) and observing that $D(x, h_n(x)) \equiv 0$, we have that

$$D(x, m) \leq D(x, h_n(x)) = 0$$

and

$$D(x, M) \geq D(x, h_n(x)) = 0,$$

for $0 \leq x \leq L$, showing that (4.20) holds. We have now shown that the functions f and b given by (4.30) satisfy all the conditions of Theorem 2. Theorem 3 follows by simply writing Theorem 2 in terms of the functions F, D and using the transformations

$$h_\epsilon(x) \equiv u_\epsilon(xL), \qquad H(x) \equiv U(xL).$$

Interpretation of Theorem 3 The above theorem is only an intermediate step, however it is extremely important to this thesis, because it defines the conditions under which we can make progress. The conditions of the theorem will be assumed to hold in what follows.

Consider a function $h \in NBV_+[0, L]$ which satisfies

$$[F(h(x))]_{x_1}^{x_2} = \int_{x_1}^{x_2} D(x, h(x))dx, \qquad \text{for all } x_1, x_2 \in [0, L]. \tag{4.31}$$

First we observe that the integral is mathematically sensible since $D(x, h(x))$ is in $L_1[0, L]$ and is bounded. For fixed x_1 the right hand side is continuous in x_2 (see [65] p.319), so it follows that $F(h(x))$ must be continuous on $[0, L]$. Thus if h is discontinuous at $x \in (0, L)$ then the jump condition

$$F(h(x-)) = F(h(x+)) \tag{4.32}$$

must hold. Following [65](p.319) it can be shown that for $x \in [0, L)$

$$\lim_{s \downarrow 0} \left(\frac{F(h(x+s)) - F(h(x))}{s} \right) = D(x, h(x+)),$$

and that for $x \in (0, L]$

$$\lim_{s \uparrow 0} \left(\frac{F(h(x+s)) - F(h(x))}{s} \right) = D(x, h(x-)).$$

Thus for an interval where h is continuous, at interior points the reduced differential equation must hold since both of the one-sided derivatives equal $D(x, h(x))$. At any end points the corresponding one-sided differential equation clearly holds. We can now give the precise mathematical definitions of what we mean by a solution of the steady flow problem for a prismatic channel.

Definition 4.1 (Type-I Solution) *A function $h \in NBV_+[0, L]$ is a type-I solution of the steady flow problem if* (4.31) *holds and at any discontinuity $x \in (0, L)$*

$$E(h(x+)) \leq E(h(x-)), \tag{4.33}$$

where E is given by (2.14).

Definition 4.2 (Type-II Solution) *A function $h \in NBV_+[0, L]$ is a type-II solution of the steady flow problem if* (4.31) *holds and at any discontinuity $x \in (0, L)$*

$$\frac{F(k) - F(h(x-))}{k - h(x-)} \leq 0, \quad \textit{for all } k \textit{ between } h(x-) \textit{ and } h(x+). \tag{4.34}$$

The definition of type-I solutions corresponds to the definition of physical solutions of the steady flow problem as introduced in section 2.2. The definition of the type-II solutions is stronger and arises naturally from our theory. Condition (4.34) is simply Oleinik's condition for a steady shock for the problem (4.6). We observed in section 4.2 that any type-II solution is also type-I solution, since $F(h(x-)) = F(h(x+))$ along with (4.34) implies (4.33) (by (2.27)). The converse of this is not necessarily true, i.e. a type-I solution is not necessarily a type-II solution. In the next section we introduce further assumptions in order that these two definitions are equivalent.

We can now give the following theorem.

Theorem 4 *For γ_0, $\gamma_1 > 0$ and under the conditions of Theorem 3, the function $h = H$ is the only type-II solution which satisfies*

$$\left.\begin{array}{ll} \textit{(1)} & \textit{For all } k \textit{ between } \gamma_0 \textit{ and } h(0) \\ & \\ & sgn(h(0) - \gamma_0)(F(h(0)) - F(k)) \geq 0. \end{array}\right\} \tag{4.35}$$

$$\left.\begin{array}{ll} \textit{(2)} & \textit{For all } k \textit{ between } \gamma_1 \textit{ and } h(L) \\ & \\ & sgn(h(L) - \gamma_1)(F(h(L)) - F(k)) \leq 0. \end{array}\right\} \tag{4.36}$$

Proof of Theorem 4 This theorem is proved in two parts. We start by demonstrating that the function $h = H$ is a type-II solution.

Firstly $H \in NBV_+[0, L]$. Using property 2 of Theorem 3 we have that $\epsilon h'_\epsilon \to 0$ in L_1 as $\epsilon \downarrow 0$. Hence there is a positive sequence $S = \{\epsilon_n\}$ with $\epsilon_n \to 0$ as $n \to \infty$ and

$$\epsilon h'_\epsilon \to 0 \text{ a.e. as } \epsilon \downarrow 0, \epsilon \in S.$$

Since $h_\epsilon \to H$ in L_1 as $\epsilon \downarrow 0 (\epsilon \in S)$, there is a subsequence S' of S such that

$$h_\epsilon \to H \text{ a.e. as } \epsilon \downarrow 0, \epsilon \in S'.$$

We define the set

$$X = \{x \in [0, L] : \epsilon h'_\epsilon(x) \to 0 \text{ and } h_\epsilon(x) \to H(x) \text{ as } \epsilon \downarrow 0, \epsilon \in S'\},$$

where $[0, L] \backslash X$ has zero measure. Suppose x'_1,$x'_2 \in X$, and integrate the differential equation (4.8) from x'_1 to x'_2 to obtain

$$[\epsilon h'_\epsilon(x)]_{x'_1}^{x'_2} + [F(h_\epsilon(x))]_{x'_1}^{x'_2} = \int_{x'_1}^{x'_2} D(x, h_\epsilon(x))dx.$$

We have $D(x, h_\epsilon(x)) \to D(x, H(x))$ a.e. as $\epsilon \downarrow 0, \epsilon \in S'$ and the sequence $D(x, h_\epsilon(x))$ is uniformly bounded in x, therefore letting $\epsilon \downarrow 0, \epsilon \in S'$ and using the Lebesgue Dominated Convergence Theorem (see ref. [65]) and the fact that F is continuous in h gives

$$[F(H(x))]_{x'_1}^{x'_2} = \int_{x'_1}^{x'_2} D(x, H(x))dx.$$

This can be shown to hold for arbitrary x_1, x_2 in $[0, L]$ by firstly letting $x'_1 \to x_1$ with $x'_1 \in X$ such that $H(x'_1) \to H(x_1)$, and secondly letting $x'_2 \to x_2$ with $x'_2 \in X$ such that $H(x'_2) \to H(x_2)$. F depends continuously on h and also $D(x, H(x)) \in L_1$ so that the integral depends continuously on its limits (see [65] p.319), hence we have

$$[F(H(x))]_{x_1}^{x_2} = \int_{x_1}^{x_2} D(x, H(x))dx.$$

Finally condition (4.34) holds at all discontinuities because of (4.28)

The second part of the proof is to show that $h = H$ is the only type-II solution which satisfies (4.35) and (4.36). Suppose that h is a type-II solution which satisfies (4.35) and (4.36), then we show that the conditions of part 4 of Theorem 3 hold

and therefore we must have $h = H$. We have already shown that (4.31) implies (4.27) and also the equalities in (4.28). The inequalities in (4.28) follow from (4.34). Condition (4.29) is clearly equivalent to (4.35) and (4.36). This completes the proof.

The main consequence of the above theorem is that there is at most one type-II solution satisfying any set of boundary values. To see this, suppose that h^1 and h^2 are type-II solutions with $h^1(0) = h^2(0)$ and $h^1(L) = h^2(L)$. Now for $\gamma_0 = h^1(0)$ and $\gamma_1 = h^1(L)$, both these solutions satisfy (4.35) and (4.36), so we must have $h^1 \equiv h^2 \equiv H$. The theorem also gives a weak existence result in that for any positive γ_0 and γ_1 there exists a type-II solution satisfying (4.35) and (4.36). In the next section this result allows us to identify all the possible physical solutions.

The difficulty with the above theory is that it concerns only type-II solutions. We would like to obtain theory regarding type-I solutions, since it is this definition which corresponds to our original concept of what is a physically allowable solution. The definitions of type-I/II solutions differ only in the cases where the channel cross-section admits multiple critical depths. It may be that in such cases that the second definition is the appropriate definition of a physical solution. Such an investigation is beyond the scope of this thesis and we avoid this difficulty by only considering the situation where the cross-section has a single critical depth, so that the two definitions are equivalent.

4.9 Cross-Sections with a Single Critical Depth

In this section we simplify the previous theory by considering only channel cross-sections with a single critical depth. The definitions of type-I and type-II solutions are then equivalent and such solutions will be referred to as physical solutions. We have already observed that any type-II solution must be a type-I solution because of the relationship (2.27). We now show that for a channel with a single critical depth that the converse is also true. Any depth h which satisfies

$$F'(h) = gA\left(1 - \frac{Q^2T}{gA^3}\right) = gA(1 - F_r^2) = 0,$$

corresponds to a critical depth, so that all local extrema of F correspond to critical depths. Suppose that h is a type-I solution, then across any discontinuity we have

$F(h(x+)) = F(h(x-))$ and $E(h(x+)) \leq E(h(x-))$. We start by showing that $F(k) \leq F(h(x-))$ for all k in the set

$$S = \{k : \min\{h(x+), h(x-)\} < k < \max\{h(x+), h(x-)\}\},$$

and furthermore $F < F(h(x-))$ for some depth in S. First suppose that $F > F(h(x-))$ for some depth in S, then clearly F has a local maxima at some depth in S, and this corresponds to a critical depth. But since $F \to \infty$ as $h \downarrow 0$, F must also have a local minima which corresponds to a second critical depth, contradicting the uniqueness assumption. If $F(k) = F(h(x-))$ for all k in S then $F'(k) = 0$ for all k in S, implying an infinite number of critical depths. Thus $F(k) \leq F(h(x-))$ for all $k \in S$ with $F < F(h(x-))$ for some depth in this range.

Since $E(h(x+)) \leq E(h(x-))$ the relationship (2.27) implies that

$$\int_{h(x-)}^{h(x+)} \frac{T}{A^2} (F - F(h(x-)))\, dh \leq 0,$$

and it follows that $h(x-) < h(x+)$ and hence (4.34) holds. We conclude that h is a type-II solution.

If the channel width does not approach zero as the depth becomes large, then the situation is as in section 2.2.1 and the F has the properties discussed there. These properties can easily be used to show that (4.37) is equivalent to (4.35) and (4.36), and Theorem 4 then reduces to the following.

Theorem 5 *Suppose that γ_0, $\gamma_1 > 0$ and that in addition to the conditions of Theorem 3 that the following hold:*

(1) $T \geq T_0 > 0$ *as* $h \to \infty$, *for some constant* T_0.

(2) There is only one positive depth h_c which satisfies

$$\frac{A(h_c)^3}{T(h_c)} = \frac{Q^2}{g}.$$

Under these conditions the function $h = H$ is the only physical solution which sat-

isfies:

$$\left.\begin{array}{ll} \textit{(1)} & \textit{If } \gamma_0 < h_c \textit{ then } h(0) = \gamma_0 \textit{ or } h(0) \geq \gamma_0^*. \\ \textit{(2)} & \textit{If } \gamma_0 \geq h_c \textit{ then } h(0) \geq h_c. \\ \textit{(3)} & \textit{If } \gamma_1 \leq h_c \textit{ then } h(L) \leq h_c. \\ \textit{(4)} & \textit{If } \gamma_1 > h_c \textit{ then } h(L) = \gamma_1 \textit{ or } h(L) \leq \gamma_1^*. \end{array}\right\} \tag{4.37}$$

The above theorem leads to the following results.

(1) There is exactly one physical solution h with $h(0) \geq h_c$ and $h(L) \leq h_c$. We define $\alpha_0 = h(0)$ and $\alpha_1 = h(L)$.

(2) For each $\gamma_0 < \alpha_0^\star$ there is exactly one physical solution h with $h(0) = \gamma_0$ and $h(L) \leq h_c$. We define $\beta_1(\gamma_0) = h(L)$.

(3) For each $\gamma_1 > \alpha_1^\star$ there is exactly one physical solution h with $h(0) \geq h_c$ and $h(L) = \gamma_1$. We define $\beta_0(\gamma_1) = h(0)$.

Part (1) is an immediate consequence of applying Theorem 5 with $\gamma_0 = \gamma_1 = h_c$. To see part (2), apply Theorem 5 for $\gamma_0 < \alpha_0^\star$ and $\gamma_1 = h_c$ to give that there is exactly one physical solution satisfying:

$$h(0) = \gamma_0 \qquad h(L) \leq h_c,$$

or

$$h(0) \geq \gamma_0^\star \qquad h(L) \leq h_c.$$

But since $\gamma_0^\star > \alpha_0$, if the second of these possibilities were satisfied it would violate part (1). Hence the first must be satisfied. The same type of argument can be used to show part (3). In fact this type of argument can be used to determine the values of $H(0)$ and $H(L)$ for any particular γ_0 and γ_1. The results are given in table (4.1). This table would be of great use if only we could determine the values of α_0, α_1 and the functions $\beta_0(\gamma_1)$ and $\beta_1(\gamma_0)$, but this is not possible analytically, and can only be done numerically. Since any physical solution must be the vanishing viscosity solution for some choice of γ_0 and γ_1, table (4.1) gives all the possible physical solutions. Any physical solution is given by one of the following four types.

Region	Subregion	$H(0)$	$H(L)$
$\gamma_0 \geq h_c, \quad \gamma_1 \leq h_c$		α_0	α_1
$\alpha_0^\star \leq \gamma_0 < h_c, \quad \gamma_1 \leq h_c$		α_0	α_1
$\gamma_0 < \alpha_0^\star, \quad \gamma_1 \leq h_c$		γ_0	$\beta_1(\gamma_0)$
$\gamma_0 \geq h_c, \quad h_c < \gamma_1 \leq \alpha_1^\star$		α_0	α_1
$\gamma_0 \geq h_c, \quad \gamma_1 > \alpha_1^\star$		$\beta_0(\gamma_1)$	γ_1
$\alpha_0^\star \leq \gamma_0 < h_c, \quad h_c < \gamma_1 \leq \alpha_1^\star$		α_0	α_1
$\gamma_0 < \alpha_0^\star, \quad h_c < \gamma_1 \leq \alpha_1^\star$	$\gamma_1 \leq \beta_1(\gamma_0)^\star$	γ_0	$\beta_1(\gamma_0)$
	$\gamma_1 > \beta_1(\gamma_0)^\star$	γ_0	γ_1
$\alpha_0^\star \leq \gamma_0 < h_c, \quad \gamma_1 > \alpha_1^\star$	$\gamma_0 \geq \beta_0(\gamma_1)^\star$	$\beta_0(\gamma_1)$	γ_1
	$\gamma_0 < \beta_0(\gamma_1)^\star$	γ_0	γ_1
$\gamma_0 < \alpha_0^\star, \quad \gamma_1 > \alpha_1^\star$	$\gamma_0 < \beta_0(\gamma_1)^\star, \quad \gamma_1 \leq \beta_1(\gamma_0)^\star$	γ_0	$\beta_1(\gamma_0)$
	$\gamma_0 \geq \beta_0(\gamma_1)^\star, \quad \gamma_1 > \beta_1(\gamma_0)^\star$	$\beta_0(\gamma_1)$	γ_1
	$\gamma_0 < \beta_0(\gamma_1)^\star, \quad \gamma_1 > \beta_1(\gamma_0)^\star$	γ_0	γ_1
	$\gamma_0 \geq \beta_0(\gamma_1)^\star, \quad \gamma_1 \leq \beta_1(\gamma_0)^\star$	Region does not exist	

Figure 4.1: Properties of limit solution for different γ_0 and γ_1

Flow subcritical at inflow and supercritical at outflow The only solution of this type satisfies $h(0) = \alpha_0$ and $h(L) = \alpha_1$. This solution can be obtained as the vanishing viscosity solution by setting $\gamma_0 \geq h_c$ and $0 < \gamma_1 \leq h_c$.

Flow supercritical at inflow and supercritical at outflow For $0 < \gamma_0 < \alpha_0^\star \leq h_c$ there is a solution with $h(0) = \gamma_0$ and $h(L) \leq h_c$. This solution can be obtained as the vanishing viscosity solution by setting $0 < \gamma_1 \leq h_c$.

Flow subcritical at inflow and subcritical at outflow For $\gamma_1 > \alpha_1^\star \geq h_c$ there is a solution with $h(0) \geq h_c$ and $h(L) = \gamma_1$. This solution can be obtained as the vanishing viscosity solution by setting $\gamma_0 \geq h_c$.

Flow supercritical at inflow and subcritical at outflow If γ_0 and γ_1 satisfy one of the conditions below, then there is a solution which satisfies $h(0) = \gamma_0$ and $h(L) = \gamma_1$.

(1) $0 < \gamma_0 \leq \alpha_0^\star \leq h_c$ and $h_c \leq \beta_1(\gamma_0)^\star < \gamma_1 \leq \alpha_1^\star$

(2) $h_c \leq \alpha_0^\star \leq \gamma_0 < \beta_0(\gamma_1)^\star$ and $\gamma_1 > \alpha_1^\star \geq h_c$

(3) $0 < \gamma_0 < \min\{\alpha_0^\star, \beta_0(\gamma_1)^\star\} \leq h_c$ and $\gamma_1 > \max\{\alpha_1^\star, \beta_1(\gamma_0)^\star\} \geq h_c$.

From the above we observe that in order to specify the depth at inflow with any degree of freedom the depth specified must at the very minimum correspond to supercritical flow (and even then there will only be solutions for certain ranges of depth). Similarly to specify the depth at outflow with any degree of freedom requires this depth to correspond to subcritical flow. This observation agrees with the theory of characteristics discussed in section 2.2.4.

We end this section by demonstrating that practical problems exist which do satisfy the conditions required by the theory. The major restrictions placed by theory are as follows:

(1) The channel must be prismatic

(2) The bed slope must be positive.

(3) The conveyance must satisfy (4.24).

(4) There must be only one critical depth.

The condition that the bed slope is positive appears to be the most restrictive. However, as we demonstrate later, when this condition is violated the uniqueness conclusions of the theory may not hold. This condition on the conveyance is only a slightly stronger version of the condition (2.31) which is used in section 2.2.2. If we again take the form (2.17) for the conveyance and and now require that

$$k_1 \geq 1/2 \text{ and } 0 \leq k_2 \leq k_1 - 1/2, \tag{4.38}$$

which includes both the Manning and Chezy forms, then conditions (4.24) are satisfied for rectangular, trapezoidal and triangular channels. This can be seen by

using (2.33) with k_1 replaced with $k_1 - 1/2$. Such cross-sections also have a unique critical depth (see section 2.2.2). There is no obvious way of showing that these conditions hold for a wider class of cross-sections and friction laws, other than testing each individual case.

4.10 Extension of the Theory

The theory derived in this chapter has certain limitations on the situations it can be applied to. In this section we discuss whether these limitations may be overcome.

Theorem 3 requires that the bed slope is positive and that (4.24) holds, in order that $D_h > 0$ for all $h > 0$ and all $0 \leq x \leq L$. If this condition is violated, then are the conclusions of the theorem still true? We demonstrate that in general they are not.

Consider a "well-behaved" channel in the sense of section 2.2.1. The channel, which need not be prismatic, has a single critical depth $h_c(x)$ at each cross-section, and a jump is allowable at x if and only if

$$h(x-) < h_c(x) < h(x+) = h(x-)^*.$$

Suppose that $\gamma_0 < h_c(0)$, $\gamma_1 > h_c(L)$ and that the following two problems have solutions:

$$\left.\begin{aligned} &F(x, h^1(x))' = D(x, h^1(x)), \quad h^1(x) < h_c(x), \quad 0 \leq x \leq L, \\ &h^1(0) = \gamma_0, \end{aligned}\right\} \tag{4.39}$$

$$\left.\begin{aligned} &F(x, h^2(x))' = D(x, h^2(x)), \quad h^2(x) > h_c(x), \quad 0 \leq x \leq L, \\ &h^2(L) = \gamma_1. \end{aligned}\right\} \tag{4.40}$$

If we define

$$J(x) = F(x, h^2(x)) - F(x, h^1(x)),$$

then any value $x^\star \in (0, L)$ such that $J(x^\star) = 0$ corresponds to a physically allowable jump, giving a physical solution

$$h(x) = \begin{cases} h^1(x) & 0 \leq x < x^\star, \\ h^2(x) & x^\star \leq x \leq L, \end{cases}$$

which satisfies both $h(0) = \gamma_0$ and $h(L) = \gamma_1$. Thus if J has more than a single root, then there is more than one physical solution satisfying the same boundary values. Observe that

$$\begin{aligned} J'(x) &= F(x, h^2(x))' - F(x, h^1(x))' &&= D(x, h^2(x)) - D(x, h^1(x)) \\ &&&= D_h(x, \hat{h}(x))(h^2(x)) - h^1(x)), \end{aligned}$$

with $h^1(x) < \hat{h}(x) < h^2(x)$, from applying the Mean Value Theorem. If $D_h > 0$ holds, then clearly $J'(x) > 0$ for $0 \leq x \leq L$, and J has at most one root. We now demonstrate a case where J has more than one root.

The most common situation where the condition $D_h > 0$ is violated is where the bed slope becomes zero or negative. Consider the following problem of a 100m long, 10m wide rectangular channel, carrying a discharge of $20\text{m}^3/\text{s}$, with bed slope given by

$$S_0(x) \equiv -\frac{1}{50}\left(\frac{x}{100} - \frac{1}{2}\right),$$

and boundary conditions $\gamma_0 = 0.45\text{m}$, $\gamma_1 = 0.9\text{m}$. We use the Manning form of the conveyance K with $n = 0.005$. It can be seen that $S_0 < 0$ for $x > 50\text{m}$. Problems (4.39) and (4.40) were numerically integrated using an automatic step size Runge-Kutta-Fehlberg method with a fourth and fifth order pair. Figure 4.2 shows $F(x, h^1(x))$ and $F(x, h^2(x))$. There are clearly two intersection points, so J has two roots. The second part of the figure shows $h^1(x)$ and $h^2(x)$ and illustrates the allowable jumps.

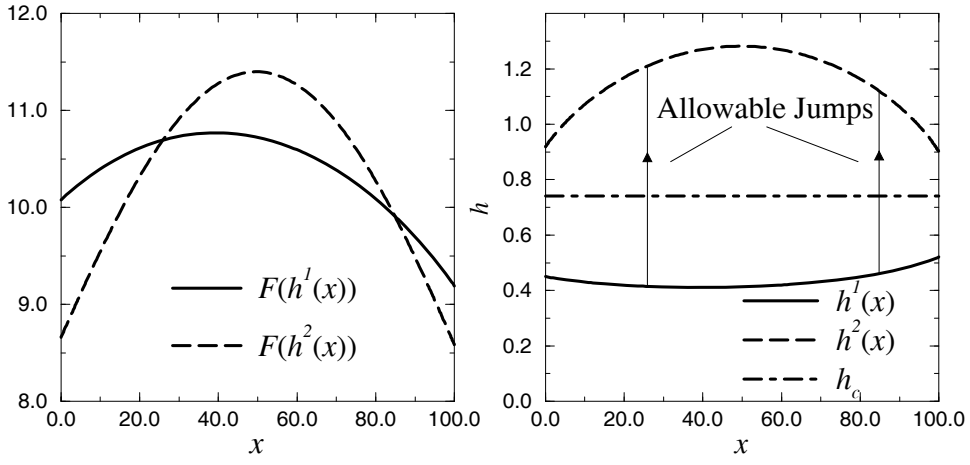


Figure 4.2: Construction of Multiple Solutions by Numerical Integration

We conclude that when $D_h > 0$ is violated, the conclusions of the theory may no longer be true, i.e. there may be more than one physical weak solutions satisfying identical boundary values. In a case where this happens, which of the solutions is the solution we require? It may only be possible to answer this question by examining the transient behaviour of the flow. For example one steady state may be arrived at from one initial state while another steady state may be arrived at from a different initial state. If this is the case then it puts an inherent limitation on the approach of computing the steady state flow without regard to the transient behaviour. It may also be true that only one of the solutions is stable in time (see section 2.2.5). In that case it may be possible to develop theory to determine the necessary conditions for stability and use these to discriminate against unstable solutions.

Theorem 3 is only applicable to prismatic channels. There is theory similar to Theorem 1 (see [36]) which allows f to depend on x as well as u, and thus could possibly be adapted to apply to the steady flow problem for nonprismatic channels. The condition $b_u \geq \delta > 0$ is replaced by the condition

$$b_u - |f_{xu}| \geq \delta > 0.$$

If the theory were adapted to the steady flow problem, we would now expect to require the condition

$$D_h - |F_{xh}| > 0.$$

However it is not clear whether this condition holds for a useful class of problems.

Chapter 5

A Class of Numerical Methods

In the previous chapter we demonstrated that under certain conditions there is at most one physical solution to the steady Saint-Venant problem, for any given boundary values, and that this solution is the vanishing viscosity solution of a second order two-point boundary value problem. In this chapter we follow on from these ideas to consider a family of finite difference approximations to the steady flow problem. As before we consider only prismatic channels, although the schemes will be extended to non-prismatic channels in Chapter 9. The basis of the theory, as in the previous chapter, is the work by Lorenz[32] although other authors, notably Abrahamsson and Osher[1] and Osher[47], have made significant contributions. Other closely related work by Lorenz can be found in [35], [34] and [33].

The steady flow equation (2.21) for a prismatic channel can be written as

$$\frac{d}{dx}f(h) = -D(x,h), \tag{5.1}$$

where $f(h) \equiv -F(h)$ and the functions F and D are given by (2.6) and (2.7) respectively. We consider approximations to this equation of the form

$$\frac{g(h_{j+1},h_j) - g(h_j,h_{j-1})}{\Delta x} = -D(x_j,h_j), \tag{5.2}$$

where $x_j = j\Delta x$, $h_j \approx h(x_j)$ and Δx is the uniform grid spacing. We require that

$$g(h,h) = f(h) \tag{5.3}$$

for all positive h, in order that the scheme be consistent with the differential equation. A motivation for considering such a scheme comes from the previous chapter

where we observed that under certain conditions the physical solutions of the steady flow problem are exactly the steady state entropy satisfying solutions of the scalar conservation law (4.6). Applying a three-point conservative finite difference scheme to this scalar conservation law and using a pointwise discretisation of the source term yields the scheme

$$\frac{h_j^{n+1} - h_j^n}{\Delta t} + \frac{g(h_{j+1}^n, h_j^n) - g(h_j^n, h_{j-1}^n)}{\Delta x} = -D(x_j, h_j^n), \tag{5.4}$$

where $h_j^n \approx h(j\Delta x, n\Delta t)$ and again (5.3) is required for consistency. At steady state this reduces to (5.2). In theory, almost any of the vast amount of numerical methods for scalar conservation laws (some of which are described in Chapter 3) could be applied to (4.6) and so be used to compute solutions of the steady flow problem, and in Chapter 7 we apply some specimen schemes from the literature. From the viewpoint of theory we consider only simple schemes of the above form. We furthermore restrict attention to the forms of the numerical flux function g which give rise to monotone schemes (in the sense of homogeneous scalar conservation laws). This includes the Engquist-Osher, Godunov and Lax-Friedrichs forms, but excludes the first-order upwind form.

For homogeneous scalar conservation laws the theory for monotone schemes is particularly rich. For example it is known that a convergent sequence of solutions (as Δx, $\Delta t \downarrow 0$ with $\Delta t/\Delta x$ fixed) must always converge to an entropy satisfying solution of the conservation law. This property often cannot be demonstrated for other types of schemes, and in many cases, such as for the first-order upwind scheme, it can be shown to be violated (see [30]). We show that the above property carries over into the steady flow problem by guaranteeing convergence to the physically allowable solution. The fact that (i) we have the addition of a source term, (ii) we are only interested in stationary solutions and (iii) we only consider a finite domain, means that the basis of the theory, as in Chapter 4, comes naturally from the field of singular perturbation problems, rather than the field of conservation laws.

5.1 Theory for Monotone Schemes

At the beginning of the 1980's, authors from the field of conservation laws, primarily Osher[47], recognised that the conservative approximation to the spatial derivative could be used in approximating singular perturbation problems of the form (4.12), yielding the scheme

$$\epsilon\frac{u_{j+1}-2u_j+u_{j-1}}{\Delta x^2}-\frac{g(u_{j+1},u_j)-g(u_j,u_{j-1})}{\Delta x}-b(x_j,u_j)=0, \qquad j=1,\ldots,N-1$$
$$u_0=\gamma_0, \quad u_N=\gamma_1, \tag{5.5}$$

where $\Delta x = 1/N$, $x_j = j\Delta x$ and u_j approximates $u_\epsilon(x_j)$. Lorenz[32] considers the above scheme under the conditions of Theorem 1 and for numerical flux functions g which give rise to monotone schemes for scalar conservation laws. The existence and uniqueness of the solution of the system of difference equations is shown for all $\epsilon \geq 0$ and $\Delta x > 0$. This arises from the fact that the system of equations form an M-function. The theory of M-functions is described in [46]. An alternative proof uses the fact that a particular mapping is a contraction mapping and follows along the lines of [47]. This second approach is more constructive since it yields a practical algorithm for computing solutions of the system of difference equations. It is also shown in [32] that the discrete solution converges in L_1 (as Δx and ϵ vanish) to the vanishing viscosity solution of problem (4.12). The family of piecewise constant extensions to the discrete solutions, $\{U_\epsilon^{\Delta x}\}_{\epsilon\geq 0}^{\Delta x>0}$ is shown to be precompact in L_1, thus any sequence of solutions has a convergent subsequence. The Lebesgue dominated convergence theorem is used to show that the limit defined by a convergent subsequence, as Δx and ϵ vanish, satisfies the properties which characterise the vanishing viscosity solution of (4.12). All that is required to demonstrate the precompactness is that the discrete solution and its total variation are shown to be uniformly bounded in Δx and ϵ.

It turns out that the order in which ϵ and Δx vanish is unimportant. For this

reason we may perform the limit as ϵ vanishes immediately to obtain the scheme

$$\begin{aligned} &\frac{g(u_{j+1},u_j)-g(u_j,u_{j-1})}{\Delta x}+b(x_j,u_j)=0, \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad j=1,\ldots,N-1 \\ &u_0=\gamma_0, \quad u_N=\gamma_1, \end{aligned} \tag{5.6}$$

and we are now only concerned with the limit as Δx vanishes. This scheme is clearly of the same form as (5.2) for the appropriate functions f and b and a transformation onto the interval $[0,L]$. The theory of Lorenz described above is summarised by the following theorem. This is essentially the discrete analogue of Theorem 1 with the role of ϵ replaced by Δx. The proof can be found in [32].

Theorem 6 ([32]) *Suppose the situation is as in Theorem 1 and consider the difference equations*

$$\begin{aligned} \mathcal{T}_j u=0, \qquad & j=1,2,\ldots,N-1 \\ u_0=\gamma_0, \quad & u_N=\gamma_1, \end{aligned} \tag{5.7}$$

where

$$\mathcal{T}_j u=\frac{g(u_{j+1},u_j)-g(u_j,u_{j-1})}{\Delta x}+b(x_j,u_j),$$

$\Delta x=1/N$, $x_j=j\Delta x$ *and the function* g *has the following properties:*

(1) $g(u,u)=f(u)$ *for all* $u\in\mathbb{R}$.

(2) For each $v\in\mathbb{R}$, $u\mapsto g(u,v)$ *is decreasing for all* $u\in\mathbb{R}$.

(3) For each $u\in\mathbb{R}$, $v\mapsto g(u,v)$ *is increasing for all* $v\in\mathbb{R}$.

(4) There exists a constant l *such that for all* $u_1,u_2,v_1,v_2\in\mathbb{R}$

$$|g(u_2,v_2)-g(u_1,v_1)|\le l(|u_2-u_1|+|v_2-v_1|).$$

Under these conditions the difference equations have exactly one solution $\boldsymbol{u}^{\Delta x}=(u_0^{\Delta x},u_1^{\Delta x},\ldots,u_N^{\Delta x})^T$ *for each* $N\in\mathbb{N}$. *If* $U^{\Delta x}\in L_1[0,1]$ *denotes the piecewise constant extension of the discrete solution given by*

$$U^{\Delta x}=u_j^{\Delta x} \quad \text{for} \quad j\Delta x\le x<(j+1)\Delta x, \quad j=0,1,\ldots,N \tag{5.8}$$

then $U^{\Delta x}\to U$ *in* L_1 *as* $\Delta x\to 0$, *where* $U\in NBV[0,1]$ *is the limiting solution of problem* (4.12) *as* $\epsilon\downarrow 0$.

The first condition placed on the numerical flux function g is simply the consistency condition. The second and third conditions on g ensure that that the scheme is monotone. To be precise, when we say that the scheme is monotone, we in fact mean that the time dependent scheme (5.19) is monotone (see section 3.4) in conjunction with an appropriate CFL condition. These conditions require g to be non-increasing in its first argument and non-decreasing in its second argument. The final condition requires g to be Lipschitz continuous, which is stronger than requiring continuity, but weaker than requiring differentiability.

As in Chapter 4 we proceed to modify this theory to hold under the conditions of Theorem 2, where the solution is restricted to being positive. To do this we construct a new problem which has identical solutions to the set of difference equations (at least in some finite region of solution space) and to which Theorem 6 is applicable. Essentially we are extending the functions g, f and b to have "well-behaved" values outside the region of interest.

Before we present the modified theorem we must introduce some notation. For $\boldsymbol{u}$, $\boldsymbol{v} \in I\!R^{N+1}$ we write $\boldsymbol{u} \leq \boldsymbol{v}$ if and only if $u_i \leq v_i$ for each i. The analogous definition applies for each of the operators $=$, $\geq$,$<$ and $>$. We define the set

$$[\boldsymbol{u}, \boldsymbol{v}] = \left\{ \boldsymbol{w} \in I\!R^{N+1} : \boldsymbol{u} \leq \boldsymbol{w} \leq \boldsymbol{v} \right\} \subset I\!R^{N+1}.$$

For the scalar quantities α and β we define the vectors

$$\begin{aligned} \boldsymbol{\alpha} &= (\alpha, \alpha, \ldots, \alpha) \in I\!R^{N+1}, \\ \boldsymbol{\beta} &= (\beta, \beta, \ldots, \beta) \in I\!R^{N+1}. \end{aligned}$$

Likewise for the vectors $\underline{\boldsymbol{u}}$ and $\overline{\boldsymbol{u}}$ constructed from the scalar quantities $\underline{u}$ and $\overline{u}$.

Theorem 7 *Suppose the situation is as in Theorem 2 and consider the difference equations* (5.7) *where now* $u_0, u_1, \ldots, u_N$ *are restricted to be positive. Suppose that for some* $0 < \alpha \leq \underline{u}$, $\beta \geq \overline{u}$ *the function* g *has the properties:*

(1) $g(u,u) = f(u)$ *for all* $u \in [\alpha, \beta]$.

(2) u_1, u_2, $v \in [\alpha, \beta]$ *with* $u_2 \geq u_1$ *implies* $g(u_2, v) \leq g(u_1, v)$.

(3) v_1, v_2, $u \in [\alpha, \beta]$ *with* $v_2 \geq v_1$ *implies* $g(u, v_2) \geq g(u, v_1)$.

(4) There exists a constant l such that for all u_1, u_2, v_1, $v_2 \in [\alpha, \beta]$

$$|g(u_2, v_2) - g(u_1, v_1)| \leq l(|u_2 - u_1| + |v_2 - v_1|).$$

Under these conditions the difference equations have a unique solution

$$\boldsymbol{u}^{\Delta x} = (u_0^{\Delta x}, u_1^{\Delta x}, \ldots, u_N^{\Delta x})^T$$

in $[\boldsymbol{\alpha}, \boldsymbol{\beta}]$ for each $N \in \mathbb{N}$. This solution satisfies the bounds:

$$0 < \underline{u} \leq u_j^{\Delta x} \leq \overline{u}, \qquad j = 0, 1, \ldots, N. \tag{5.9}$$

If $U^{\Delta x} \in L_1[0,1]$ denotes the piecewise constant extension of this discrete solution given by (5.8), then $U^{\Delta x} \to U$ in L_1 as $\Delta x \to 0$, where $U \in NBV_+[0,1]$ is the limiting solution of problem P_ϵ (4.18) as $\epsilon \downarrow 0$.

Proof of Theorem 7 The proof of Theorem 2 shows that the problem $P_\epsilon^{\alpha,\beta}$ satisfies Theorem 1 and that problems P_ϵ and $P_\epsilon^{\alpha,\beta}$ have identical solutions and hence identical limiting solutions. We consider the following set of difference equations

$$\begin{gathered} \mathcal{T}_j^{\alpha,\beta} u = 0, \qquad j = 1, 2, \ldots, N-1 \\ u_0 = \gamma_0, \quad u_N = \gamma_1, \end{gathered} \tag{5.10}$$

where

$$\mathcal{T}_j^{\alpha,\beta} u = \frac{g^{\alpha,\beta}(u_{j+1}, u_j) - g^{\alpha,\beta}(u_j, u_{j-1})}{\Delta x} + b^{\alpha,\beta}(x_j, u_j).$$

The function $g^{\alpha,\beta}$ is constructed so as to match up with g on the region $\alpha \leq u \leq \beta$, $\alpha \leq v \leq \beta$ and also to satisfy the conditions of Theorem 6. Define the function $g^{\alpha,\beta}$ by

$$\begin{aligned} g^{\alpha,\beta}(u,v) &= g(\mu_1(u), \mu_1(v)) + q_1(\mu_2(u); \alpha) + q_1(\mu_3(u); \beta) \\ &\quad + q_2(\mu_2(v); \alpha) + q_2(\mu_3(v); \beta), \end{aligned}$$

where

$$\begin{aligned} q_1(u; z) &= \int_z^u \min\{(f^{\alpha,\beta})'(s), 0\} ds, \\ q_2(v; z) &= \int_z^v \max\{(f^{\alpha,\beta})'(s), 0\} ds, \end{aligned}$$

$$\mu_1(u) = \begin{cases} \beta & u > \beta \\ u & \alpha \leq u \leq \beta \\ \alpha & u < \alpha, \end{cases}$$

$$\mu_2(u) = \begin{cases} \alpha & u > \alpha \\ u & u \leq \alpha, \end{cases}$$

and

$$\mu_3(u) = \begin{cases} \beta & u < \beta \\ u & u \geq \beta. \end{cases}$$

We start by showing that conditions (1)-(4) of Theorem 6 hold for $g^{\alpha,\beta}$. Firstly we have that

$$\begin{aligned} g^{\alpha,\beta}(u,u) &= g(\mu_1(u),\mu_1(u)) + q_1(\mu_2(u);\alpha) + q_1(\mu_3(u);\beta) \\ &\quad + q_2(\mu_2(u);\alpha) + q_2(\mu_3(u);\beta) \\ &= f^{\alpha,\beta}(\mu_1(u)) + \int_{\alpha}^{\mu_2(u)} (f^{\alpha,\beta})'(s)ds + \int_{\beta}^{\mu_3(u)} (f^{\alpha,\beta})'(s)ds \\ &= f^{\alpha,\beta}(\mu_1(u)) + f^{\alpha,\beta}(\mu_2(u)) - f^{\alpha,\beta}(\alpha) + f^{\alpha,\beta}(\mu_3(u)) - f^{\alpha,\beta}(\beta) \\ &= f^{\alpha,\beta}(u), \end{aligned}$$

so condition (1) holds. Secondly for $u_2 \geq u_1$ we have

$$\begin{aligned} g^{\alpha,\beta}(u_2,v) - g^{\alpha,\beta}(u_1,v) &= g(\mu_1(u_2),\mu_1(v)) - g(\mu_1(u_1),\mu_1(v)) \\ &\quad + \int_{\mu_2(u_1)}^{\mu_2(u_2)} \min\{(f^{\alpha,\beta})'(s),0\}ds \\ &\quad + \int_{\mu_3(u_1)}^{\mu_3(u_2)} \min\{(f^{\alpha,\beta})'(s),0\}ds \\ &\leq 0, \end{aligned}$$

since $\mu_j(u_2) \geq \mu_j(u_1)$ for $j = 1,2,3$. Similarly for $v_2 \geq v_1$ we have

$$\begin{aligned} g^{\alpha,\beta}(u,v_2) - g^{\alpha,\beta}(u,v_1) &= g(\mu_1(u),\mu_1(v_2)) - g(\mu_1(u),\mu_1(v_1)) \\ &\quad + \int_{\mu_2(v_1)}^{\mu_2(v_2)} \max\{(f^{\alpha,\beta})'(s),0\}ds \\ &\quad + \int_{\mu_3(v_1)}^{\mu_3(v_2)} \max\{(f^{\alpha,\beta})'(s),0\}ds \\ &\geq 0, \end{aligned}$$

since $\mu_j(v_2) \geq \mu_j(v_1)$ for $j = 1,2,3$. It follows that $g^{\alpha,\beta}(u,v)$ is decreasing in u and increasing in v giving properties (2) and (3). Finally we demonstrate that $g^{\alpha,\beta}(u,v)$

is Lipschitz continuous. The first term is Lipschitz continuous since for $j = 1, 2, 3$ and for each $u_1, u_2 \in I\!R$ we have

$$|\mu_j(u_2) - \mu_j(u_1)| \leq |u_2 - u_1|,$$

so that for $u_1, u_2, v_1, v_2 \in I\!R$ we have

$$\begin{aligned} |g(\mu_1(u_2), \mu_1(v_2)) - g(\mu_1(u_1), \mu_1(v_1))| &\leq l(|\mu_1(u_2) - \mu_1(u_1)| + |\mu_1(v_2) - \mu_1(v_1)|) \\ &\leq l(|u_2 - u_1| + |v_2 - v_1|). \end{aligned}$$

Next choose a constant l' such that $|(f^{\alpha,\beta})'(u)| \leq l'$ for all u. Such a constant exists since $(f^{\alpha,\beta})'(u) \to 0$ as $u \to \pm\infty$. Now

$$\begin{aligned} |q_1(\mu_2(u_2); \beta) - q_1(\mu_2(u_1); \beta)| &= \left| \int_{\mu_2(u_1)}^{\mu_2(u_2)} \min\{(f^{\alpha,\beta})'(s), 0\} ds \right| \\ &\leq l' |\mu_2(u_2) - \mu_2(u_1)| \\ &\leq l' |u_2 - u_1|. \end{aligned}$$

The same argument shows that the remaining terms also have a Lipschitz constant l'

Applying Theorem 6 we deduce that the difference equations (5.10) have a unique solution. We show that this solution is in $[\boldsymbol{\alpha}, \boldsymbol{\beta}]$ and so must also solve equations (5.7). Furthermore this is the only solution contained in this set. Suppose that $\boldsymbol{u} = (u_0, u_1, \ldots, u_N)^T$ is the solution to (5.10) and let j_1, j_2 be such that

$$u_{j_1} \leq u_j \leq u_{j_2}, \qquad j = 0, 1, \ldots, N.$$

If $u_{j_2} > \overline{u}$ then using the difference equation for $j = j_2$ yields

$$\begin{aligned} -b^{\alpha,\beta}(x_{j_2}, u_{j_2}) &= \frac{g^{\alpha,\beta}(u_{j_2+1}, u_{j_2}) - g^{\alpha,\beta}(u_{j_2}, u_{j_2-1})}{\Delta x} \\ &= \frac{g^{\alpha,\beta}(u_{j_2+1}, u_{j_2}) - g^{\alpha,\beta}(u_{j_2}, u_{j_2}) + g^{\alpha,\beta}(u_{j_2}, u_{j_2}) - g^{\alpha,\beta}(u_{j_2}, u_{j_2-1})}{\Delta x} \\ &\geq 0, \end{aligned}$$

since $u_{j_2-1} \leq u_{j_2}$ and $u_{j_2+1} \leq u_{j_2}$. But $b^{\alpha,\beta}(x, u) > 0$ for $u > \overline{u}$ which is a contradiction, so we must have $u_{j_2} \leq \overline{u}$. Next suppose that $u_{j_1} < \underline{u}$, then using the difference

equation for $j = j_1$ yields

$$\begin{aligned} -b^{\alpha,\beta}(x_{j_1}, u_{j_1}) &= \frac{g^{\alpha,\beta}(u_{j_1+1}, u_{j_1}) - g^{\alpha,\beta}(u_{j_1}, u_{j_1-1})}{\Delta x} \\ &= \frac{g^{\alpha,\beta}(u_{j_1+1}, u_{j_1}) - g^{\alpha,\beta}(u_{j_1}, u_{j_1}) + g^{\alpha,\beta}(u_{j_1}, u_{j_1}) - g^{\alpha,\beta}(u_{j_1}, u_{j_1-1})}{\Delta x} \\ &\leq 0, \end{aligned}$$

since $u_{j_1-1} \geq u_{j_1}$ and $u_{j_1+1} \geq u_{j_1}$. But $b^{\alpha,\beta}(x,u) < 0$ for $u < \underline{u}$ which is again a contradiction, so we must have $u_{j_1} \geq \underline{u}$. Since $\boldsymbol{u} \in [\underline{\boldsymbol{u}}, \overline{\boldsymbol{u}}] \subset [\boldsymbol{\alpha}, \boldsymbol{\beta}]$ we have

$$\mathcal{T}_j u = \mathcal{T}_j^{\alpha,\beta} u = 0, \quad \text{for} \quad j = 1, \ldots, N-1.$$

Hence the solution of difference equations (5.10) is also a solution of difference equations (5.7). Clearly this can be the only solution of (5.7) in $[\boldsymbol{\alpha}, \boldsymbol{\beta}]$ since any other solution must also be a solution of (5.10) violating the uniqueness given by Theorem 6.

Finally we deduce the required convergence result for the limit Δx tends to zero. If $\boldsymbol{u}^{\Delta x}$ denotes the unique solution of (5.7) in $[\boldsymbol{\alpha}, \boldsymbol{\beta}]$, then from the above this is also the unique solution of (5.10). Application of Theorem 6 to the difference equations (5.10) states that the piecewise constant extension of this solution converges in L_1 to the limit solution of problem $P_\epsilon^{\alpha,\beta}$ as $\Delta x \to 0$. But from the proof of Theorem 2 this limit solution is also the limit solution of problem P_ϵ. This gives the required result and completes the proof of the theorem.

In the above theorem the conditions on the numerical flux function are essentially those of the original theory; however in this case we only require these conditions to hold over a finite range and we say that the scheme is monotone over this range. In return the theorem gives only the uniqueness of the solution in this range and does not preclude the existence of other solutions not wholly contained in this range. However if conditions 1-4 of Theorem 7 hold for all positive α and β then the system of difference equations will have only one positive solution, although this will not be the case for all the forms of g we consider.

Theorem 7 can straightforwardly be applied to the steady flow problem, under the conditions of Theorem 3, by transforming problem (4.8) onto the unit interval. We postpone this until section 5.3 and first consider a method for computing the solution of the system of difference equations.

5.2 The Time Stepping Iteration

In this section we consider a method for solving the system of difference equations. The theory relies on the following lemma which is simply an application of the contraction mapping theorem.

Lemma 5.1 *Consider the function* $\boldsymbol{G} : [\boldsymbol{c}, \boldsymbol{d}] \longrightarrow I\!R^N$ *where* $\boldsymbol{c}, \boldsymbol{d} \in I\!R^N$. *If* $\boldsymbol{G}$ *is such that*

(1) $\boldsymbol{G}(\boldsymbol{c}) \geq \boldsymbol{c}$ *and* $\boldsymbol{G}(\boldsymbol{d}) \leq \boldsymbol{d}$

(2) For each $\boldsymbol{u}_1$, $\boldsymbol{u}_2 \in [\boldsymbol{c}, \boldsymbol{d}]$,

$$\boldsymbol{G}(\boldsymbol{u}_2) - \boldsymbol{G}(\boldsymbol{u}_1) = M(\boldsymbol{u}_2 - \boldsymbol{u}_1),$$

where the matrix M *has all non-negative elements and* $\|M\|_1 \leq k < 1$ *with* k *independent of* $\boldsymbol{u}_1$ *and* $\boldsymbol{u}_2$

then we have

(1) $\boldsymbol{G}([\boldsymbol{c}, \boldsymbol{d}]) \subset [\boldsymbol{c}, \boldsymbol{d}]$

(2) The mapping $\boldsymbol{G}$ *has exactly one fixed point*

(3) Given any starting vector $\boldsymbol{u}^0 \in [\boldsymbol{c}, \boldsymbol{d}]$, *the sequence*

$$\boldsymbol{u}^{n+1} = \boldsymbol{G}(\boldsymbol{u}^n), \qquad n = 0, 1, 2, \ldots \tag{5.11}$$

converges to the fixed point $\boldsymbol{u}$ *as* $n \to \infty$ *and we have the convergence rate estimate*

$$\|\boldsymbol{u}^n - \boldsymbol{u}\|_1 \leq k^n \|\boldsymbol{u}^0 - \boldsymbol{u}\|_1 \leq \|\boldsymbol{u}^0 - \boldsymbol{u}\|_1 e^{-n(1-k)}.$$

Proof

Given any vectors $\boldsymbol{u}_1$, $\boldsymbol{u}_2 \in [\boldsymbol{c}, \boldsymbol{d}]$ with $\boldsymbol{u}_2 \geq \boldsymbol{u}_1$ we have

$$\boldsymbol{G}(\boldsymbol{u}_2) - \boldsymbol{G}(\boldsymbol{u}_1) = M(\boldsymbol{u}_2 - \boldsymbol{u}_1) \geq 0,$$

because the matrix M has all non-negative elements. Hence for arbitrary $\boldsymbol{u}_1 \in [\boldsymbol{c}, \boldsymbol{d}]$ we have by definition $\boldsymbol{c} \leq \boldsymbol{u}_1 \leq \boldsymbol{d}$ which implies

$$\boldsymbol{c} \leq \boldsymbol{G}(\boldsymbol{c}) \leq \boldsymbol{G}(\boldsymbol{u}_1) \leq \boldsymbol{G}(\boldsymbol{d}) \leq \boldsymbol{d}.$$

This demonstrates that $\boldsymbol{G}$ maps onto its own domain. The contractivity of $\boldsymbol{G}$ then gives the uniqueness and existence of the fixed point by the contraction mapping theorem (for example see ref. [46] section 5.1.3). The convergence of the sequence given by (5.11) arises from observing observing that

$$\|\boldsymbol{u}^n - \boldsymbol{u}\|_1 = \|\boldsymbol{G}(\boldsymbol{u}^{n-1}) - \boldsymbol{G}(\boldsymbol{u})\|_1 \leq k\|\boldsymbol{u}^{n-1} - \boldsymbol{u}\|_1,$$

and hence by induction

$$\|\boldsymbol{u}^n - \boldsymbol{u}\|_1 \leq k^n\|\boldsymbol{u}^0 - \boldsymbol{u}\|_1 \leq \|\boldsymbol{u}^0 - \boldsymbol{u}\|_1 e^{-n(1-k)}.$$

Here we have used the fact that $k \leq e^{-(1-k)}$ for $k \in [0,1]$. This completes the proof.

Under the conditions of Theorem 7 we apply the above lemma to the mapping

$$\boldsymbol{G} : [\boldsymbol{\alpha}, \boldsymbol{\beta}] \longrightarrow I\!R^{N+1},$$

given by

$$\boldsymbol{G}(\boldsymbol{u}) = \begin{bmatrix} \gamma_0 \\ u_1 - \Delta t \mathcal{T}_1 u \\ \vdots \\ u_j - \Delta t \mathcal{T}_j u \\ \vdots \\ u_{N-1} - \Delta t \mathcal{T}_{N-1} u \\ \gamma_1 \end{bmatrix}, \tag{5.12}$$

where $\Delta t > 0$. Notice that a vector $\boldsymbol{u}$ is a fixed point of this mapping if and only it is a solution of the difference equations (5.7). We first show that $\boldsymbol{G}$ has property (1) of the lemma. We have

$$\boldsymbol{G}(\boldsymbol{\alpha}) = \begin{bmatrix} \gamma_0 \\ \alpha - \Delta t b(x_1, \alpha) \\ \vdots \\ \alpha - \Delta t b(x_j, \alpha) \\ \vdots \\ \alpha - \Delta t b(x_{N-1}, \alpha) \\ \gamma_1 \end{bmatrix} \geq \boldsymbol{\alpha},$$

since $\alpha \leq \underline{u} = \min\{\gamma_0, \gamma_1, m\}$ and hence $b(x_j, \alpha) \leq 0$. Also we have

$$\boldsymbol{G}(\boldsymbol{\beta}) = \begin{bmatrix} \gamma_0 \\ \beta - \Delta t b(x_1, \beta) \\ \vdots \\ \beta - \Delta t b(x_j, \beta) \\ \vdots \\ \beta - \Delta t b(x_{N-1}, \beta) \\ \gamma_1 \end{bmatrix} \leq \boldsymbol{\beta},$$

since $\beta \geq \overline{u} = \max\{\gamma_0, \gamma_1, M\}$ and hence $b(x_j, \beta) \geq 0$.

We next investigate the circumstances under which $\boldsymbol{G}$ satisfies condition (2) of Lemma 5.1. For $u_1, u_2, v_1, v_2 \in [\alpha, \beta]$ we define the functions

$$l^u(u_2, u_1; v_1) = \begin{cases} \dfrac{g(u_2, v_1) - g(u_1, v_1)}{u_2 - u_1} & \text{if} \quad u_2 \neq u_1 \\ 0 & \text{if} \quad u_2 = u_1, \end{cases} \tag{5.13}$$

and

$$l^v(u_2, u_1; v_1) = \begin{cases} \dfrac{g(v_1, u_2) - g(v_1, u_1)}{u_2 - u_1} & \text{if} \quad u_2 \neq u_1 \\ 0 & \text{if} \quad u_2 = u_1, \end{cases} \tag{5.14}$$

which from the properties of g are bounded and satisfy

$$l^u(u_2, u_1; v_1) \leq 0, \qquad l^v(u_2, u_1; v_1) \geq 0.$$

We can now write

$$\begin{aligned} g(u_2, v_2) - g(u_1, v_1) &= g(u_2, v_2) - g(u_1, v_2) + g(u_1, v_2) - g(u_1, v_1) \\ &\quad l^u(u_2, u_1; v_2)(u_2 - u_1) + l^v(v_2, v_1; u_1)(v_2 - v_1). \end{aligned}$$

Using this relationship and applying the mean value theorem to the difference in the term involving b, we can for $\boldsymbol{u}, \boldsymbol{v} \in [\boldsymbol{\alpha}, \boldsymbol{\beta}]$ write

$$\boldsymbol{G}(\boldsymbol{u}) - \boldsymbol{G}(\boldsymbol{v}) = M(\boldsymbol{u} - \boldsymbol{v}),$$

where the tri-diagonal matrix M is given by

$$\begin{pmatrix} 0 & & & & & & \\ p_1 & q_1 & r_1 & & & & \\ & \ddots & \ddots & \ddots & & & \\ & & p_j & q_j & r_j & & \\ & & & \ddots & \ddots & \ddots & \\ & & & & p_{N-1} & q_{N-1} & r_{N-1} \\ & & & & & & 0 \end{pmatrix}, \tag{5.15}$$

with

$$p_j = \frac{\Delta t}{\Delta x} l^v(u_{j-1}, v_{j-1}; v_j),$$

$$\begin{aligned} q_j &= 1 - \Delta t \left(\frac{l^v(u_j, v_j; v_{j+1}) - l^u(u_j, v_j; u_{j-1})}{\Delta x} + b_u(x_j, \hat{u}_j) \right) \\ &= 1 - \Delta t b_u(x_j, \hat{u}_j) - p_{j+1} - r_{j-1}, \end{aligned}$$

$$r_j = -\frac{\Delta t}{\Delta x} l^u(u_{j+1}, v_{j+1}; u_j),$$

and $\hat{u}_j \in [\alpha, \beta]$. We have that $p_j, r_j \geq 0$, and if we require that Δt be chosen such that

$$\begin{aligned} \Delta t \left(\frac{l^v(u_1, u_2; u_3) - l^u(u_1, u_2; u_4)}{\Delta x} + b_u(x_j, u_5) \right) \leq 1, \\ \text{for all } u_1, u_2, u_3, u_4, u_5 \in [\alpha, \beta] \text{ and } 0 \leq j \leq N, \end{aligned} \tag{5.16}$$

then it follows also that $q_j \geq 0$. It is always possible to choose such a value for Δt since the coefficient of Δt is positive and bounded.

To compute the L_1 norm of the matrix M we need to compute the sum of each column. The sum of the first column is p_1. If

$$\delta = \min_{\substack{0 \leq j \leq N \\ \alpha \leq u \leq \beta}} \{b_u(x_j, u)\}, \tag{5.17}$$

then using condition (5.16) with the correct values we obtain

$$p_1 \leq 1 - \Delta t b_u(x_0, u_0) + \frac{l^u(u_0, v_0; u_0)}{\Delta x} \leq 1 - \Delta t \delta,$$

since $l^u \leq 0$. The sum of the second column is given by

$$q_1 + p_2 = 1 - \Delta t b_u(x_1, \hat{u}_1) - r_0 \leq 1 - \Delta t \delta,$$

since $r_0 \geq 0$. For the j^{th} column $(3 \leq j \leq N-2)$ the sum is given by

$$r_{j-2} + q_{j-1} + p_j = 1 - \Delta t b_u(x_{j-1}, \hat{u}_{j-1}) \leq 1 - \Delta t \delta.$$

The same argument shows that the remaining two column sums satisfy the same bound, hence we conclude that

$$\|M\|_1 \leq 1 - \Delta t \delta < 1.$$

We can obtain a slightly less restrictive requirement on the parameter Δt than given by (5.16) if the function $g(u, v)$ is assumed to be continuously differentiable for all $u, v \in [\alpha, \beta]$. In this case the function $\boldsymbol{G}$ is Frechet-differentiable and for $\boldsymbol{u}, \boldsymbol{v} \in [\boldsymbol{\alpha}, \boldsymbol{\beta}]$ we can write

$$\boldsymbol{G}(\boldsymbol{u}) - \boldsymbol{G}(\boldsymbol{v}) = M(\boldsymbol{u} - \boldsymbol{v}),$$

where

$$M = \int_0^1 \boldsymbol{G}'(\boldsymbol{u} + s(\boldsymbol{v} - \boldsymbol{u}))ds,$$

(see [46], sections 3.2.6 and 3.2.8). The Jacobian $\boldsymbol{G}'(\boldsymbol{u})$ is again of the form (5.15) where now

$$p_j = \frac{\Delta t}{\Delta x} g_v(u_j, u_{j-1}),$$

$$\begin{aligned} q_j &= 1 - \Delta t \left(\frac{g_v(u_{j+1}, u_j) - g_u(u_j, u_{j-1})}{\Delta x} + b_u(x_j, u_j) \right) \\ &= 1 - \Delta t b_u(x_j, u_j) - p_{j+1} - r_{j-1}, \end{aligned}$$

$$r_j = -\frac{\Delta t}{\Delta x} g_u(u_{j+1}, u_j).$$

As before $p_j, r_j \geq 0$, but in this case the condition

$$\Delta t \left(\frac{g_v(u_1, u_2) - g_u(u_2, u_3)}{\Delta x} + b_u(x_j, u_2) \right) \leq 1, \qquad \text{for all } u_1, u_2, u_3 \in [\alpha, \beta] \text{ and } 0 \leq j \leq N, \tag{5.18}$$

is sufficient to ensure $q_j \geq 0$.

We estimate the L_1 norm of the matrix $\boldsymbol{G}'(u)$ by computing the sum of each column. The sum of the first column is p_1, and using condition (5.18) with the correct values we obtain

$$p_1 \leq 1 - \Delta t b_u(x_0, u_0) + \frac{g_u(u_0, u_0)}{\Delta x} \leq 1 - \Delta t \delta,$$

since $g_u \leq 0$. The sum of the second column is given by

$$q_1 + p_2 = 1 - \Delta t b_u(x_1, u_1) - r_0 \leq 1 - \Delta t \delta,$$

since $r_0 \geq 0$. For the j^{th} column $(3 \leq j \leq N-2)$ the sum is given by

$$r_{j-2} + q_{j-1} + p_j = 1 - \Delta t b_u(x_{j-1}, u_{j-1}) \leq 1 - \Delta t \delta.$$

The same argument shows that the remaining two column sums satisfy the same bound, hence we conclude that

$$\|\boldsymbol{G}'(u)\|_1 \leq 1 - \Delta t \delta.$$

It follows that

$$\|M\|_1 = \left\| \int_0^1 \boldsymbol{G}'(\boldsymbol{u} + s(\boldsymbol{v} - \boldsymbol{u}))ds \right\|_1 \leq \int_0^1 \|\boldsymbol{G}'(\boldsymbol{u} + s(\boldsymbol{v} - \boldsymbol{u}))\|_1 ds \leq 1 - \Delta t \delta < 1.$$

From the above discussion we obtain the following Theorem.

Theorem 8 *Suppose the situation is as in Theorem 7 and that either (i) $\Delta t > 0$ satisfies condition (5.16) or (ii) the function $g(u, v)$ is continuously differentiable for all $u, v > 0$ and $\Delta t > 0$ satisfies condition (5.18). Under these conditions the the mapping (5.12) has exactly one fixed point $\boldsymbol{u}$ which is the only solution in $[\boldsymbol{\alpha}, \boldsymbol{\beta}]$ of the difference equations (5.7). For any initial guess $\boldsymbol{u}^0 \in [\boldsymbol{\alpha}, \boldsymbol{\beta}]$ the iteration*

$$\boldsymbol{u}^{n+1} = \boldsymbol{G}(\boldsymbol{u}^n), \qquad n = 0, 1, 2, \ldots$$

converges to the fixed point as $n \to \infty$ and we have the convergence rate estimate

$$\|\boldsymbol{u}^n - \boldsymbol{u}\|_1 \leq (1 - \Delta t \delta)^n \|\boldsymbol{u}^0 - \boldsymbol{u}\|_1 \leq \|\boldsymbol{u}^0 - \boldsymbol{u}\|_1 e^{-n\Delta t \delta},$$

where δ is given by (5.17) and $0 \leq 1 - \Delta t \delta < 1$.

The above theorem demonstrates that we may compute the solution of the system of difference equations simply by computing the sequence of vectors $\boldsymbol{u}^{n+1} = \boldsymbol{G}(\boldsymbol{u}^n)$, $n = 0, 1, \ldots$ with $\boldsymbol{u}^0$ an arbitrary starting vector. We can rearrange this iteration to obtain

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + \frac{g(u_{j+1}^n, u_j^n) - g(u_j^n, u_{j-1}^n)}{\Delta x} = -b(x_j, u_j^n), \tag{5.19}$$

$j = 1, 2 \ldots, N-1$, where $u_0^n = \gamma_0$ and $u_N^n = \gamma_1$. This is a first order time accurate approximation to the partial differential equation

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x} f(u) = -b(x,u), \quad 0 \leq x \leq 1,$$

and the condition on the time step Δt plays the role of the CFL condition.

5.3 Application to the Steady Flow Problem

In this section we apply the theory from the previous section to the steady flow problem. Application of Theorem 7 under the conditions of Theorem 3 yields the following theorem. Any discussion throughout the remainder of this chapter assumes that the problem satisfies the conditions of Theorem 3 and also the additional assumptions of Theorem 5.

Theorem 9 *Suppose the situation is as in Theorem 3 and consider the difference equations*

$$\begin{aligned} \mathcal{T}_j h = 0, \qquad & j = 1, 2, \ldots, N-1 \\ h_0 = \gamma_0, \quad & h_N = \gamma_1, \end{aligned} \tag{5.20}$$

where

$$\mathcal{T}_j h = \frac{g(h_{j+1}, h_j) - g(h_j, h_{j-1})}{\Delta x} + D(x_j, h_j),$$

$\Delta x = L/N$ *and* $x_j = j\Delta x$. *Suppose that for some* $0 < \alpha \leq \underline{h}$, $\beta \geq \overline{h}$ *the function* g *has the properties:*

(1) $g(u,u) = -F(u)$ *for all* $u \in [\alpha, \beta]$.

(2) u_1, u_2, $v \in [\alpha, \beta]$ *with* $u_2 \geq u_1$ *implies* $g(u_2, v) \leq g(u_1, v)$.

(3) v_1, v_2, $u \in [\alpha, \beta]$ *with* $v_2 \geq v_1$ *implies* $g(u, v_2) \geq g(u, v_1)$.

(4) There exists a constant l *such that for all* u_1, u_2, v_1, $v_2 \in [\alpha, \beta]$

$$|g(u_2, v_2) - g(u_1, v_1)| \leq l(|u_2 - u_1| + |v_2 - v_1|).$$

Under these conditions the difference equations have a unique solution

$$\boldsymbol{h}^{\Delta x} = (h_0^{\Delta x}, h_1^{\Delta x}, \ldots, h_N^{\Delta x})^T$$

in $[\boldsymbol{\alpha}, \boldsymbol{\beta}]$ *for each* $N \in \mathbb{N}$. *This solution satisfies the bounds:*

$$0 < \underline{h} \leq h_j^{\Delta x} \leq \overline{h}, \qquad j = 0, 1, \ldots, N. \tag{5.21}$$

If $H^{\Delta x} \in L_1[0, L]$ *denotes the piecewise constant extension of this discrete solution given by*

$$H^{\Delta x} = h_j^{\Delta x} \quad \textit{for} \quad j\Delta x \leq x < (j+1)\Delta x, \quad j = 0, 1, \ldots, N \tag{5.22}$$

then $H^{\Delta x} \to H$ *in* L_1 *as* $\Delta x \to 0$, *where* $H \in NBV_+[0, L]$ *is the limiting solution of problem* (4.8) *as* $\epsilon \downarrow 0$.

Proof of Theorem 9 As in the proof of Theorem 3 we write

$$\begin{aligned} f(u) &= -LF(u), \\ b(x, u) &= L^2 D(xL, u), \end{aligned}$$

to transform the continuous problem (4.8) onto the unit interval. This problem then satisfies the conditions of Theorem 2. Likewise transform the discrete problem onto the unit interval by writing

$$\begin{aligned} u_j &= h_j, \\ \tilde{g}(u, v) &= Lg(u, v), \\ \Delta\tilde{x} &= \Delta x / L, \end{aligned}$$

giving

$$\tilde{\mathcal{T}}_j u = \frac{\tilde{g}(u_{j+1}, u_j) - \tilde{g}(u_j, u_{j-1})}{\Delta\tilde{x}} + b(x_j, u_j) = 0, \qquad j = 1, 2, \ldots, N-1$$
$$u_0 = \gamma_0, \quad u_N = \gamma_1.$$

Theorem 7 is directly applicable to this system and the result is obtained by writing the conclusions of Theorem 7 in terms of the original variables.

For a scheme which is monotone on the interval $[\alpha, \beta]$, the above theorem gives the existence and uniqueness of the system of difference equation in $[\boldsymbol{\alpha}, \boldsymbol{\beta}]$. More importantly the theorem demonstrates the convergence of the discrete solution to the vanishing viscosity solution of problem (4.8), which is a physical solution of the

steady flow problem. Moreover under the conditions of section 4.9 we demonstrated that any physical solution of the steady flow problem is a vanishing viscosity solution of (4.8) for the appropriate choice of γ_0 and γ_1. Hence by choosing the correct values for these boundary values the solution of the difference equations can approximate any physical solution of the steady flow problem we require. The convergence result is not as strong as we would like, since in general we compute the solution for only one grid spacing and the theory gives no indication of the "closeness" of the discrete solution to the exact solution. It may be that N is required to be unrealistically large before an acceptable approximation is obtained. This is not expected to be the case, since such schemes have been successful in many other applications. The following corollary to Theorem 9 gives conditions under which the solution of the difference equations is globally unique. The proof follows trivially from Theorem 9.

Corollary 9.1 *Suppose that the conditions of Theorem 3 hold and that g satisfies the conditions 1-4 of Theorem 9 for any positive α and β, then the system of difference equations* (5.20) *have exactly one positive solution.*

We next apply the theory for the time stepping iteration to the steady flow problem to obtain the following Theorem.

Theorem 10 *Suppose the situation is as in Theorem 7 and that* at least one *of the following conditions holds.*

(1) $\Delta t > 0$ *satisfies*

$$\Delta t \left(\frac{l^v(h_1, h_2; h_3) - l^u(h_1, h_2; h_4)}{\Delta x} + D_h(x_j, h_5) \right) \leq 1, \quad \text{for all } h_1, h_2, h_3, h_4, h_5 \in [\alpha, \beta] \text{ and } 0 \leq j \leq N, \tag{5.23}$$

where l^u, l^v are given by (5.13) *and* (5.14).

(2) The function $g(u, v)$ is continuously differentiable for all $u, v > 0$ and $\Delta t > 0$ satisfies

$$\Delta t \left(\frac{g_v(h_1, h_2) - g_u(h_2, h_3)}{\Delta x} + D_h(x_j, h_2) \right) \leq 1, \quad \text{for all } h_1, h_2, h_3 \in [\alpha, \beta] \text{ and } 0 \leq j \leq N. \tag{5.24}$$

Under these conditions the mapping (5.12) *has exactly one fixed point* $\boldsymbol{h}$ *which is the only solution in* $[\boldsymbol{\alpha}, \boldsymbol{\beta}]$ *of the difference equations* (5.20). *For any initial guess* $\boldsymbol{h}^0 \in [\boldsymbol{\alpha}, \boldsymbol{\beta}]$ *the iteration*

$$\boldsymbol{h}^{n+1} = \boldsymbol{G}(\boldsymbol{h}^n), \qquad n = 0, 1, 2, \ldots$$

converges to the fixed point as $n \to \infty$ *and we have the convergence rate estimate*

$$\|\boldsymbol{h}^n - \boldsymbol{h}\|_1 \leq (1 - \Delta t\delta)^n \|\boldsymbol{h}^0 - \boldsymbol{h}\|_1 \leq \|\boldsymbol{u}^0 - \boldsymbol{u}\|_1 e^{-n\Delta t\delta}, \tag{5.25}$$

where

$$\delta = \min_{\substack{0 \leq j \leq N \\ \alpha \leq h \leq \beta}} \{D_h(x_j, h)\},$$

and $0 \leq 1 - \Delta t\delta < 1$.

Proof of Theorem 10 We apply the same transformation as in the proof of Theorem 9 to the mapping (5.12) with $\Delta t = L^2 \Delta \tilde{t}$ to gain a mapping of the same form with the addition of tildes to the appropriate variables. Applying Theorem 8 and writing the conditions on $\Delta \tilde{t}$ in terms of the original variables, gives the required result.

5.4 Numerical Flux Functions giving Monotone Schemes

In this section we consider the various well known forms of the numerical flux function g which satisfy conditions 1-4 of Theorem 9. These are the Engquist-Osher, Godunov and Lax-Friedrichs forms and are those which give rise to monotone schemes for scalar conservation laws. Although the functions can easily be written in terms of the quantity F, we write them in terms of the quantity $f = -F$ so that the forms can be easily recognized from the literature.

The Engquist-Osher numerical flux function is given by

$$g(u, v) = f_-(u) + f_+(v) + f(c), \tag{5.26}$$

where

$$\begin{aligned} f_-(u) &= \int_c^u \min\{f'(s), 0\} ds, \\ f_+(u) &= \int_c^u \max\{f'(s), 0\} ds \end{aligned}$$

and $c > 0$ is arbitrary. This choice of g is continuously differentiable and we have

$$\begin{aligned} g_u(u,v) &= f'_-(u) = \min\{f'(u), 0\} \leq 0, \\ g_v(u,v) &= f'_+(v) = \max\{f'(v), 0\} \geq 0. \end{aligned}$$

Also we have $g(u,u) \equiv f(u)$ for positive u, so that conditions 1-4 of Theorem 9 are satisfied for any choice of α and β. Hence under the conditions of Theorem 3 the conclusions of Theorem 9 are true. Theorem 9 does not show the global uniqueness of the discrete solution, however since the conditions on g hold for arbitrary α and β, the corollary to the theorem is valid, giving the global uniqueness.

Using the fact that g is continuously differentiable, the stronger form of the CFL condition (condition 5.24) is sufficient to guarantee convergence of the time stepping iteration. This condition reduces to the requirement that

$$\Delta t \left(\frac{|f'(h)|}{\Delta x} + D_h(x_j, h) \right) \leq 1, \tag{5.27}$$

for all $h \in [\alpha, \beta]$ and $0 \leq j \leq N$. This is now more recognisable as being a CFL condition. The second term, due to the presence of the source term, diminishes in influence as the grid is refined. For small Δx we essentially return to the traditional requirement that

$$\frac{\Delta t}{\Delta x} |f'(h)| \leq 1,$$

at all times. In general the function f is unbounded as the depth tends to zero or tends to infinity, so we cannot choose a single time step which satisfies the CFL condition (5.27) for all values of α and β. Thus the allowable time step will be dependent on the a-priori bounds on the solution and the choice of initial data. We discuss in the next section how to find an allowable range for Δt in practice.

The Godunov form of the numerical flux function is given by

$$g(u,v) = \begin{cases} \max\{f(w) : u \leq w \leq v\} & \text{for } u \leq v \\ \min\{f(w) : v \leq u \leq v\} & \text{for } v \leq u. \end{cases} \tag{5.28}$$

In this case g is not everywhere differentiable. It is not difficult to demonstrate that if $h_1, h_2, h_3 \in [\alpha, \beta]$, then

$$0 \leq -l^u(h_1, h_2; h_3) \leq \overline{|f'|} \tag{5.29}$$

and

$$0 \leq l^v(h_1, h_2; h_3) \leq \overline{|f'|}, \tag{5.30}$$

where

$$\overline{|f'|} = \max_{\alpha \leq h \leq \beta} \{|f'(h)|\}.$$

Since the consistency condition also holds then g satisfies conditions 1-4 of Theorem 9, hence under the conditions of Theorem 3 the conclusions of Theorem 9 are true. Again the conditions on g hold for arbitrary α and β, thus the corollary to the Theorem gives the global uniqueness of the discrete solution.

Since in this case the numerical flux function is not everywhere differentiable, we must use the weaker form of the CFL condition (condition 5.23) to ensure the convergence of the time stepping iteration. Using the bounds (5.29) and (5.30), the requirement that

$$\Delta t \left(\frac{2|f'(h_1)|}{\Delta x} + D_h(x_j, h_2) \right) \leq 1, \tag{5.31}$$

for all $h_1, h_2 \in [\alpha, \beta]$ and $0 \leq j \leq N$ can be seen to be sufficient. The difference between this condition and that for the Engquist-Osher form is the addition of the factor two in the first term. Hence as Δx becomes small the condition will only allow a time step of half that allowed by the Engquist-Osher scheme. It is likely that more thorough analysis, using the fact that g is only non-differentiable on isolated curves in the u-v space, can eliminate this extra factor from the CFL condition.

The Lax-Friedrichs form of the numerical flux function is given by

$$g(u, v) = \frac{1}{2} \left(f(u) + f(v) + \lambda(v - u) \right), \tag{5.32}$$

where λ is some parameter to be chosen. This form of g is continuously differentiable and we have

$$g_u(u, v) = \frac{1}{2} \left(f'(u) - \lambda \right)$$

and

$$g_v(u, v) = \frac{1}{2} \left(f'(v) + \lambda \right).$$

Since in general f' is unbounded as the depth tends to zero or infinity, we cannot choose a single value of the parameter λ such that $g_u \leq 0$ and $g_v \geq 0$ for all positive

values of u and v. The best we can achieve is to enforce these conditions to hold over the finite range $\alpha \leq u, v \leq \beta$, by taking λ such that

$$\lambda \leq |f'(h)|, \quad \text{for all } \alpha \leq h \leq \beta. \tag{5.33}$$

Conditions 1-4 of Theorem 9 are then satisfied, and under the conditions of Theorem 3 the conclusions of Theorem 9 are true. In this case, however, the corollary to Theorem 9 is not valid and thus the discrete solution may not be globally unique. The theory does not preclude the existence of other solutions which are not contained in the set $[\boldsymbol{\alpha}, \boldsymbol{\beta}]$. This is not significant if we intend to use the time stepping iteration to solve the difference equations, since this is guaranteed to converge to the solution which is contained in $[\boldsymbol{\alpha}, \boldsymbol{\beta}]$, and is the solution that converges to the vanishing viscosity solution as $\Delta x \downarrow 0$. For other methods there is usually no such guarantee and the possibility of obtaining unphysical solutions cannot be ruled out.

For the Lax-Friedrichs form the stronger form of the CFL condition reduces to the requirement that

$$\Delta t \left(\frac{\lambda}{\Delta x} + D_h(x_j, h) \right) \leq 1, \tag{5.34}$$

for all $h \in [\alpha, \beta]$ and $0 \leq j \leq N$. The requirement (5.33) means that the condition (5.27) is sufficient to ensure convergence.

In practice the channel cross-sections considered in this thesis all satisfy not only the conditions of Theorem 3 but also the extra requirements of Theorem 5 which essentially require there to be only one critical depth h_c. Under these conditions we have

$$\begin{aligned} f'(h) > 0 \quad &\text{for} \quad h < h_c, \\ f'(h_c) = 0, & \\ f'(h) < 0 \quad &\text{for} \quad h > h_c \end{aligned}$$

These properties hold if f is concave and the Engquist-Osher and Godunov numerical flux functions simplify to the same forms as for a concave f. In the case of the

Engquist-Osher scheme we have

$$g(u,v) = \begin{cases} f(v) & u \le h_c, v \le h_c \\ f(h_c) & u \le h_c, v \ge h_c \\ f(u) + f(v) - f(h_c) & u \ge h_c, v \le h_c \\ f(u) & u \ge h_c, v \ge h_c \end{cases} \tag{5.35}$$

and for the Godunov scheme we have

$$g(u,v) = \begin{cases} f(v) & u \le h_c, v \le h_c \\ f(h_c) & u \le h_c, v \ge h_c \\ \min\{f(u), f(v)\} & u \ge h_c, v \le h_c \\ f(u) & u \ge h_c, v \ge h_c. \end{cases} \tag{5.36}$$

These two forms differ in only one quadrant of the u-v plane. If we consider the terms in the difference equations of the form $g(h_{j+1}, h_j)$, then the two forms are only different in the case $h_{j+1} \ge h_c$, $h_j \le h_c$ which corresponds to a hydraulic jump. In a sense the Godunov form is the one to compare other forms with since, for the Riemann problem for scalar homogeneous conservation laws, it gives the exact flux across the center for both rarefaction waves and shocks (see section 3.2). The Engquist-Osher form gives the correct flux for a rarefaction wave, but not in general for a shock. Compare also the numerical flux function for the first-order upwind scheme which reduces to

$$g(u,v) = \begin{cases} f(v) & u \le h_c, v \le h_c \\ \max\{f(u), f(v)\} & u \le h_c, v \ge h_c \\ \min\{f(u), f(v)\} & u \ge h_c, v \le h_c \\ f(u) & u \ge h_c, v \ge h_c. \end{cases} \tag{5.37}$$

Again this form only differs from the Godunov version in one quadrant, in this case corresponding to a smooth transition (steady rarefaction wave). The above form of g does not satisfy conditions 2 and 3 of Theorem 9 for any range of depths which includes the critical depth. For example if $u < h_c$ and $v > h_c$ with $f(u) \ge f(v)$, then we have $g(u,v) = \max\{f(u), f(v)\} = f(u)$ and $g_u(u,v) = f'(u) > 0$.

5.5 Theory into Practice

In this section we describe how to carry out the necessary steps to utilise the theory and obtain an efficient, robust and practical algorithm for computing solutions to the steady flow problem. We consider the following five steps.

(1) Choose the values for γ_0 and γ_1.

(2) Determine bounds on the normal depth for the problem and hence find bounds on the exact solution.

(3) Choose the starting vector $\boldsymbol{h}^0$ for the time stepping iteration and then appropriate values for α and β.

(4) Ensure the numerical flux function satisfies conditions 1-4 of Theorem 9.

(5) Find a value of Δt which satisfies the CFL condition and hence guarantees the convergence of the time stepping iteration.

For a given problem, the first step is to choose values for γ_0 and γ_1 in order to give the required solution. In section 4.9 we observed that there are essentially four types of problem to solve. These are (i) the flow is supercritical at inflow and supercritical at outflow with the depth specified at inflow, (ii) the flow is subcritical at inflow and outflow with the depth specified at outflow, (iii) the flow is supercritical at inflow and is subcritical at outflow with the depth specified at both inflow and outflow, and (iv) the flow is subcritical at inflow and supercritical at outflow with the depth specified at neither end of the reach. Section 4.9 describes the appropriate values for γ_0 and γ_1 such that in each case the vanishing viscosity solution of the singular perturbation problem is the required solution of the steady flow problem. From Theorem 9 the solution of the difference equations converges to the vanishing viscosity solution as $\Delta x \downarrow 0$, so we retain the same choice of γ_0 and γ_1 for the discrete system. To recap briefly: if the depth is specified at inflow then γ_0 is taken to be this value otherwise we set $\gamma_0 \geq h_c$. If the depth is specified at outflow then we take γ_1 to be this value otherwise we set $0 < \gamma_1 \leq h_c$. We will see later in this section that when the depth is not specified at a boundary, then it is advantageous to take

the corresponding γ_j as the critical depth since this minimises the range over which the CFL condition must hold.

The next step is to obtain bounds for the normal depth h_n. The normal depth is defined by

$$K(h_n(x)) = \frac{Q}{\sqrt{S_0(x)}}.$$

The conditions of Theorem 9 imply that the conveyance K is a strictly increasing function of depth, so that the normal depth is a strictly decreasing function of bed slope. Thus if S_0^1 and S_0^2 are such that

$$0 < S_0^1 \leq S_0(x) \leq S_0^2, \quad \text{for } 0 \leq x \leq L,$$

and h_n^1, h_n^2 solve

$$K(h_n^1) = \frac{Q}{\sqrt{S_0^2}} \tag{5.38}$$

and

$$K(h_n^2) = \frac{Q}{\sqrt{S_0^1}}, \tag{5.39}$$

respectively, then we have

$$h_n^1 \leq h_n(x) \leq h_n^2, \quad \text{for } 0 \leq x \leq L.$$

The bounds on the exact solution given by Theorem 9 are

$$\underline{h} \leq h(x) \leq \overline{h}, \quad \text{for } 0 \leq x \leq L,$$

where we have

$$\overline{h} = \max_{0 \leq x \leq 1} \{h_n(x), \gamma_0, \gamma_1\} \leq \min\{h_n^2, \gamma_0, \gamma_1\}$$

and

$$\underline{h} = \min_{0 \leq x \leq 1} \{h_n(x), \gamma_0, \gamma_1\} \geq \min\{h_n^1, \gamma_0, \gamma_1\}.$$

Equations (5.38) and (5.39) can be solved simply by using Newton-Raphson.

The initial vector $\boldsymbol{h}^0$ for the time stepping iteration can be taken as any positive vector and the α and β are then required to be such that $0 < \alpha \leq \underline{h}$, $\beta \geq \overline{h}$ and $\boldsymbol{h}^0 \in [\boldsymbol{\alpha}, \boldsymbol{\beta}]$. The values

$$\alpha = \min_{0 \leq j \leq N} \{h_j^0, h_n^1, \gamma_0, \gamma_1\}$$

and

$$\beta = \max_{0 \le j \le N} \{h_j^0, h_n^2, \gamma_0, \gamma_1\}$$

achieve this. Since the CFL condition is to hold over the range $\alpha \le h \le \beta$, we would like this range to be as small as possible. To achieve this we take the initial vector such that $\gamma_0 \le h_j^0 \le \gamma_1$ for $0 \le j \le N$. For example we could vary the initial data linearly between the end values γ_0 and γ_1. We can now use the values

$$\alpha = \min\{h_n^1, \gamma_0, \gamma_1\}$$

and

$$\beta = \max\{h_n^2, \gamma_0, \gamma_1\}.$$

Note also that if the depth is not specified at a particular end of the channel, then the tightest bounds on the solution and the smallest interval $[\alpha, \beta]$ are obtained by setting the corresponding γ_j to the critical depth.

The next step is to ensure that the numerical flux function g satisfies conditions 1-4 of Theorem 9 for the choice of α and β. This step is achieved automatically for the Engquist-Osher and Godunov forms. For the Lax-Friedrichs form we must choose the parameter λ such that (5.33) is satisfied. In general to find a bound for $|f'|$ we must use a graphical method or an automated algorithm with some kind of sampling procedure. However the bound can be found analytically for certain types of cross-section. For example if we consider the trapezoidal family of channels, where the width is given by

$$T(h) = B + zh \qquad (B, z \ge 0, B + z > 0), \tag{5.40}$$

then $f'(h_c) = 0$ and

$$f''(h) = -\frac{2Q^2 z^2}{h(B+hz)^3} - \frac{2Q^2 z}{h^2(B+hz)^2} - \frac{2Q^2}{h^3(B+hz)} - g(B+2hz) < 0.$$

In this case we have

$$\max_{\alpha \le h \le \beta} \{|f'(h)|\} = \max\left\{f'(\min\{\alpha, h_c\}), -f'(\max\{\beta, h_c\})\right\}. \tag{5.41}$$

The final step is find a value for the time step which satisfies the appropriate CFL condition and hence ensures the convergence of the time stepping iteration.

The convergence rate estimate (5.25) indicates that a larger value of Δt can yield a faster rate of convergence, so it is of interest to find the greatest time step allowable by the CFL condition. For all the three forms of g discussed in the previous section, the CFL condition can be written as

$$\Delta t \Gamma(x_j, h_1, h_2) \leq 1,$$

for $j = 0, 1, \ldots, N$ and all $h_1, h_2 \in [\alpha, \beta]$. The greatest allowable time step is then given by

$$\Delta t_{\text{opt}} = \left(\max_{\substack{0 \leq j \leq N \\ h_1, h_2 \in [\alpha, \beta]}} \{\Gamma(x_j, h_1, h_2)\} \right)^{-1}.$$

We are therefore required to maximise the function Γ over at most three parameters (actually two in the case of the Engquist-Osher and Lax-Friedrichs forms). The first step is to observe that in all three cases

$$\frac{\partial \Gamma}{\partial S_0} = \frac{\partial D_h}{\partial S_0} = gT > 0,$$

so that Γ is increasing in bed slope. Thus if j_1 and j_2 are such that

$$S_0(x_{j_1}) \leq S_0(x_j) \leq S_0(x_{j_2}), \quad 0 \leq j \leq N,$$

then

$$\max_{\substack{0 \leq j \leq N \\ h_1, h_2 \in [\alpha, \beta]}} \{\Gamma(x_j, h_1, h_2)\} = \max_{h_1, h_2 \in [\alpha, \beta]} \{\Gamma(x_{j_2}, h_1, h_2)\},$$

reducing by one the number of variables over which we must maximise. This leaves only one variable for the Engquist-Osher and Lax-Friedrichs forms. For a general channel cross-section this must be carried out graphically or automated using a sampling procedure. However in some cases further progress can be made analytically. For the Engquist-Osher form we have

$$\begin{aligned} \Gamma(x_{j_2}, h_1, h_2) &= \frac{|f'(h_1)|}{\Delta x} + D_h(x_{j_2}, h_1) \\ &\leq \frac{1}{\Delta x} \max_{h \in [\alpha, \beta]} \{|f'(h)|\} + \max_{h \in [\alpha, \beta]} \{D_h(x_{j_2}, h)\}. \end{aligned}$$

For the Godunov scheme we have

$$\begin{aligned} \Gamma(x_{j_2}, h_1, h_2) &= \frac{2|f'(h_1)|}{\Delta x} + D_h(x_{j_2}, h_2) \\ &\leq \frac{2}{\Delta x} \max_{h \in [\alpha, \beta]} \{|f'(h)|\} + \max_{h \in [\alpha, \beta]} \{D_h(x_{j_2}, h)\}. \end{aligned}$$

For the Lax-Friedrichs scheme we have

$$\begin{aligned}\Gamma(x_{j_2}, h_1, h_2) &= \frac{\lambda}{\Delta x} + D_h(x_{j_2}, h_1) \\ &\leq \frac{\lambda}{\Delta x} + \max_{h\in[\alpha,\beta]}\{D_h(x_{j_2}, h)\}.\end{aligned}$$

To compute an upper bound for $\Gamma(x_{j_2}, h_1, h_2)$ in each case only requires upper bounds for the terms $|f'(h)|$ and $D_h(x_{j_2}, h)$. If we again consider the trapezoidal class of channel cross-sections, then we have already observed that the maximum of $|f'|$ is given by (5.41). For the choice of Manning friction or Chezy friction it can be shown that

$$\frac{d^2}{dh^2}\left(\frac{A}{K^2}\right) > 0,$$

hence since

$$D_h = g\left(T\,S_0 - Q^2\frac{d}{dh}\left(\frac{A}{K^2}\right)\right)$$

and $T' > 0$, it follows that

$$\max_{h\in[\alpha,\beta]}\{D_h(x_{j_2}, h)\} \leq g\left(T(\beta)S_0(x_{j_2}) - Q^2\frac{d}{dh}\left(\frac{A}{K^2}\right)\Big|_{h=\alpha}\right).$$

Combining the above bounds yields an upper bound for $\Gamma(x_{j_2}, h_1, h_2)$, and thus a lower bound for Δt_{opt}. The difference between this lower bound and the optimum value decreases as Δx becomes smaller since the relative importance of the term D_h decreases. We can also use the same idea as above to obtain a lower bound for the convergence rate δ in equation (5.25). This is given by

$$\delta = \min_{\substack{0\leq j\leq N \\ h\in[\alpha,\beta]}}\{D_h(x_j, h)\} \geq g\left(T(\alpha)S_0(x_{j_1}) - Q^2\frac{d}{dh}\left(\frac{A}{K^2}\right)\Big|_{h=\beta}\right).$$

Chapter 6

Test Problems with Analytic Solutions

In many areas of computational fluid dynamics there are benchmark test problems which have known analytic solutions. The performance of a particular numerical scheme can be evaluated by using some measure of the difference between the numerical solution and the exact solution. Important features of the solution can be compared with the exact solution; for example the position and strength of any shocks can be assessed. An acceptable level of performance over a wide range of such benchmark test problems leads to confidence that a numerical scheme will perform acceptably for any practical problems which are not too dissimilar. Altogether, benchmark test problems with known solutions are an extremely useful tool.

For the steady open channel problem, because of the non-linear nature of the differential equation, particularly the friction term, even the simplest problem of flow in a uniform rectangular channel with zero bed slope cannot be solved analytically. To obtain a solvable problem it is necessary to assume zero friction or at least to simplify the friction term significantly. However features of the problem will then be lost. For this reason, until now, the performance of methods for steady computation have mostly been judged only qualitatively.

This chapter presents a simple method for constructing test problems with known analytic solutions to the full steady Saint-Venant equation. The method is an "inverse method" in that some hypothetical depth profile is chosen and the bed slope that

makes this profile an actual solution of the steady equation is then found. The method can be used to construct test problems with almost any desired features, including hydraulic jumps. Hence these test problems can be used to compare the numerical results, for any algorithm, with an exact solution. The method is also useful for evaluating unsteady solvers, since, if an unsteady model is given steady boundary conditions, the limiting steady solution can be compared with the analytic steady solution. The method presented in this chapter fits in well with the validation documentation initiative of the European hydraulics laboratories (see [12]), since it enables the creation of benchmark test problems which can be used as a standard measure for the performance of commercial software packages.

6.1 Test Problems with Smooth Solutions

It is convenient to write equation (2.25) as

$$S_0(x) = f_1(x, h(x))h'(x) + f_2(x, h(x)), \tag{6.1}$$

where

$$f_1 = 1 - \frac{Q^2 T}{gA^3} = 1 - F_r^2 \tag{6.2}$$

and

$$f_2 = \frac{Q^2}{K^2} - \frac{Q^2}{gA^3}\int_0^h \sigma_x d\eta. \tag{6.3}$$

The crux of the work in this chapter depends on the following argument. Suppose that for some reach the function T representing channel width is arbitrarily defined. For example for a rectangular channel we would define $T = B$, where $B(x) > 0$ gives the width. If the conveyance function K is completely specified and a value for the discharge Q is given, then the functions f_1 and f_2 given by (6.2) and (6.3) are completely defined. The main part of the method is to choose a hypothetical depth profile $\hat{h}(x)$ for the reach, which at this stage we assume to be smooth. We then use the following formula to determine the bed slope for the reach:

$$S_0(x) = f_1(x, \hat{h}(x))\hat{h}'(x) + f_2(x, \hat{h}(x)). \tag{6.4}$$

It is not difficult to conclude that, for the above situation, the function $h = \hat{h}$ satisfies the differential equation (6.1) for the entire reach.

We can now use the above argument to specify a benchmark test problem for which the exact solution is known. The following information is required:

- The length of the reach L
- The width of the channel T as a function of depth and distance
- The conveyance K as a function of depth and distance
- The discharge Q
- The bed slope (which is given by (6.4))
- The appropriate boundary conditions

The required boundary conditions are found as follows: If the depth $\hat{h}(0)$ corresponds to supercritical flow, then this depth must be specified at inflow. Similarly if the depth $\hat{h}(L)$ corresponds to subcritical flow, then this depth must be specified at outflow. The above problem has as a solution $h = \hat{h}$.

Problems 1-4 in section 6.3 are examples of test problems constructed with the above technique. Note that even though the function $\hat{h}$ is assumed to be smooth we can still construct solutions with transcritical flow via a smooth transition. At any point where the depth profile smoothly crosses the critical depth, (6.4) automatically ensures that the bed slope has the critical bed slope S_{0c} (see section 2.2.2) at this point. Problem 4 in section 6.3 illustrates a smooth transition.

For many computational models the bed level z_b is required rather than the bed slope. This cannot normally be found analytically from S_0, so equation (2.8) must be integrated with a high accuracy ODE solver. For this purpose a starting value such as $z_b(L) = 0$ is required.

6.2 Test Problems with Hydraulic Jumps

The argument in the previous section requires the hypothetical depth profile $\hat{h}$ and hence the solution to be smooth since (6.4) contains the derivative of the function. This appears to prohibit solutions with hydraulic jumps. We now show how to

get round this difficulty. Suppose that the hypothetical depth profile is now only piecewise smooth, where all the discontinuities represent physical hydraulic jumps, i.e. satisfy (2.23) and (2.24). Consider a discontinuity at $x = x^\star$. Using (6.4) the bed slope is not defined at $x^\star$ and this corresponds to a discontinuity in the bed slope, i.e.

$$S_0(x^\star-) \neq S_0(x^\star+).$$

This is not a great difficulty since this yields a perfectly realistic bed profile and one may go ahead and use this as test problem. However we feel that it is worthwhile taking further steps to improve the quality of the bed slope.

In general for a problem where a hydraulic jump is triggered by a bed slope discontinuity, the position of the hydraulic jump will not coincide exactly with the position of the discontinuity in the bed slope. Hence to add realism we take steps to ensure that the position of the jump does not correspond to a bed slope discontinuity. To achieve this requires the following to hold.

$$\begin{aligned} S_0(x^\star-) &= f_1(x^\star, \hat{h}(x^\star-))\hat{h}'(x^\star-) + f_2(x^\star, \hat{h}(x^\star-)) \\ &= f_1(x^\star, \hat{h}(x^\star+))\hat{h}'(x^\star+) + f_2(x^\star, \hat{h}(x^\star+)) \quad = \quad S_0(x^\star+). \end{aligned} \tag{6.5}$$

The procedure we use is to specify $\hat{h}$ for $x \leq x^*$ which then gives $\hat{h}(x^\star-)$ and $\hat{h}'(x^\star-)$. The jump condition (2.23) determines the value of $\hat{h}(x^\star+)$ and the linear equation (6.5) then determines $\hat{h}'(x^\star+)$. The hypothetical depth downstream of the jump is now chosen to satisfy these values. For most cross-section shapes equation (2.23) must be solved numerically. In practice it is often found that the quality of the bed slope and also the solution increases with the smoothness of the bed slope at the jump. Provided the necessary derivatives exist, equation (6.4) can be differentiated a number of times to find a condition ensuring the continuity of any order derivative of the bed slope. If the bed slope is to be M times differentiable then this fixes the values of $\hat{h}(x^\star+), \hat{h}'(x^\star+), \hat{h}''(x^\star+), \ldots, \hat{h}^{(M+1)}(x^\star+)$. If we compute these values in order, then other than solving the jump condition, only linear relationships must be solved at each step. One reason to require the bed slope to be smoother than just continuous is in order to satisfy the theory of Chapter 4, which requires the bed slope to be continuously differentiable everywhere. The examples given in section 6.3

and Appendix B satisfy this at minimum and in many cases are constructed so as to have even greater smoothness.

To construct a solution for the entire reach (with possibly multiple hydraulic jumps) we work downstream from the inflow boundary. There are very few constraints on the behaviour of the hypothetical depth profile up until the position of the first hydraulic jump. Obviously the depth must be positive. In this work we also require the resulting bed slope to be everywhere positive (satisfying the conditions of the theory in Chapter 4). Downstream of a jump the situation become significantly more difficult. Suppose x^* is the position of the jump and x^{**} denotes the position of the next jump downstream (or end of reach if there are no more jumps). We require the depth profile to satisfy the following:

- Have the required values for the derivatives at the jump
- Be positive for $x^\star \leq x \leq x^{\star\star}$
- Yield a positive bed slope for $x^\star \leq x \leq x^{\star\star}$

In addition to this we must be able to obtain a solution with the desired features, e.g. maxima, minima, critical sections etc. The strategy we use is to choose the functional form of the depth downstream of the jump, allowing free parameters. Two sets of free parameters are required. The values of the first set are determined from the constraints on the depth profile at the jump. The remaining set of parameters are used to control the behaviour of the depth profile. Choosing the values for the second set of parameters is essentially a matter of trial and error, however the initial choice of the functional form is found to be crucial.

The most obvious functional form of the depth profile is a polynomial in x. For such a form it is trivial to satisfy the constraints at the jump and we took this route in [37]. The downstream depth gradient at the jump is often required to be relatively large. This results in the polynomial having a large amplitude oscillation and it is difficult to prevent the depth profile from becoming negative. The problem becomes more severe as the distance the profile is required to cover increases. Because of these difficulties, in [38], [39] and [40] we chose to use a series of exponential functions. These have the advantage that the high derivatives that may be required downstream

of the jump can be restricted to the locality of the jump. The exact functional form still made it difficult to control the solution away from the jump. The examples in this thesis still use exponential functions, however the exact form allows more systematic control over the solution. The form can be written as

$$\hat{h}(x) = \exp\left(-p(x - x^\star)\right) \sum_{i=0}^{M} k_i \left(\frac{x - x^\star}{x^{\star\star} - x^\star}\right)^i + \phi(x). \tag{6.6}$$

The parameters $k_0, k_1, \ldots, k_M$ are used satisfy the constraints at the jump. Calculating these values only involves solving a small linear system. The positive parameter p influences the rate at which the high derivatives and curvature near the jump decay and essentially controls how local the effect of the jump is. The behaviour of the solution away from the jump is controlled by the function ϕ, since $\hat{h}(x) \sim \phi(x)$ as $x - x^\star$ becomes large. Problems 5-8 in section 6.3 are examples of test problems constructed using the above method.

6.3 Test Problems for Prismatic Channels

We now give eight examples of test problems for prismatic channels (details of six more examples for non-prismatic channels are also given in Appendix B). These are used throughout the remainder of this thesis to evaluate the performance of the various numerical methods considered. Many other examples of test problems can be found in our previous work, for example see [38], [39] and [40].

All the channels have a cross-section of the form (2.32) and Manning's friction law (see section 2.1.3) is used. The conditions of the theory in section 4.9 are then satisfied so long as the bed slope is continuously differentiable and positive. The test cases here are all constructed in order to satisfy this. Table 6.1 gives all the required information (except the bed slope) for the eight cases. To recap the notation: B and Z are bottom width and side slope of the channel, respectively, n is the Manning friction coefficient, Q is the discharge and h_{in} and h_{out} are the depths to be specified at inflow and outflow, respectively (if any). In each case the bed slope is given by

$$S_0(x) = \left(1 - \frac{Q^2\left(B + 2Z\hat{h}(x)\right)}{9.08665\left(\hat{h}(x)\right)^3\left(B + Z\hat{h}(x)\right)^3}\right)\hat{h}'(x) + \frac{Q^2 n^2\left(B + 2\hat{h}(x)\sqrt{1+Z^2}\right)^{4/3}}{\left(\hat{h}(x)\right)^{10/3}\left(B + Z\hat{h}(x)\right)^{10/3}},$$

Problem	B/m	Z	L/m	n	$Q/(\mathrm{m^3 s^{-1}})$	h_{in}/m	h_{out}/m
1	10	0	150	0.03	20		0.800054
2	10	2	300	0.03	20		0.710000
3	10	2	200	0.03	20	0.400013	
4	10	2	200	0.03	20		
5	0	10	100	0.03	20	0.700000	1.900000
6	10	0	150	0.03	20		1.700225
7	5	5	200	0.03	20	0.750000	
8	5	5	650	0.03	20	0.850000	

Table 6.1: Information for test problems 1-8

where it only remains to specify the function $\hat{h}$, which is also the solution of the problem.

Problem 1 (subcritical flow) In this case we have

$$\hat{h}(x) = 0.8 + 0.25 \exp\left(-33.75\left(\frac{x}{150} - \frac{1}{2}\right)^2\right).$$

Figure 6.1(a) shows $\hat{h}$, Figure 6.1(b) shows the corresponding bed slope and Figure 6.1(c) shows the bed level and the free surface elevation. The channel flattens as we approach the mid-point of the reach, having the least gradient at this point. The channel then steepens again, returning to the initial gradient. The solution of this problem corresponds to entirely subcritical flow. The depth rises to a maximum at the center of the reach and approaches the critical depth at both ends.

Problem 2 (subcritical flow) In this case we have

$$\hat{h}(x) = 0.71 + 0.25 \sin^2\left(\frac{3\pi x}{300}\right)$$

and the problem is illustrated by Figure 6.2. The gradient of the channel flattens and then steepens again three times. As in the previous case the solution corresponds to entirely subcritical flow. The depth has local maxima corresponding to each local minima of the bed slope.

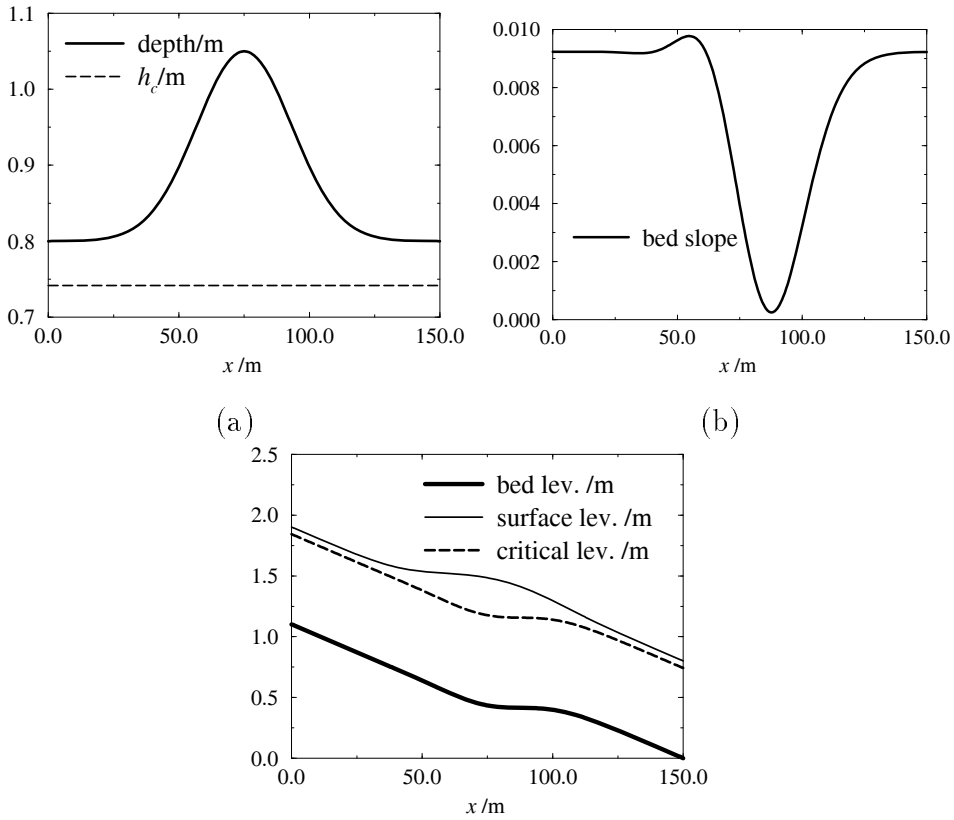


Figure 6.1: Depth, bed slope, bed level and surface level for test problem 1

Problem 3 (supercritical flow) In this case

$$\hat{h}(x) = 0.4 + 0.29 \exp\left(-40\left(\frac{x}{200} - \frac{1}{2}\right)^2\right)$$

and the problem is illustrated by Figure 6.3. Again the channel becomes flatter and then steepens. In this case the flow is entirely supercritical and the depth rises to a maximum at the middle of the reach.

Problem 4 (a smooth transition) In this case we have

$$\hat{h}(x) = 0.706033 - 0.25\tanh\left(\frac{x - 100}{50}\right)$$

and the problem is illustrated by Figure 6.4. The gradient of the channel steadily increases along the length of the reach. The flow is subcritical for the first half of the reach and supercritical for the second half. The transition between the two regimes is via a smooth transition.

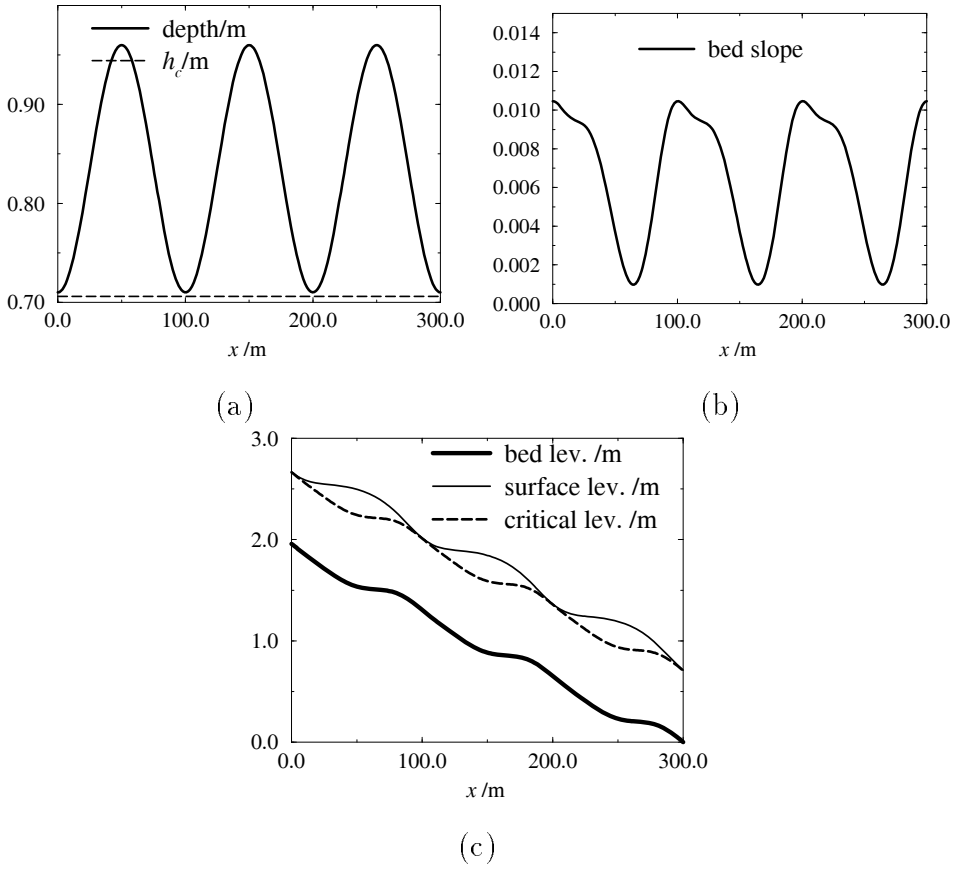


Figure 6.2: Depth, bed slope, bed level and surface level for test problem 2

Problem 5 (a hydraulic jump) In this case for $x \leq 50\text{m}$

$$\hat{h}(x) = 0.7,$$

while for $50\text{m} < x \leq 100\text{m}$ the depth is of the form (6.6) with $x^* = 50\text{m}$, $x^{**} = 100\text{m}$, $M = 4$, $k_0 = 1.279$, $k_1 = 0.0771155$, $k_2 = -0.0062375$, $k_3 = 0.00391708$, $k_4 = -0.00368501$, $p = 0.5$ and

$$\phi(x) = 1.9 \exp\left(0.0005(x - 100)\right).$$

This problem is illustrated in Figure 6.5. For the first half of the reach, the channel has constant bed slope. After the mid-point the channel flattens out steadily. The solution to this problem changes from supercritical to subcritical via a hydraulic jump at the mid-point which is triggered by the reducing gradient of the channel causing the flow to decelerate.

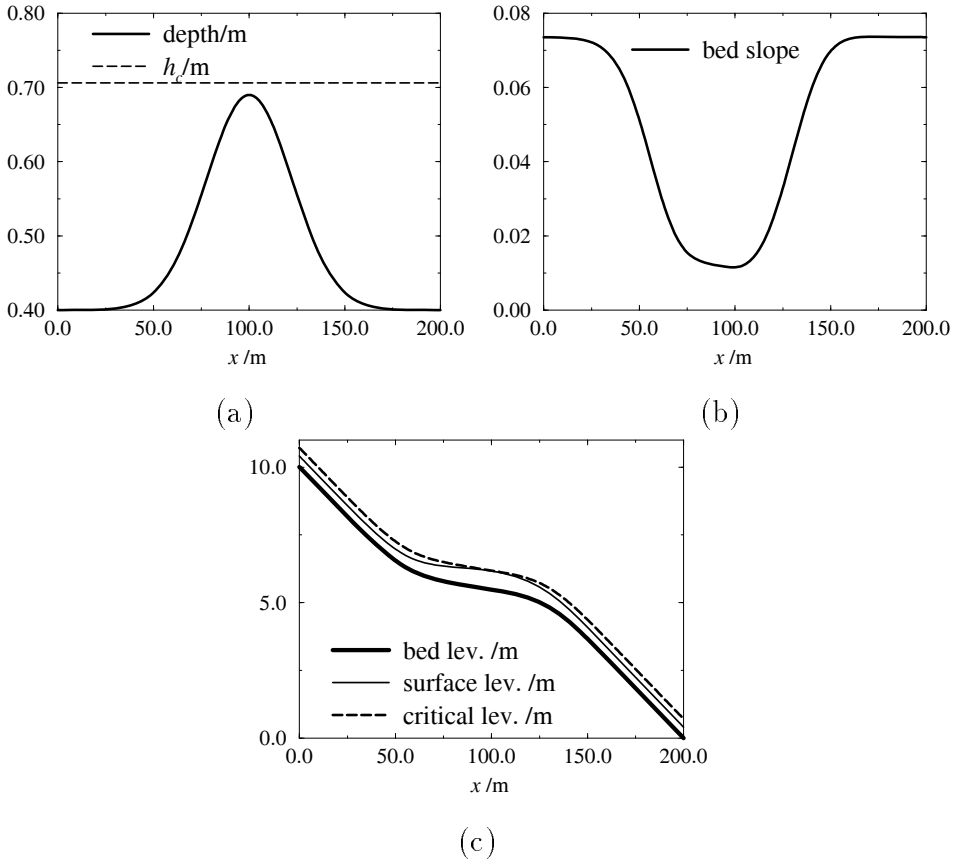


Figure 6.3: Depth, bed slope, bed level and surface level for test problem 3

Problem 6 (a smooth transition followed by a jump) In this case for $x \leq 100\text{m}$

$$\hat{h}(x) = 0.741617 - \frac{0.25}{\tanh(3)}\tanh\left(3\frac{(x-50)}{50}\right),$$

while for $100\text{m} < x \leq 200\text{m}$ the depth is of the form (6.6) with $x^* = 100\text{m}$, $x^{**} = 200\text{m}$, $M = 4$, $k_0 = 1.0656$, $k_1 = 0.0604859$, $k_2 = -0.00423834$, $k_3 = 0.00198394$, $k_4 = -0.00144967$, $p = 0.3$ and

$$\phi(x) = 1.7\exp\left(0.005(x-200)\right).$$

This problem is illustrated in Figure 6.6. The channel steepens and the flattens out again. The solution changes smoothly from subcritical flow to supercritical flow at one third distance, and then returns via a hydraulic jump to subcritical flow at two thirds distance.

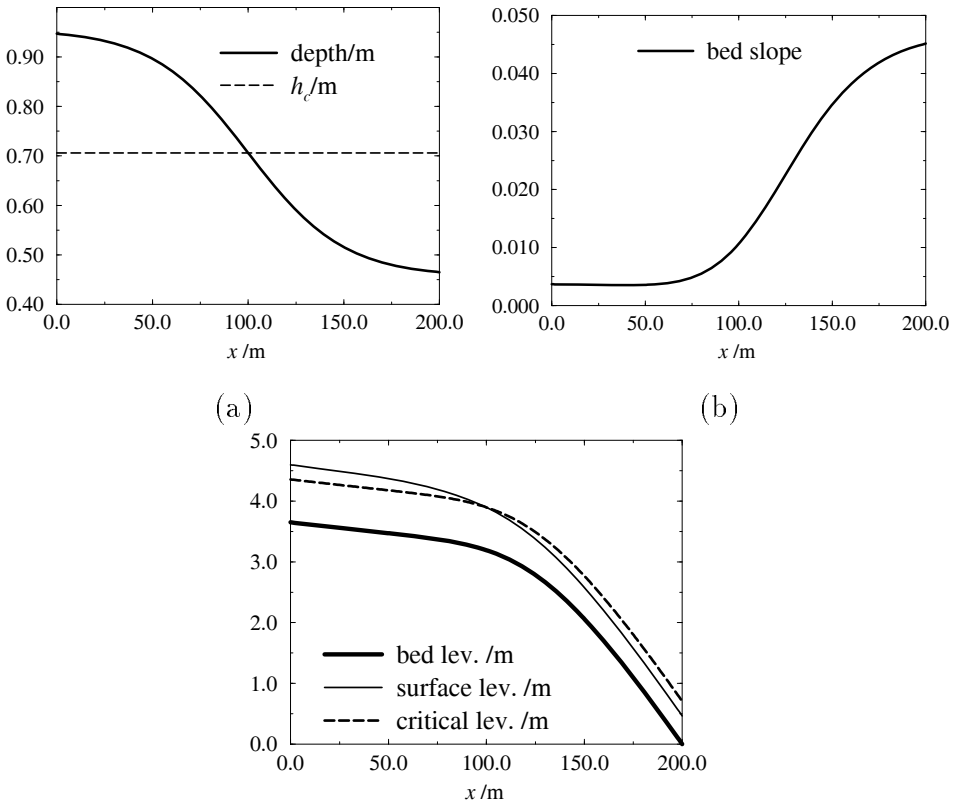


Figure 6.4: Depth, bed slope, bed level and surface level for test problem 4

Problem 7 (a jump followed by a smooth transition) In this case for $x \leq 50\text{m}$

$$\hat{h}(x) = 0.75,$$

while for $50\text{m} < x \leq 200\text{m}$ the hypothetical depth is of the form (6.6) with $x^* = 50\text{m}$, $x^{\star\star} = 200\text{m}$, $M = 4$, $k_0 = 1.01906$, $k_1 = 0.0330202$, $k_2 = -0.00362889$, $k_3 = 0.0021294$, $k_4 = -0.002026$, $p = 0.5$ and

$$\phi(x) = 1.0 - 0.25\tanh(0.03(x - 130)).$$

This problem is the opposite of problem 6 and is shown in Figure 6.7. The channel flattens out and then steepens again. The solution changes from supercritical flow to subcritical flow via a hydraulic jump at one quarter distance, and then returns smoothly to supercritical flow at roughly two thirds distance.

Problem 8 (many transitions) In this case for $x \leq 50\text{m}$

$$\hat{h}(x) = 0.75 + 0.1\exp(-0.1x).$$

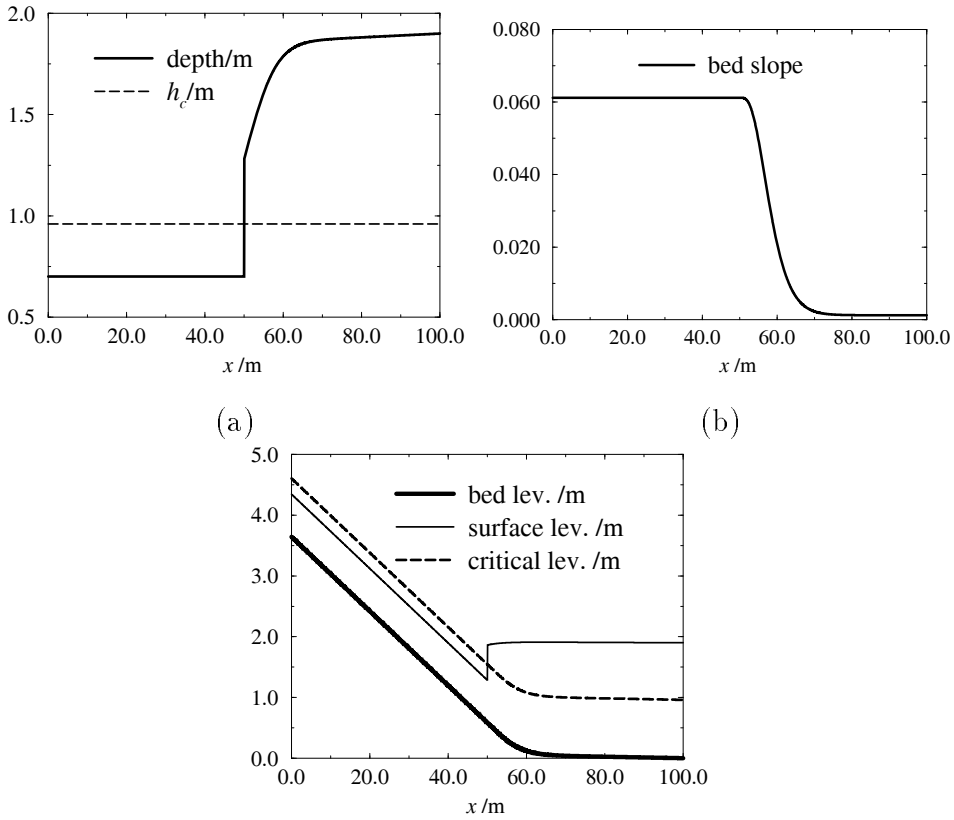


Figure 6.5: Depth, bed slope, bed level and surface level for test problem 5

For 50m $< x \leq$ 200m the depth is of the form (6.6) with $x^* = 50$m, $x^{**} = 200$m, $M = 3$, $k_0 = 1.01826$, $k_1 = 0.0330684$, $k_2 = -0.00366582$, $k_3 = 0.00216754$, $p = 0.5$ and

$$\phi(x) = 1.0 - 0.25\tanh(0.03(x - 130)).$$

For 200m $< x \leq$ 350m the depth is of the form (6.6) with $x^* = 200$m, $x^{**} = 350$m, $M = 3$, $k_0 = 1.01031$, $k_1 = 0.032978$, $k_2 = -0.00387227$, $k_3 = 0.00243084$, $p = 0.5$ and

$$\phi(x) = 1.0 - 0.23\tanh(0.03(x - 280)).$$

For 350m $< x \leq$ 500m the depth is of the form (6.6) with $x^* = 350$m, $x^{**} = 500$m, $M = 3$, $k_0 = 0.987768$, $k_1 = 0.0308504$, $k_2 = -0.00416726$, $k_3 = 0.00294851$, $p = 0.5$ and

$$\phi(x) = 0.98 - 0.20\tanh(0.03(x - 430)).$$

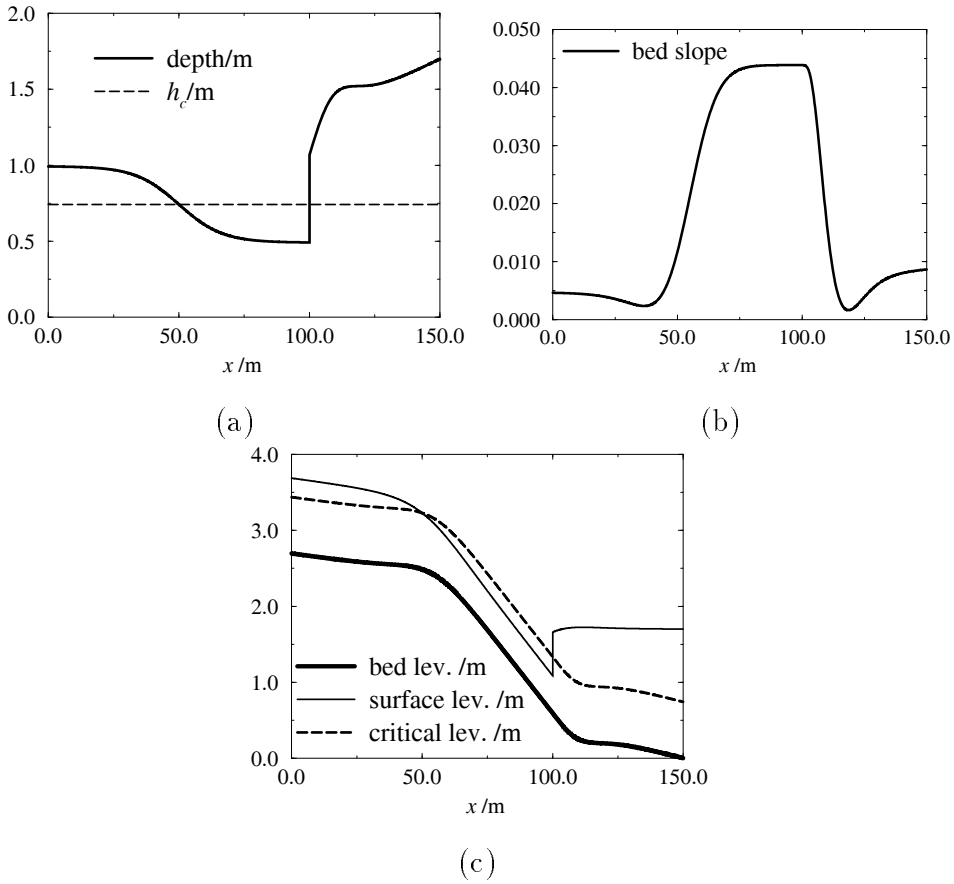


Figure 6.6: Depth, bed slope, bed level and surface level for test problem 6

For 500m $< x \leq$ 650m the depth is of the form (6.6) with $x^* = 500$m, $x^{**} = 650$m, $M = 3$, $k_0 = 0.977391$, $k_1 = 0.029882$, $k_2 = -0.00435568$, $k_3 = 0.00329588$, $p = 0.5$ and

$$\phi(x) = 0.96 - 0.20\tanh(0.03(x - 580)).$$

The solution to this problem (shown in Figure 6.8) has altogether eight transitions, corresponding to four hydraulic jumps and four smooth transitions. The flow is supercritical at both inflow and outflow.

6.4 Conclusions

In this chapter we have given a method for constructing steady open channel test problems to which the exact solution of the steady Saint-Venant equation is known. To the author's knowledge this is the first time that non-trivial exact solutions have been made available to the modeller. Moreover, the method can create a useful

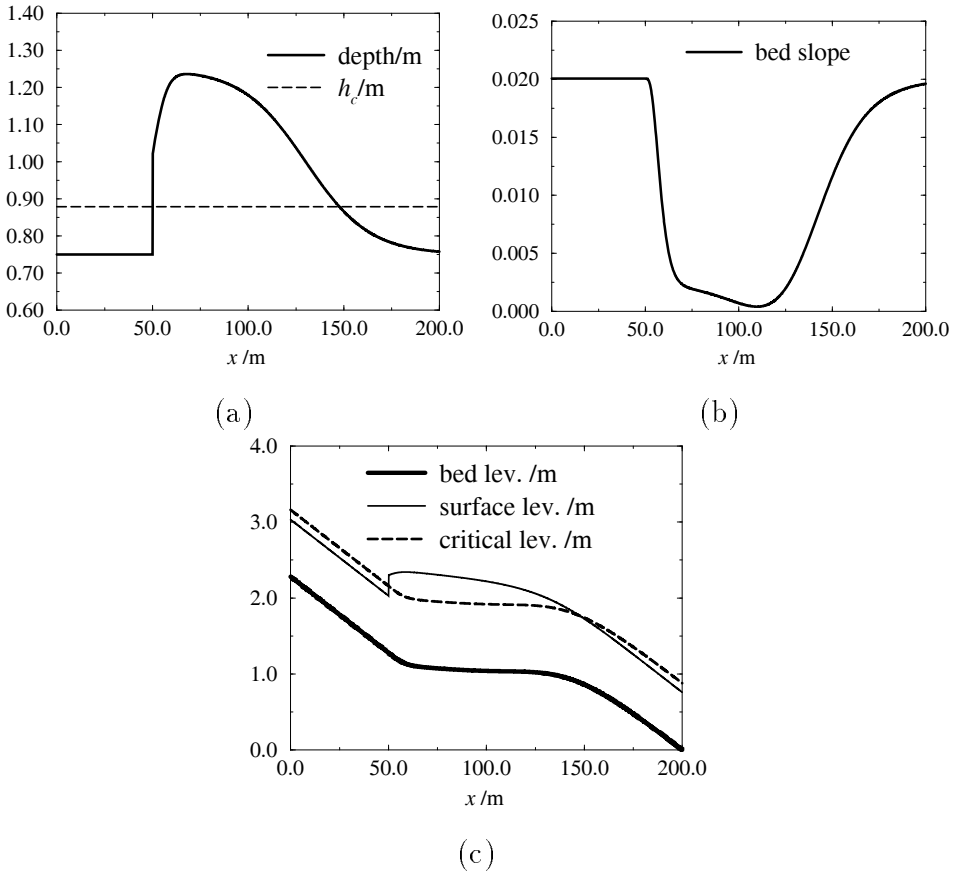


Figure 6.7: Depth, bed slope, bed level and surface level for test problem 7

range of test problems, including almost all channel geometries and all flow types. In particular, techniques for constructing problems with hydraulic jumps have been described and it has been shown that jumps must satisfy certain conditions depending on how smooth the bed slope is required to be. For brevity, the test examples given are restricted to rectangular, triangular and trapezoidal cross-sections. This is not a restriction on the method, although for complicated channel shapes the expressions for the bed slope become even more large and unwieldy. The symbolic computation package Mathematica (see [69]) helped greatly to facilitate the algebraic construction of the test problems. We consider the method described as a valuable tool for developing, validating or comparing steady open-channel solvers. The method can also be used to test the performance of unsteady models as the solution tends to a steady state.

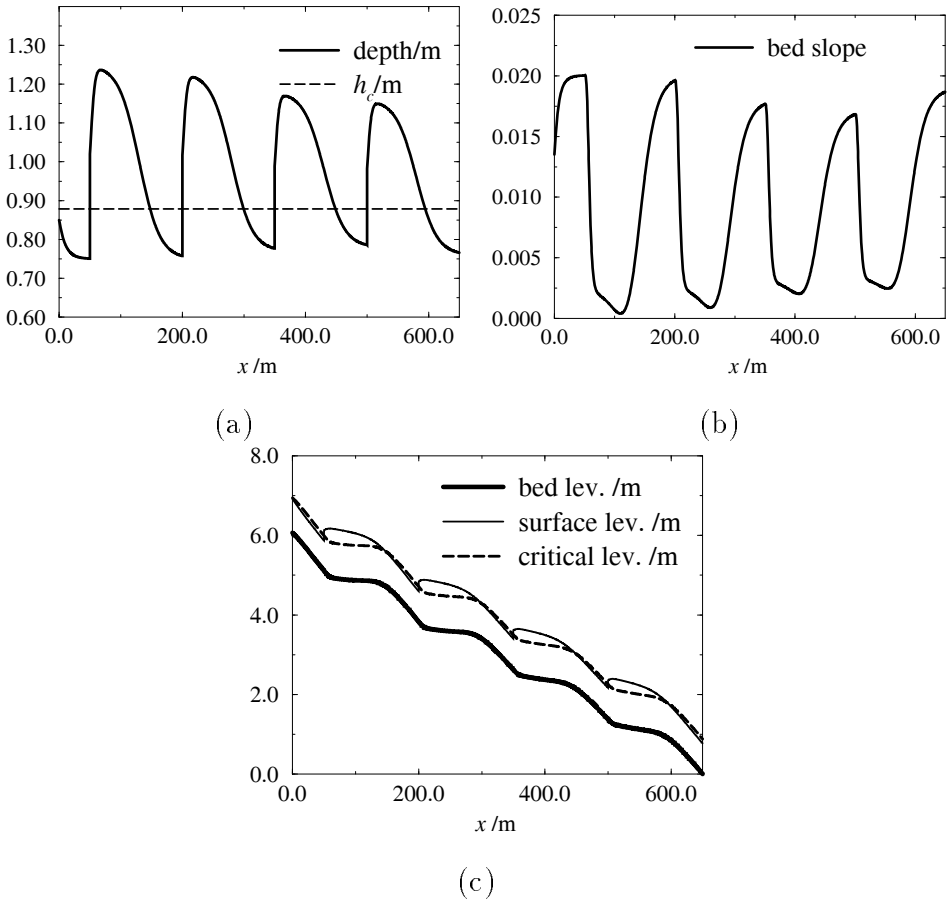


Figure 6.8: Depth, bed slope, bed level and surface level for test problem 8

Chapter 7

Numerical Experiments

In Chapter 5 we presented theory for a family of numerical methods for computing steady solutions to the Saint-Venant equations. We refer to this approach as the "scalar approach" since we proceed as for computing the steady solution of a scalar partial differential equation. We present results for this approach for some of the benchmark test problems described in the previous chapter. The usefulness of the theory is assessed in terms of the quality of the a-priori estimates for the bounds on the solution, the time step sufficient for convergence of the time-stepping iteration and the bound on the convergence rate for this iteration. The accuracy of the various scalar schemes are then compared against a well known time accurate solver. This being the approximate Riemann solver of Roe[55]. We find that for a certain discretisation of the source term, Roe's scheme gives second order accuracy at steady state. We use this idea to also obtain second order accuracy for the scalar approach whilst retaining a three-point scheme. We then consider more traditional methods of obtaining high order accuracy, namely the high order TVD approach discussed in section 3.5.

7.1 Application of some Monotone Schemes

We now apply the methods discussed in Chapter 5 to the test problems 1-8 given in the previous chapter. The schemes we consider are the monotone schemes of Engquist-Osher, Godunov and Lax-Friedrichs. We also consider the first-order up-

wind scheme which is not a monotone scheme, but nevertheless is still well behaved. We use the strategy described in section 5.5 to compute a-priori bounds for the solution, and allowable time steps for the time stepping iteration.

Problem 1 Consider the test problem 1. It is given that the flow at inflow is subcritical and that the depth at outflow is $\hat{h}(L) \approx 0.80$m. Following the strategy described in section 5.5, we take $\gamma_0 = h_c \approx 0.74$m and γ_1 to be the depth specified at outflow. This yields the bounds on the solution

$$\underline{h} \leq h(x) \leq \overline{h}, \qquad 0 \leq x \leq L,$$

where $\underline{h}$ =0.74m and $\overline{h}$ =2.64m. Comparing the bounds and the actual solution (shown in Figure 7.1) we find that the upper bound is not at all tight. The actual solution does not rise above 1.05m. In general the bounds given by the theory cannot be expected to be tight, for they depend solely on the extreme values of the bed-slope. In the current example the upper bound must therefore take into account the worst case scenario for which the bed slope is at its minimum value for a great enough distance for the solution to asymptote to the corresponding normal depth. In reality though, the bed slope (see Figure 6.1(b)) is only close to it's minimum value for a small fraction of the reach.

Figure 7.1 shows results for the Engquist-Osher scheme for problem 1 with $\Delta x =$ 10m. The Godunov and the first-order upwind schemes give identical results because the difference equations reduce to an identical form for purely subcritical or purely supercritical solutions. The numerical solution gives a reasonable representation of the solution. The numerical solution is slightly skew, whereas the exact solution is symmetric about the middle of the reach. The numerical solution also fails to reach the correct maximum depth by a few centimeters.

The initial guess for the time-stepping iteration is taken to be the linear depth profile joining the end values γ_0 and γ_1. Thus as discussed in section 5.5 the depth range of interest, and that over which the CFL condition must hold, can be taken as $[\alpha, \beta] = [\underline{h}, \overline{h}]$. It is found that the time stepping iteration converges for significantly higher time steps than that predicted by the theory. For example for the Engquist-Osher scheme with Δx =10m, the CFL condition in Theorem 10 is equivalent to

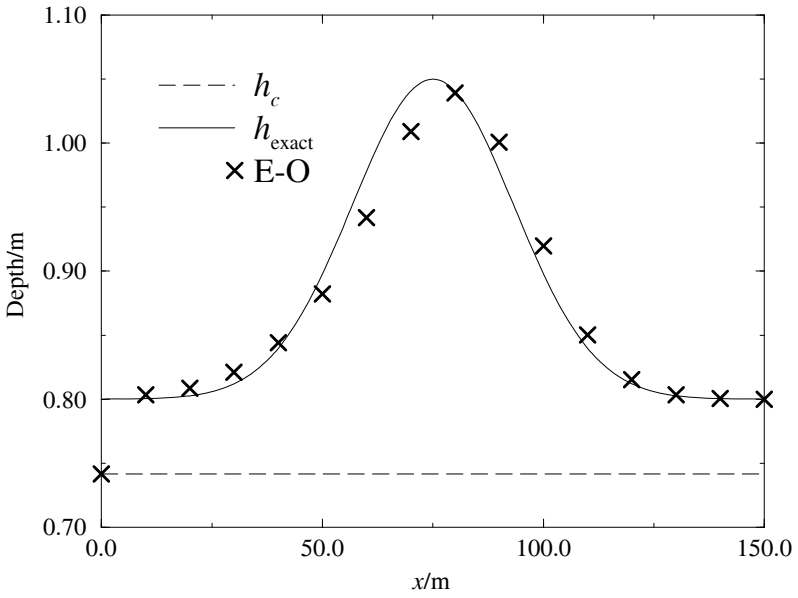


Figure 7.1: Problem 1 using the Engquist-Osher scheme and $\Delta x = 10$m.

the requirement that $\Delta t \leq 0.038$. However in practice the iteration is observed to converge for time steps up until around 0.235, although the optimum performance is found to be at about 0.172. Theorem 9 also gives a lower bound for the convergence rate of the iteration. We have that

$$\|\boldsymbol{h}^n - \boldsymbol{h}\| \leq \|\boldsymbol{h}^0 - \boldsymbol{h}\| \exp(-n\delta\Delta t),$$

where $\delta = 0.083$. This is found to be far too pessimistic since, for $\Delta x = 10$m and $\Delta t = 0.038$, we find that

$$\|\boldsymbol{h}^n - \boldsymbol{h}\| \approx \|\boldsymbol{h}^0 - \boldsymbol{h}\| \exp(-n\hat{\delta}\Delta t),$$

for $\hat{\delta}$ =3.6. It is already clear that above a-priori estimates may be of pretty poor quality and thus may not be of too great practical use. The main cause of this appears to be the looseness of the upper bound $\overline{h}$. This has the consequence that the CFL condition is required to hold over a far too large a range of depths.

Figure 7.2 shows results for the Lax-Friedrichs scheme with $\Delta x = 10$m for a selection of values of the parameter λ. Clearly the lower the value of this parameter, the more diffusive the scheme is. For the scheme to be monotone on the depth interval of interest $[\alpha, \beta]$, it is required that $\lambda|F'(h)| \leq 1$ on this range. This in turn requires that $\lambda \leq 0.004$. Unfortunately the solution is far too diffusive for values inside this

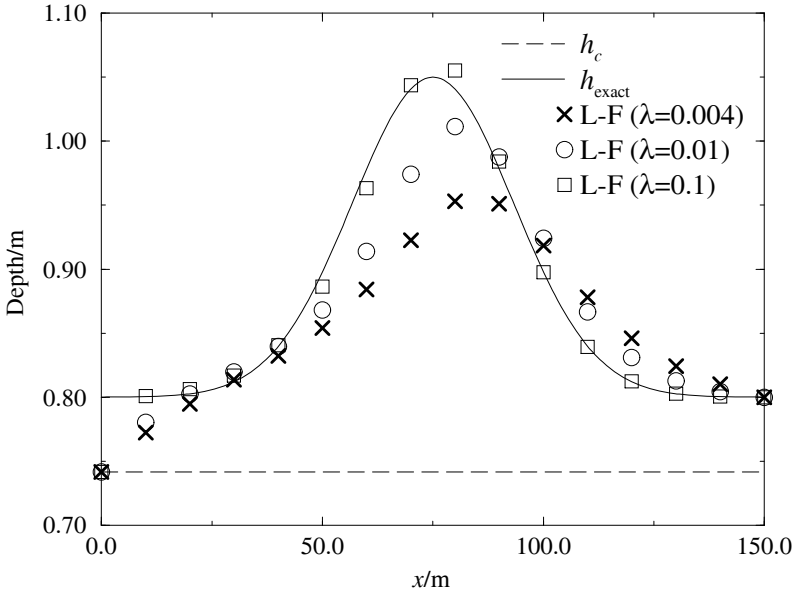


Figure 7.2: Lax-Friedrichs scheme for Problem 1 and various λ ($\Delta x = 10$m).

range. For example Figure 7.2 shows results for $\lambda = 0.004$. The solution can be improved by increasing λ above this range. The figure also shows the solution for $\lambda = 0.1$ which appears to be more accurate than the corresponding solution from the Engquist-Osher scheme. Unfortunately there is no way to estimate such a value of λ in advance. Increasing the value of λ too far (for a fixed Δt) causes the time stepping iteration to diverge.

Problem 2 As in problem 1, it is given that the flow is subcritical at inflow. The depth at outflow is given to be $\hat{h}(L)$ =0.71m. Taking $\gamma_0 = h_c \approx$0.71m and $\gamma_1 = \hat{h}(L)$ yields the lower bound $\underline{h}$ =0.70m and the upper bound $\overline{h}$ =1.41m. Figure 7.3 shows results for the Engquist-Osher scheme with $\Delta x = 10$m. Again the solution does not reach the correct extrema and the solution is shifted to the right.

Problem 3 For this problem the depth at inflow is given to be $\hat{h}(0) \approx$0.40m and the flow at outflow is given to be supercritical. In this case we take $\gamma_0 = \hat{h}(0)$ and $\gamma_1 = h_c \approx$0.71m, to yield bounds $\underline{h}$ =0.39m and $\overline{h}$ =0.71m. The bounds are much tighter for this problem than in the previous cases. Figure 7.4 shows the solution for the Engquist-Osher scheme with $\Delta x = 10$m. Again the maximum is too low, but the solution is almost symmetric in this case. The a priori estimates for this case are

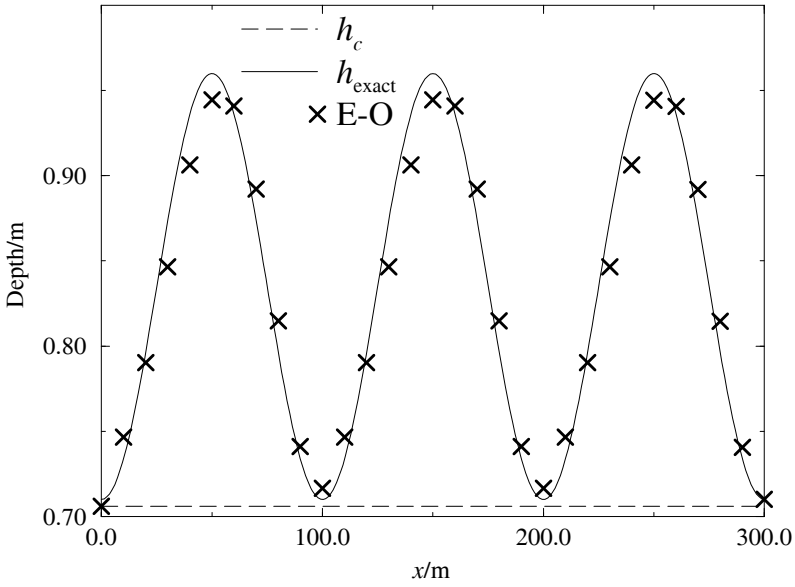


Figure 7.3: Problem 2 using the Engquist-Osher scheme and $\Delta x = 10$m.

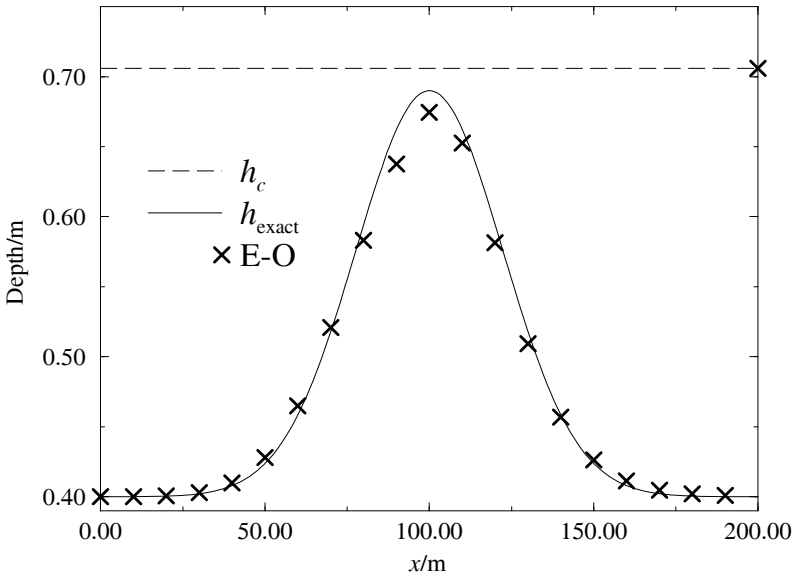


Figure 7.4: Problem 3 using the Engquist-Osher scheme and $\Delta x = 10$m.

found to be of much better quality than in the previous cases, and this must be due to the tightness of the given solution bounds.

Problem 4 For this problem it is given that the flow is subcritical at inflow and supercritical at outflow, hence we take $\gamma_0 = \gamma_1 = h_c \approx 0.71$m. This yields the bounds $\underline{h} = 0.46$m and $\overline{h} = 0.98$m. Figure 7.5 shows results for the Engquist-Osher, Godunov and first-order upwind schemes with $\Delta x = 40$m. The exact solution to this problem

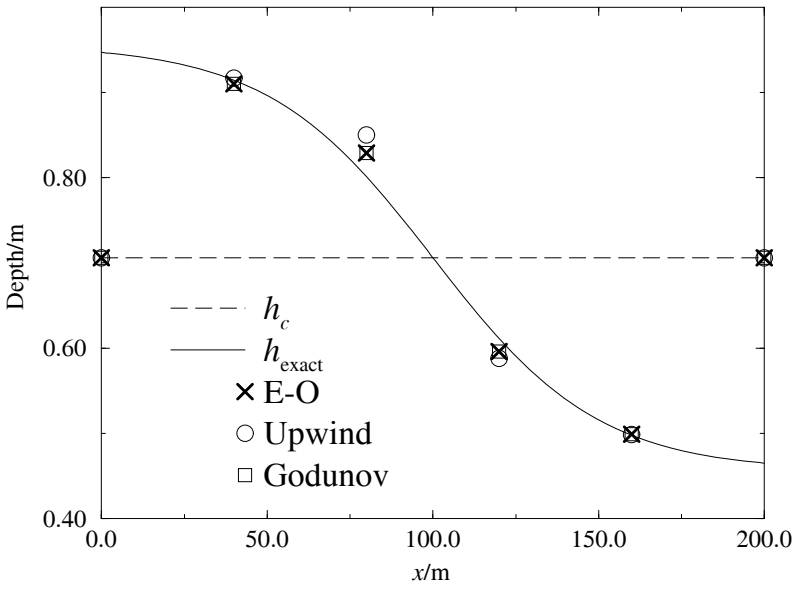


Figure 7.5: Comparison of E-O, Godunov and upwind schemes for problem 4 (Δx =40m).

changes smoothly from subcritical to supercritical at the midpoint of the channel. The methods give a good representation of the solution even with so few grid points. The upwind scheme is found to be less accurate than for the Engquist-Osher or Godunov schemes at the grid points on either side of the transition. The smooth transition is a steady expansion wave for the scalar equation (4.6). The fact that the first-order upwind scheme is less accurate near the transition stems from the fact that the numerical flux assumes any transition is a shock. The Godunov and Engquist-Osher numerical fluxes, however, are designed to correctly recognise an expansion wave. In fact the Engquist-Osher and Godunov schemes have identical solutions for this type of solution. This can be seen by comparing (5.35) and (5.36) for the case across a smooth transition

Problem 5 In this case the depth is $\hat{h}(0)$ =0.70m at inflow and $\hat{h}(L)$ =1.9m at outflow. Hence we take $\gamma_0 = \hat{h}(0)$ and $\gamma_1 = \hat{h}(L)$ which gives $\underline{h}$ =0.70m and $\overline{h}$ =1.90m. Figure 7.6 shows results for the Engquist-Osher, Godunov and first-order upwind schemes with Δx =5m. The exact solution to this problem has a hydraulic jump at the midpoint of the channel and all three methods capture this jump satisfactorily. In fact the jump is captured even with only a few grid points. It

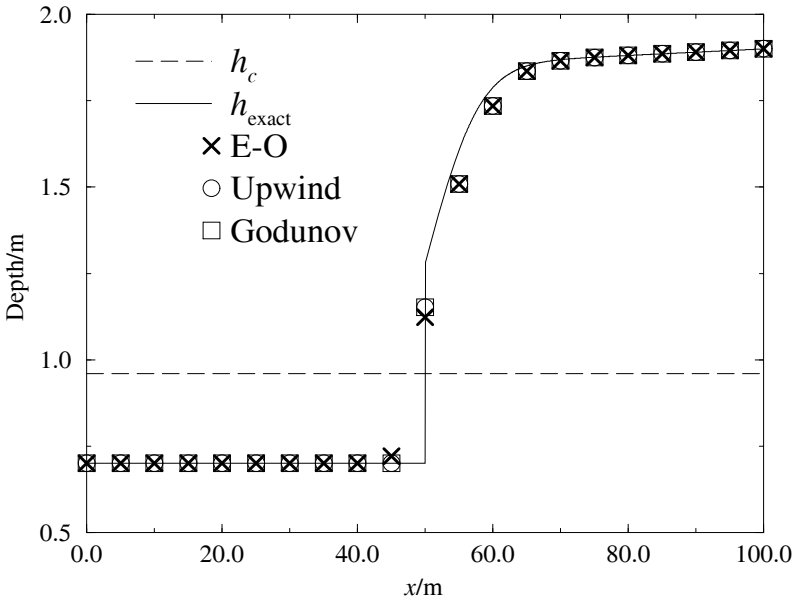


Figure 7.6: Comparison of E-O, Godunov and upwind schemes for problem 5 (Δx =5m).

can be seen that the Godunov scheme and the first-order upwind scheme give slightly better resolved jumps than does the Engquist-Osher scheme. In particular at the grid point immediately before the jump, the Engquist-Osher solution has visibly moved away from the constant state, whereas the other schemes have not. It is well known that the Engquist-Osher scheme smears discontinuities to a greater extent than the other schemes. The upwind and Godunov methods have identical solutions for this type of solution (compare (5.37) and (5.36) for the case across a jump).

The Lax-Friedrichs scheme requires $\lambda \leq 0.0029$, in order to be monotone over the range of depths of interest. Figure 7.7 shows results for $\lambda = 0.0029$. The results for values inside the monotone range, as in problem 1, are far too diffusive relative to the other schemes. The results can be improved by taking values above this range. For example Figure 7.7 also shows results for $\lambda = 0.01$. Increasing λ further results in divergence of the iteration for a fixed Δt.

Problem 6 For this problem it is given that the flow is subcritical at inflow and has depth $\hat{h}(L) \approx$1.70m at outflow. Taking $\gamma_0 = h_c \approx$0.74m and $\gamma_1 = \hat{h}(L)$ gives the bounds $\underline{h}$ =0.49m and $\overline{h}$ =1.70m. Figure 7.8 shows results for the Engquist-Osher, Godunov and first-order upwind schemes with Δx =6m. The solution to this problem

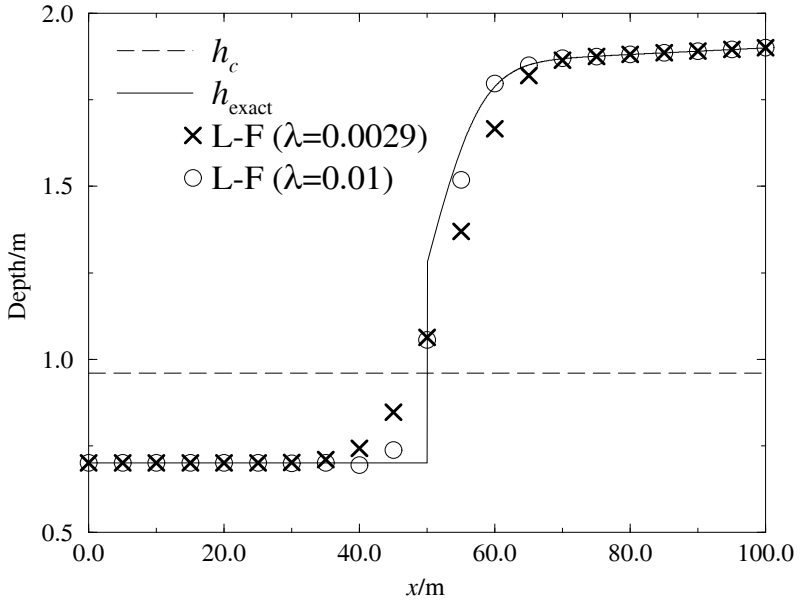


Figure 7.7: Lax-Friedrichs scheme for Problem 5 for various λ (Δx =5m).

changes smoothly from subcritical to supercritical at one third distance and back to subcritical at two thirds distance via a hydraulic jump. As in the previous examples the Engquist-Osher and Godunov schemes are more accurate than the first-order upwind scheme near the smooth transition. On the other hand the Godunov and the first-order upwind schemes give a sharper jump than the Engquist-Osher scheme.

Problem 7 For this problem it is given that the depth is $\hat{h}(0)$ =0.75m at inflow and the flow is supercritical at outflow. Hence we can take $\gamma_0 = \hat{h}(0)$ and $\gamma_1 = h_c \approx$0.88m, to give the bounds $\underline{h}$ =0.75m and $\overline{h}$ =1.9m. Figure 7.9 shows results for the Engquist-Osher, Godunov and first-order upwind schemes for Δx =8m. The solution to this problem changes from supercritical to subcritical via a hydraulic jump at one third distance and returns smoothly back to supercritical at two thirds distance. Again the first-order upwind scheme is least accurate near the smooth transition and the Engquist-Osher scheme gives the most smeared jump.

Problem 8 For this problem it is given that the depth is $\hat{h}(L)$ =0.75m at inflow and that the flow is supercritical at outflow. Taking $\gamma_0 = \hat{h}(0)$ and $\gamma_1 = h_c \approx$0.88m yields the bounds $\underline{h}$ =0.75m and $\overline{h}$ =1.9m. Figure 7.10 shows results for the Engquist-Osher scheme with Δx =10m. The solution to this problem has eight transitions

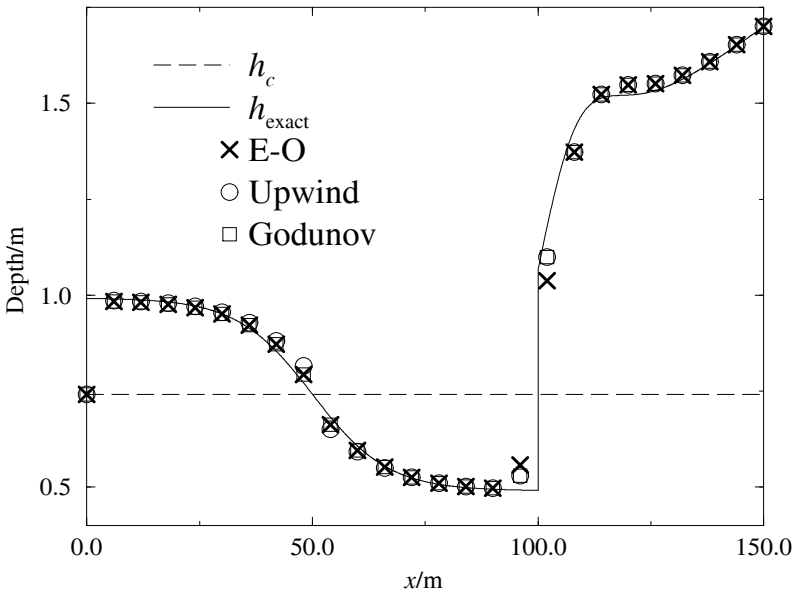


Figure 7.8: Comparison of E-O, Godunov and upwind schemes for problem 6 (Δx =6m).

(4 hydraulic jumps and 4 smooth transitions). The solution demonstrates that the scheme is successful at solving multiple transition problems.

From the eight test problems we conclude that the time steps required to satisfy Theorem 10 and hence guarantee convergence of the time stepping iteration are in general too pessimistic to be of great practical use. A more effective approach may be to allow a variable time step which is chosen to satisfy a CFL condition at each particular time level.

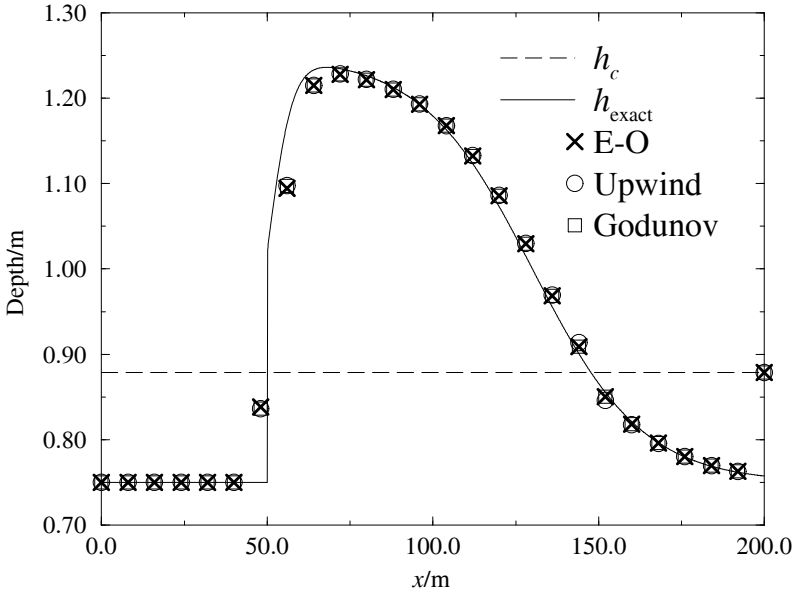


Figure 7.9: Comparison of E-O, Godunov and upwind schemes for problem 7 (Δx =6m).

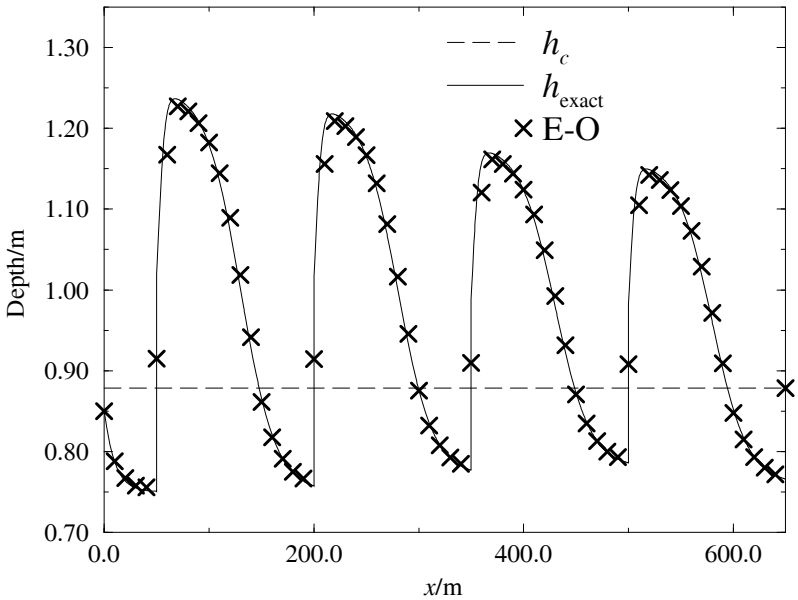


Figure 7.10: Results for the Engquist-Osher scheme for problem 8 (Δx =10m).

7.2 Comparison with Roe's Approximate Riemann Solver

We now compare the accuracy of the schemes in the previous section against that of Roe's approximate Riemann solver[55] which is described in sections 3.3 and 3.8. The latter scheme is a time accurate solver of the time dependent Saint-Venant system and we model the transient flow until steady state is attained. The scheme is a natural generalisation of the first-order upwind scheme to systems of equations and is designed specifically for the computation of discontinuous flows. In section 3.7 we discussed two different methods of discretising a source term. Here we apply both the pointwise discretisation and the upwind discretisation. To average the source term at an interface we use (3.20). The upwind discretisation is motivated by the fact that, for the pointwise discretisation, the resulting discharge varies wildly from the expected constant discharge. Upwinding the source term is found to not only remedy this but, for the particular choice of source term averaging, have the side effect of giving second order accuracy at the steady state.

Figure 7.11 shows results for Roe's scheme for both the pointwise and upwind discretisation of the source term for problem 1 ($\Delta x = 10m$). The corresponding results for the Engquist-Osher scheme are also shown. Part (a) of the figure shows the depth field. Roe's scheme with pointwise source term appears to be on average slightly more accurate than the Engquist-Osher scheme, although the latter scheme gives a more accurate maximum depth. Roe's scheme with upwinded source term, however, is clearly seen to be the most accurate method by a significant margin. Part (b) of the figure shows the discharge field. For the scheme with pointwise source term, the solution deviates by a significant amount from the expected constant discharge. It thus seems unreasonable that this method should give better (or even comparable) accuracy in the depth field to that of the Engquist-Osher scheme, when in the case of the latter scheme, the discharge is (by definition) exact. The discharge for Roe's scheme with upwinded source term is correct everywhere to within $10^{-5}\mathrm{m}^3\mathrm{s}^{-1}$, and these errors are most-likely to be only rounding errors since they do not decrease as the grid is refined. This evidence indicates that the difference equations do have an

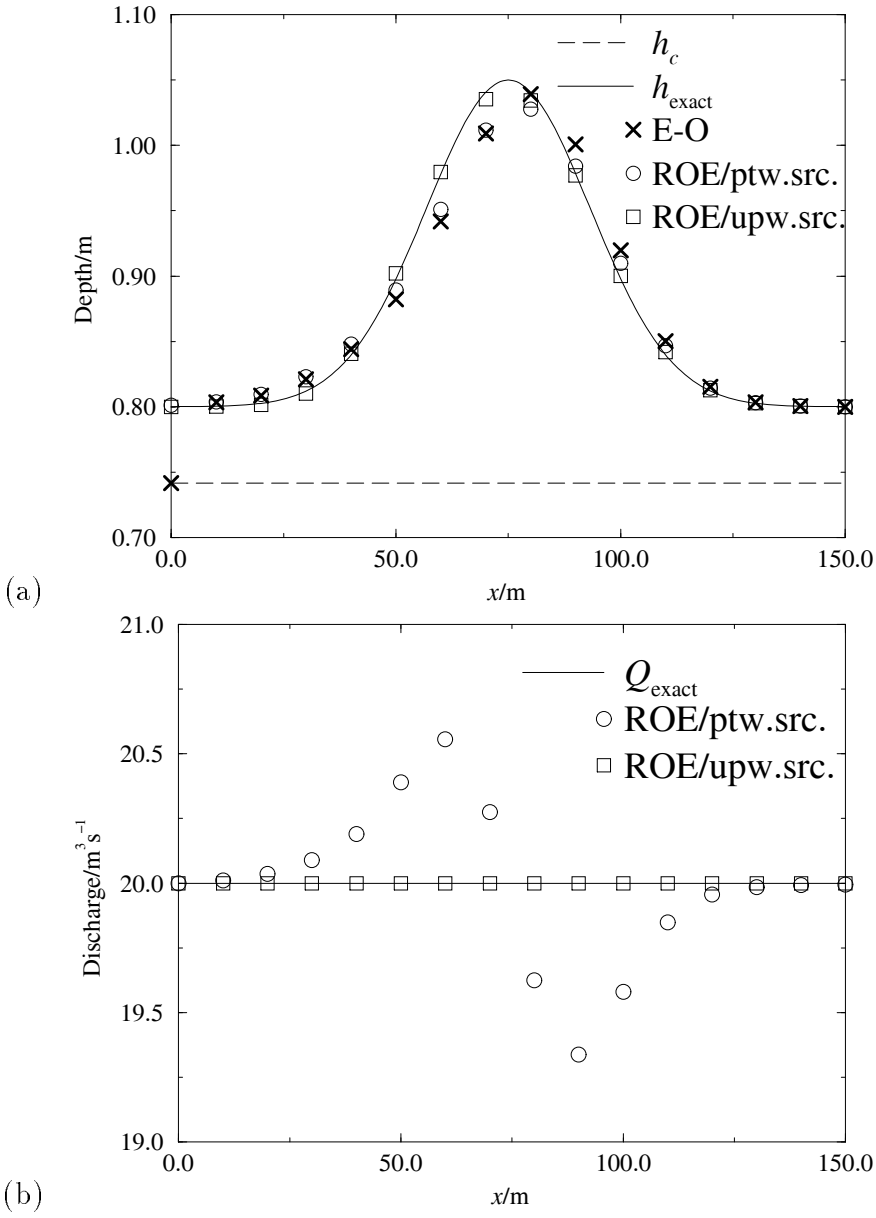


Figure 7.11: Roe's scheme for problem 1 (Δx =10m).

exact solution with $Q \equiv$constant at steady state. We show later that this is in fact the case.

We consider two different measures of the accuracy for the schemes, which are the L_1 (or mean) error given by

$$\frac{1}{N-1}\sum_{i=1}^{N-1}|\hat{h}(x_i)-h_i|$$

and the L_∞ (or maximum) error given by

$$\max_{0<i<N}|\hat{h}(x_i)-h_i|,$$

where the function $\hat{h}$ is the exact solution. In order to allow a fair comparison of the two distinct approaches, the end points of the reach are not included in the error measures. This is because the solution is not in general approximated at these points for the scalar approach, since we fix $h_0 = \gamma_0$ and $h_N = \gamma_1$, where these values do not necessarily arise from physical boundary conditions.

Figures 7.12 (a) and (b) show the L_∞ and L_1 errors as a function of the number of grid-points N. The figures confirm the observation that Roe's scheme with upwinded source term is more accurate than the version with pointwise source term, which in turn is more accurate than the Engquist-Osher scheme. Since this data is plotted using logarithmic axes, a measure of the order of accuracy of a particular scheme with respect to a particular measure is given by the negative of the slope of the straight line passing through the data points. The best straight line fit is obtained by the method of least squares and is illustrated along with its slope for each of the data sets. The Engquist-Osher scheme and Roe's scheme with pointwise source term illustrate first order accuracy in both measures as expected. Unexpectedly, however, Roe's scheme with upwinded source term demonstrates second order accuracy. We show why this is the case later.

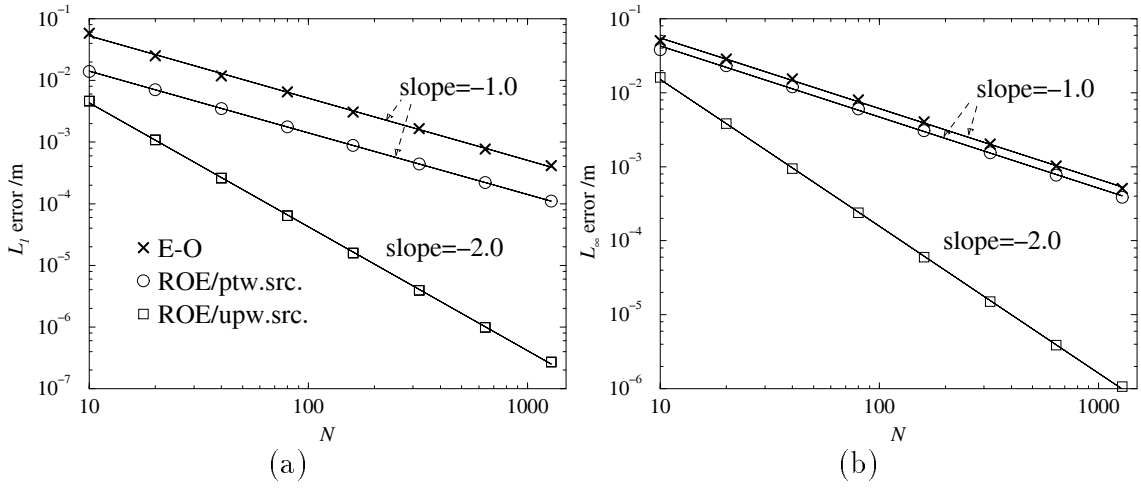


Figure 7.12: L_1 and L_∞ errors for problem 1.

Figure 7.13 compares the results for the Engquist-Osher scheme with those of Roe's scheme for problem 6 and $\Delta x = 6$m. Part (a) of the figure shows the depth field. At the grid-points on either side of the smooth transition, Roe's scheme with upwinded source term gives rise to the largest errors. Roe's scheme with upwinded

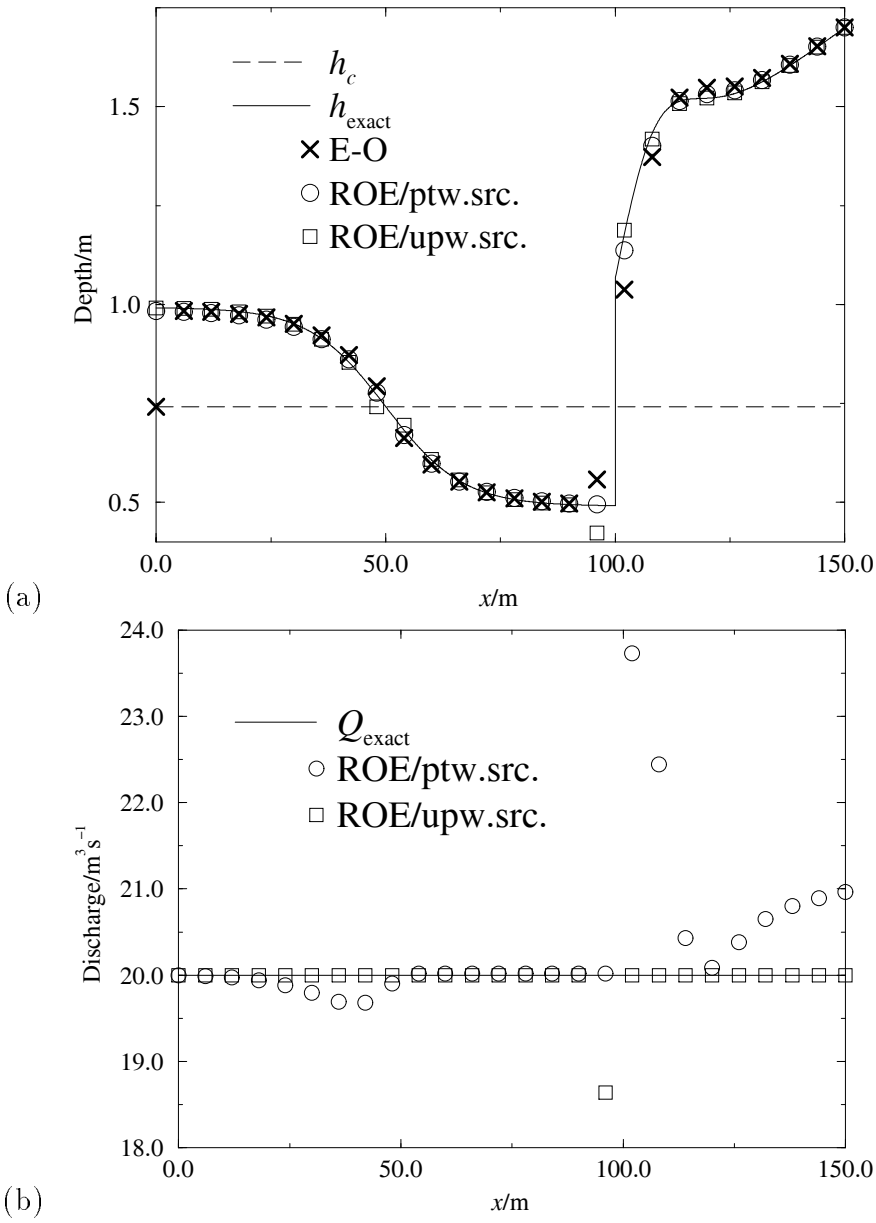


Figure 7.13: Roe's scheme for problem 6 (Δx =6m).

source term is by far the most accurate method at the grid-point immediately downstream of the jump, but the opposite is true at the grid-point immediately upstream of the jump, where the solution undershoots the jump. Such features are found to occur relatively frequently for the upwind source term discretisation.

Part (b) of the figure shows the discharge field. As in the previous test case, Roe's scheme with pointwise source term deviates considerably for much of the reach from the expected constant discharge. In the notation of section 3.8 this scheme can

be written as

$$\frac{\boldsymbol{w}_j^{n+1} - \boldsymbol{w}_j^n}{\Delta t} + \left(\tilde{J}^-_{j+\frac{1}{2}}\right)^n \frac{(\boldsymbol{w}_{j+1}^n - \boldsymbol{w}_j^n)}{\Delta x} + \left(\tilde{J}^+_{j-\frac{1}{2}}\right)^n \frac{(\boldsymbol{w}_j^n - \boldsymbol{w}_{j-1}^n)}{\Delta x} = \boldsymbol{D}_j^n,$$

where $\boldsymbol{D}_j^n = \boldsymbol{D}(x_j, \boldsymbol{w}_j)$. At steady state this reduces to

$$\tilde{J}^-_{j+\frac{1}{2}} \frac{(\boldsymbol{w}_{j+1} - \boldsymbol{w}_j)}{\Delta x} + \tilde{J}^+_{j-\frac{1}{2}} \frac{(\boldsymbol{w}_j - \boldsymbol{w}_{j-1})}{\Delta x} = \boldsymbol{D}_j. \tag{7.1}$$

In a region of supercritical flow, i.e. $\tilde{\lambda}_{1,j-\frac{1}{2}}$, $\tilde{\lambda}_{1,j+\frac{1}{2}} > 0$, $\tilde{\lambda}_{2,j+\frac{1}{2}}$, $\tilde{\lambda}_{2,j+\frac{1}{2}} > 0$, we have $\tilde{J}^+_{j\pm\frac{1}{2}} = \tilde{J}_{j\pm\frac{1}{2}}$ and $\tilde{J}^-_{j\pm\frac{1}{2}} = 0$. The scheme now reduces to

$$\frac{\boldsymbol{F}(\boldsymbol{w}_j) - \boldsymbol{F}(\boldsymbol{w}_{j-1})}{\Delta x} = \boldsymbol{D}_j,$$

which gives

$$\begin{aligned} Q_j &= Q_{j-1}, \\ \frac{F_j - F_{j-1}}{\Delta x} &= D_j, \end{aligned}$$

hence Q_i =constant is a solution. This explains the reason for the region of constant discharge between about 50m and 100m for problem 6, because this region corresponds to supercritical flow. Although the discharge will be constant for such a region, if the region is separated from the inflow boundary (where the discharge is specified as a boundary condition), then the constant discharge will not in general be at the correct level. Here for example the discharge is slightly above the correct level. In the case of a completely supercritical flow, at steady state the scheme is identical to the Engquist-Osher, Godunov and first-order upwind schemes.

For subcritical flow, i.e. $\tilde{\lambda}_{1,j-\frac{1}{2}}$, $\tilde{\lambda}_{1,j+\frac{1}{2}} < 0$, $\tilde{\lambda}_{2,j-\frac{1}{2}}$, $\tilde{\lambda}_{2,j+\frac{1}{2}} > 0$, the situation is more complicated. We have that

$$\tilde{J}^+_{j-\frac{1}{2}} = \frac{\tilde{\lambda}_{2,j-\frac{1}{2}}}{\tilde{\lambda}_{2,j-\frac{1}{2}} - \tilde{\lambda}_{1,j-\frac{1}{2}}} \begin{pmatrix} -\tilde{\lambda}_{1,j-\frac{1}{2}} & 1 \\ -\tilde{\lambda}_{1,j-\frac{1}{2}}\tilde{\lambda}_{2,j-\frac{1}{2}} & \tilde{\lambda}_{2,j-\frac{1}{2}} \end{pmatrix}$$

and

$$\tilde{J}^-_{j+\frac{1}{2}} = \frac{\tilde{\lambda}_{1,j+\frac{1}{2}}}{\tilde{\lambda}_{2,j+\frac{1}{2}} - \tilde{\lambda}_{1,j+\frac{1}{2}}} \begin{pmatrix} \tilde{\lambda}_{2,j+\frac{1}{2}} & -1 \\ \tilde{\lambda}_{1,j+\frac{1}{2}}\tilde{\lambda}_{2,j+\frac{1}{2}} & \tilde{\lambda}_{1,j+\frac{1}{2}} \end{pmatrix}.$$

If we attempt to obtain a constant discharge solution by setting $Q_{i-1} = Q_i = Q_{i+1}$, then the first component of equation (7.1) reduces to

$$\frac{\tilde{\lambda}_{2,j+\frac{1}{2}}\tilde{\lambda}_{1,j+\frac{1}{2}}}{\tilde{\lambda}_{2,j+\frac{1}{2}} - \tilde{\lambda}_{1,j+\frac{1}{2}}}(A_{j+1} - A_j) - \frac{\tilde{\lambda}_{2,j-\frac{1}{2}}\tilde{\lambda}_{1,j-\frac{1}{2}}}{\tilde{\lambda}_{2,j-\frac{1}{2}} - \tilde{\lambda}_{1,j-\frac{1}{2}}}(A_j - A_{j-1}) = 0.$$

The coefficients of $(A_{j+1} - A_j)$ and $(A_j - A_{j-1})$ must be positive, so this precludes the wetted area from having extrema in a subcritical region of flow. This is clearly nonsense, so we conclude that the difference equations are not in general consistent at steady state with a constant discharge solution.

Figure 7.13(b) also shows the discharge for Roe's scheme with upwinded source term. As for problem 1, the discharge is correct everywhere to within $10^{-5}\mathrm{m}^3\mathrm{s}^{-1}$ (except at an isolated point at the jump). Thus it appears that away from any transitions, the scheme is consistent with a constant discharge. The scheme can be written as

$$\begin{aligned} \frac{\boldsymbol{w}_j^{n+1} - \boldsymbol{w}_j^n}{\Delta t} &+ \left(\tilde{J}^-_{j+\frac{1}{2}}\right)^n \frac{(\boldsymbol{w}_{j+1}^n - \boldsymbol{w}_j^n)}{\Delta x} + \left(\tilde{J}^+_{j-\frac{1}{2}}\right)^n \frac{(\boldsymbol{w}_j^n - \boldsymbol{w}_{j-1}^n)}{\Delta x} \\ &= \Gamma\left(\tilde{J}^n_{j-\frac{1}{2}}\right)\tilde{\boldsymbol{D}}^n_{j-\frac{1}{2}} + \left(I - \Gamma\left(\tilde{J}^n_{j+\frac{1}{2}}\right)\right)\tilde{\boldsymbol{D}}^n_{j+\frac{1}{2}}, \end{aligned}$$

where

$$\tilde{\boldsymbol{D}}_{j+\frac{1}{2}} = \frac{\boldsymbol{D}_j + \boldsymbol{D}_{j+1}}{2}. \tag{7.2}$$

At steady state this reduces to

$$\tilde{J}^-_{j+\frac{1}{2}} \frac{(\boldsymbol{w}_{j+1} - \boldsymbol{w}_j)}{\Delta x} + \tilde{J}^+_{j-\frac{1}{2}} \frac{(\boldsymbol{w}_j - \boldsymbol{w}_{j-1})}{\Delta x} = \Gamma\left(\tilde{J}_{j-\frac{1}{2}}\right)\tilde{\boldsymbol{D}}_{j-\frac{1}{2}} + \left(I - \Gamma\left(\tilde{J}_{j+\frac{1}{2}}\right)\right)\tilde{\boldsymbol{D}}_{j+\frac{1}{2}}.$$

This scheme simplifies for supercritical flow as for the case with pointwise source term, since $\Gamma\left(\tilde{J}_{j-\frac{1}{2}}\right) = \Gamma\left(\tilde{J}_{j+\frac{1}{2}}\right) = I$, giving

$$\begin{aligned} Q_j &= Q_{j-1}, \\ \frac{F_j - F_{j-1}}{\Delta x} &= \tilde{D}_{j-\frac{1}{2}}, \end{aligned}$$

where

$$\tilde{D}_{j+\frac{1}{2}} = \frac{D_j + D_{j+1}}{2}. \tag{7.3}$$

This scheme is simply the trapezium rule applied to the steady flow equation in conservative form. As in the pointwise case the situation is considerably more complex in the case of subcritical flow, since the mass and momentum equations do not immediately decouple at steady state. For subcritical flow we have

$$\Gamma\left(\tilde{J}_{j-\frac{1}{2}}\right) = \frac{1}{\tilde{\lambda}_{2,j-\frac{1}{2}} - \tilde{\lambda}_{1,j-\frac{1}{2}}} \begin{pmatrix} -\tilde{\lambda}_{1,j-\frac{1}{2}} & 1 \\ -\tilde{\lambda}_{1,j-\frac{1}{2}}\tilde{\lambda}_{2,j-\frac{1}{2}} & \tilde{\lambda}_{2,j-\frac{1}{2}} \end{pmatrix}$$

and

$$I - \Gamma\left(\tilde{J}_{j+\frac{1}{2}}\right) = \frac{1}{\tilde{\lambda}_{2,j+\frac{1}{2}} - \tilde{\lambda}_{1,j+\frac{1}{2}}} \begin{pmatrix} \tilde{\lambda}_{2,j+\frac{1}{2}} & -1 \\ \tilde{\lambda}_{1,j+\frac{1}{2}}\tilde{\lambda}_{2,j+\frac{1}{2}} & -\tilde{\lambda}_{1,j+\frac{1}{2}} \end{pmatrix}.$$

If we try to obtain the constant discharge solution $Q_{i-1} = Q_i = Q_{i+1}$, then the scheme reduces to the following

$$\left(\tilde{\lambda}_{2,j+\frac{1}{2}}\tilde{\lambda}_{1,j+\frac{1}{2}}\frac{A_{j+1} - A_j}{\Delta x} + \tilde{D}_{j+\frac{1}{2}}\right)\frac{\boldsymbol{r}_{1,j+\frac{1}{2}}}{\tilde{\lambda}_{2,j+\frac{1}{2}} - \tilde{\lambda}_{1,j+\frac{1}{2}}}$$

$$- \left(\tilde{\lambda}_{2,j-\frac{1}{2}}\tilde{\lambda}_{1,j-\frac{1}{2}}\frac{A_j - A_{j-1}}{\Delta x} + \tilde{D}_{j-\frac{1}{2}}\right)\frac{\boldsymbol{r}_{2,j-\frac{1}{2}}}{\tilde{\lambda}_{2,j-\frac{1}{2}} - \tilde{\lambda}_{1,j-\frac{1}{2}}} = 0.$$

Clearly a solution to this relationship is given by

$$-\tilde{\lambda}_{2,j+\frac{1}{2}}\tilde{\lambda}_{1,j+\frac{1}{2}}\frac{A_{j+1} - A_j}{\Delta x} = \tilde{D}_{j+\frac{1}{2}},$$
$$-\tilde{\lambda}_{2,j-\frac{1}{2}}\tilde{\lambda}_{1,j-\frac{1}{2}}\frac{A_j - A_{j-1}}{\Delta x} = \tilde{D}_{j-\frac{1}{2}}.$$

Remarkably these reduce to

$$\frac{F_{j+1} - F_j}{\Delta x} = \tilde{D}_{j+\frac{1}{2}},$$
$$\frac{F_j - F_{j-1}}{\Delta x} = \tilde{D}_{j-\frac{1}{2}}.$$

We arrive at the conclusion that, at steady state and for subcritical flow, the scheme with upwinded source term may again reduce to the trapezium rule applied to the steady flow equation (in conservative form). For a purely subcritical flow, the solution can be solved by simply integrating upstream as follows:

$$\begin{aligned} Q_{j-1} &= Q_0 \\ \frac{F_j - F_{j-1}}{\Delta x} &= \frac{D_{j-1} + D_j}{2} \\ & \qquad j = N, N-1, \ldots, 1 \end{aligned}$$

where A_N is given by the depth boundary condition at outflow and Q_0 is the inflow discharge. Each step upstream requires the solution of a nonlinear equation which could be carried out using Newton-Raphson.

We have seen that at steady state and for both subcritical and supercritical flow, the scheme in effect reduces to the trapezium rule. This explains the second order

accuracy observed for problem 1, since the trapezium rule is a second order accurate discretisation.

Figure 7.14 compares the L_1 accuracy of the three schemes considered above for problem 6. This clearly illustrates that Roe's scheme with upwinded source term is the most accurate and the Engquist-Osher scheme is the least accurate of the methods. Both Roe's scheme with pointwise source term and the Engquist-Osher scheme demonstrated first order accuracy, whilst Roe's scheme with upwinded source term gives a slightly higher order of about 1.3. The scheme does not give full second order accuracy for this problem because of the nonuniform convergence at the jump. Figure 7.15(a-d) compares the accuracy of the three schemes at the point $x = 30$m,

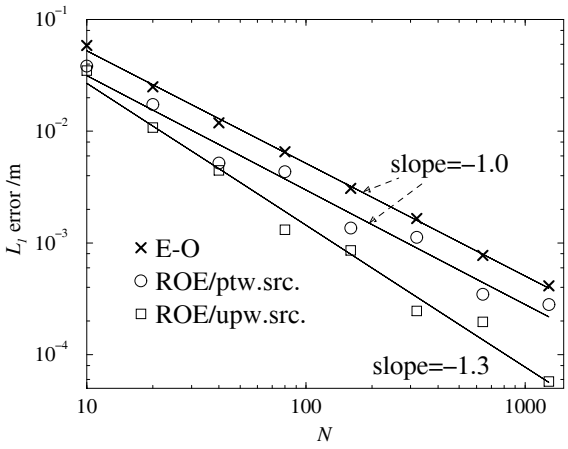


Figure 7.14: L_1 errors for Problem 6.

60m, 90m, 120m, respectively. At all of the these point, except $x = 30$m, Roe's scheme with upwinded source term is by far the most accurate of the three methods. At $x = 30$m the Engquist-Osher scheme is the most accurate for N up until around 300 when it is overtaken by Roe's scheme with upwinded source term. This is most likely due to closeness of this point to the smooth transition, since the Engquist-Osher scheme is in general found to give superior solutions at such transitions. In general the Engquist-Osher scheme and Roe's scheme with pointwise source term demonstrate first order accuracy at the points considered. The only anomaly is at $x = 30$m where the Engquist-Osher scheme gives a higher than expected order of about 1.5. From the previous arguments we would expect Roe's scheme with upwinded source term to show second order accuracy. This indeed the case, except

at $x = 120$m where the error initially decreases at a rate consistent with an order of accuracy of 2.5. For N above 160 the error remains roughly constant at about 10^{-6}. The reason that the error decreases no further is most likely due to the presence of rounding errors.

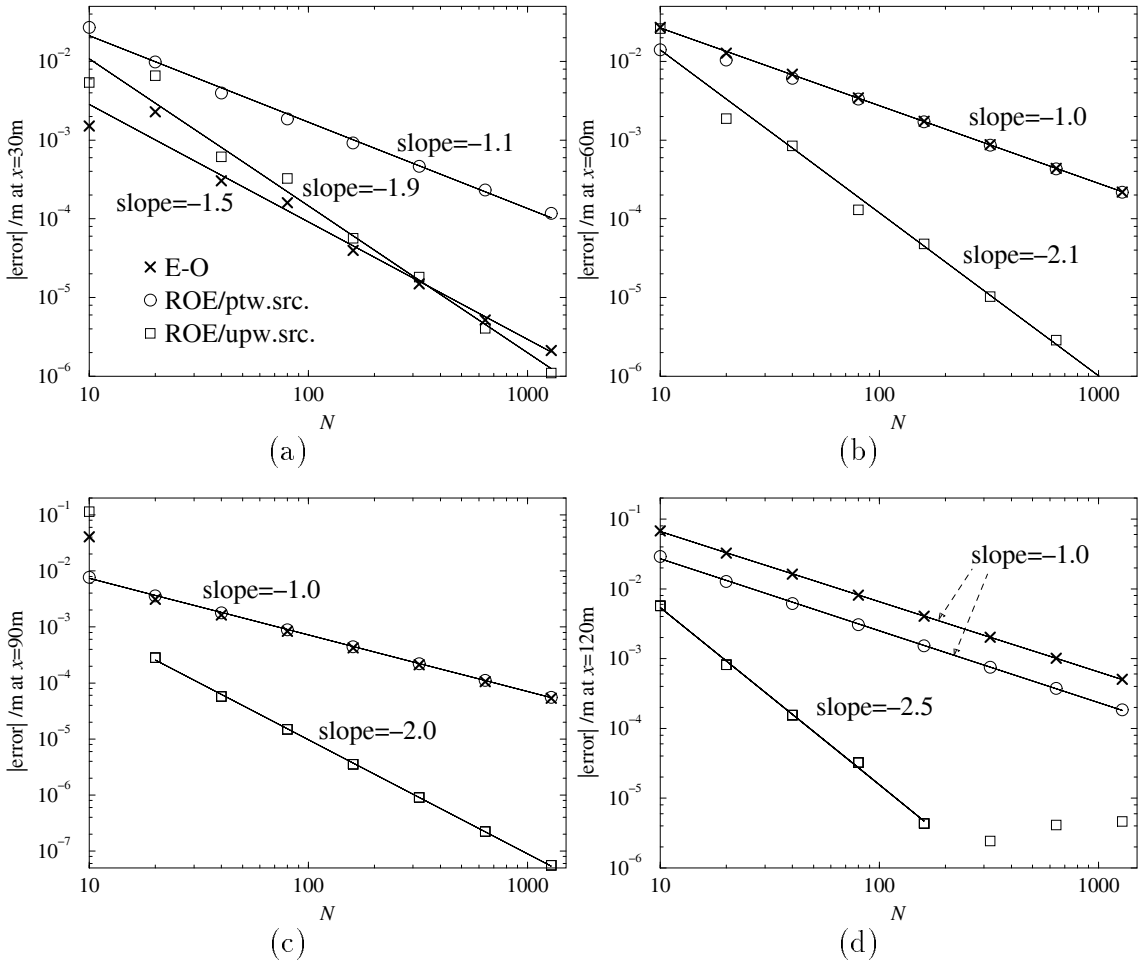


Figure 7.15: Errors for problem 6.

7.3 Higher Order Accuracy

This section generalises the "scalar approach" in order to obtain higher order approximations to the steady state solutions. In the previous section we obtained second order accuracy using Roe's approximate Riemann solver combined with a particular method of upwinding the source term. We apply the direct analogue of this scheme to the scalar equation

$$\frac{\partial h}{\partial t} + \frac{\partial}{\partial x} f(h) = -D, \tag{7.4}$$

where $f(h) = -F(h)$. This gives the first-order upwind scheme with upwinded source term. Next a generalisation of the Engquist-Osher scheme is considered which again comes down to upwinding the source term. These methods achieve higher order accuracy by solely modifying the method of discretisation of the source term, using the fact that the higher order accuracy is only required at the steady state. More usually, methods are designed for the homogeneous equations and are required to give higher order accuracy not just at steady state, but also in the transient state of the solution. Schemes which achieve this include the high order TVD schemes which are discussed in section 3.5. We apply two examples of such schemes to the equation (7.4) and the accuracy is compared against the source term upwinding approach.

7.3.1 Upwinded Source Terms

The analogue of Roe's scheme with upwinded source terms for the scalar equation (7.4) can be written as

$$\begin{aligned} \frac{h_j^{n+1} - h_j^n}{\Delta t} + \mathcal{T}_j^{\text{UPW-2}} h^n = 0, \qquad j = 1, 2, \ldots, N-1 \\ u_0^n = \gamma_0, \quad u_N^n = \gamma_1, \end{aligned} \tag{7.5}$$

where

$$\begin{aligned} \mathcal{T}_j^{\text{UPW-2}} h &= s^-_{j+\frac{1}{2}} \frac{(h_{j+1} - h_j)}{\Delta x} + s^+_{j-\frac{1}{2}} \frac{(h_j - h_{j-1})}{\Delta x} \\ &+ \Gamma\left(s_{j-\frac{1}{2}}\right) \tilde{D}_{j-\frac{1}{2}} + \left(1 - \Gamma\left(s_{j+\frac{1}{2}}\right)\right) \tilde{D}_{j+\frac{1}{2}}, \end{aligned}$$

and $s_{j+\frac{1}{2}}$ is given by (3.14). We can also write

$$\begin{aligned}\mathcal{T}_j^{\text{UPW-2}} h &= \frac{g^{\text{FOU}}_{j+\frac{1}{2}} - g^{\text{FOU}}_{j-\frac{1}{2}}}{\Delta x} \\ &+ \Gamma\left(s_{j-\frac{1}{2}}\right)\tilde{D}_{j-\frac{1}{2}} + \left(1-\Gamma\left(s_{j+\frac{1}{2}}\right)\right)\tilde{D}_{j+\frac{1}{2}}.\end{aligned}$$

At steady state the scheme reduces to

$$\mathcal{T}_j^{\text{UPW-2}} h = 0, \qquad j = 1, 2, \ldots, N-1$$

$$u_0 = \gamma_0, \qquad u_N = \gamma_1.$$

We observe that for subcritical flow ($s_{j\pm\frac{1}{2}} < 0$) this reduces to

$$\frac{F_{j+1} - F_j}{\Delta x} = \tilde{D}_{j+\frac{1}{2}},$$

and for supercritical flow ($s_{j\pm 1} > 0$) it reduces to

$$\frac{F_j - F_{j-1}}{\Delta x} = \tilde{D}_{j-\frac{1}{2}}.$$

Therefore for the choice (7.3), at steady state and away form transitions, the scheme reduces to the trapezium rule. This is exactly the behaviour encountered in the previous section for Roe's scheme with upwinded source term. In fact for flows which are entirely subcritical or entirely supercritical the schemes will give identical solutions (except possibly points at the ends of the reach where Roe's scheme is treated differently to the "scalar approach").

Figure 7.16 shows results for the scheme discussed above, which we refer to as the upwind-2 scheme. The accuracy can clearly be seen to be superior to that of the Engquist-Osher scheme.

A similar idea to the above is used in [33] to extend the Engquist-Osher scheme to a second order accurate method for solving the singular perturbation problem (4.12). The scheme for the case $\epsilon = 0$ is given by

$$\mathcal{T}_j^{\text{E-O-2}} h = 0, \qquad j = 1, 2, \ldots, N-1$$

$$u_0 = \gamma_0, \qquad u_N = \gamma_1,$$

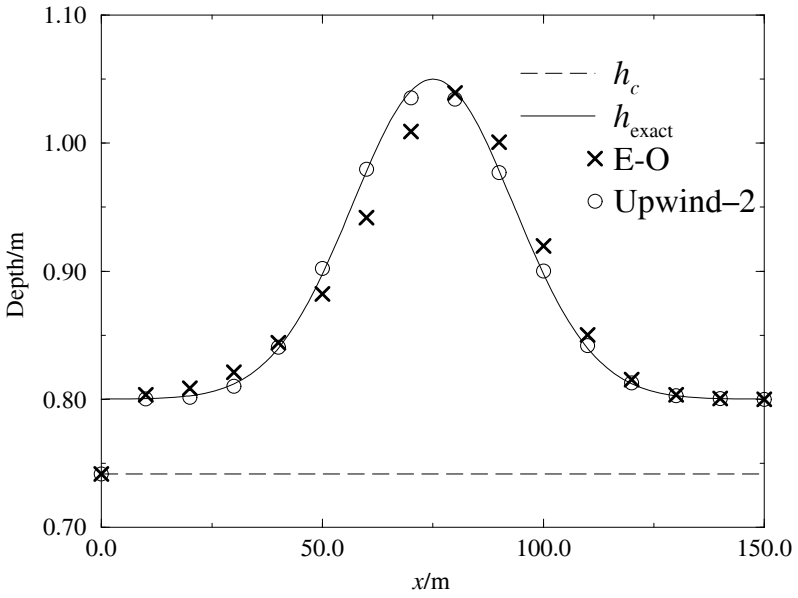


Figure 7.16: The upwind-2 and Engquist-Osher schemes for problem 1 (Δx =10m).

the operator $\mathcal{T}_j^{\text{E-O-2}}$ being given by

$$\mathcal{T}_j^{\text{E-O-2}} h \;=\; \frac{g^{\text{E-O}}_{j+\frac{1}{2}} - g^{\text{E-O}}_{j-\frac{1}{2}}}{\Delta x} + \chi_j^- D_{j-1} + \chi_j^0 D_j + \chi_j^+ D_{j+1},$$

where

$$\begin{aligned}
\chi_j^- &= \chi\left(\frac{pf'(h_{j-1})}{\sqrt{\Delta x}}\right), \\
\chi_j^0 &= 1 - \chi\left(\frac{pf'(h_j)}{\sqrt{\Delta x}}\right) - \chi\left(\frac{-pf'(h_j)}{\sqrt{\Delta x}}\right) = 1 - \chi_{j+1}^- - \chi_{j-1}^+, \\
\chi_j^+ &= \chi\left(\frac{-pf'(h_{j+1})}{\sqrt{\Delta x}}\right),
\end{aligned}$$

$p \geq 0$ is a parameter and χ is the smooth increasing function

$$\chi(r) = \begin{cases} 0 & r < 0 \\ r^2 & 0 \leq r \leq \frac{1}{2} \\ \frac{1}{2} - (1-r)^2 & \frac{1}{2} \leq r \leq 1 \\ \frac{1}{2} & r > 1, \end{cases}$$

connecting the values 0 and $\frac{1}{2}$. The precise form of this function is unimportant and any monotone C^1 function connecting 0 to $\frac{1}{2}$ could be taken. The case $p = 0$ corresponds to the Engquist-Osher scheme with pointwise source term, so we consider

the case where p is positive. To interpret the scheme we observe the following. For $f'(h_{j-1}),\ f'(h_j),\ f'(h_{j+1}) \leq -\sqrt{\Delta x}/p$ the scheme reduces to

$$\frac{F_{j+1} - F_j}{\Delta x} = \frac{D_j + D_{j+1}}{2}.$$

This case corresponds to subcritical flow with depths above a certain distance greater then the critical depth. For $f'(h_{j-1}),\ f'(h_j),\ f'(h_{j+1}) \geq \sqrt{\Delta x}/p$ the scheme reduces to

$$\frac{F_j - F_{j-1}}{\Delta x} = \frac{D_{j-1} + D_j}{2}.$$

This case corresponds to supercritical flow with depths above a certain distance below the critical depth. Essentially the scheme is upwinding the source term as for the upwind-2 scheme. The difference here is that the switching is performed in a smooth manner. We will see that this smoothness makes the difference equations more amenable to solution than for the upwind-2 scheme The parameter p controls the rate at which the source terms switch between the subcritical and supercritical forms across transitions. The higher the value of p, the greater the speed the switching occurs and the more accurate the scheme is, since the scheme corresponds to the second order accurate trapezium rule for a greater part of the solution. If p is too great the authors of [33] predict that the scheme may become badly behaved, for example by having no solutions or multiple solutions. In Appendix A we adapt the theory in [33], which shows the uniqueness of the discrete solution, to also find conditions which guarantee convergence of the time stepping iteration

$$\frac{h_j^{n+1} - h_j^n}{\Delta t} + \mathcal{T}_j^{\text{E-O-2}} h^n = 0.$$

Under the conditions of Theorem 3 it is shown that if

$$p\sqrt{h}M_1 \leq 1, \qquad 4p^2 M_2 \leq 1, \tag{7.6}$$

where $|D_h(x_j, h)| \leq M_1$ and $|f''(h)D(x_j, h)| \leq M_2$ for all $0 \leq j \leq N$ and $\underline{h} \leq h \leq \overline{h}$, then the system of difference equations has exactly one solution satisfying $\underline{h} \leq h_i \leq \overline{h}$, $i = 0, \ldots, N$. The modified CFL condition then requires that if

$$\alpha = \min_{0 \leq i \leq N} \{h_i^0, \underline{h}\}, \qquad \beta = \max_{0 \leq i \leq N} \{h_i^0, \overline{h}\},$$

then

$$\Delta t \left(\frac{|f'(h)|}{\Delta x} + D_h(x_j, h) + \frac{p}{\sqrt{\Delta x}} |f''(h) D(x_j, h)| \right) \leq 1,$$

for all $h \in [\alpha, \beta]$ and $0 \leq j \leq N$.

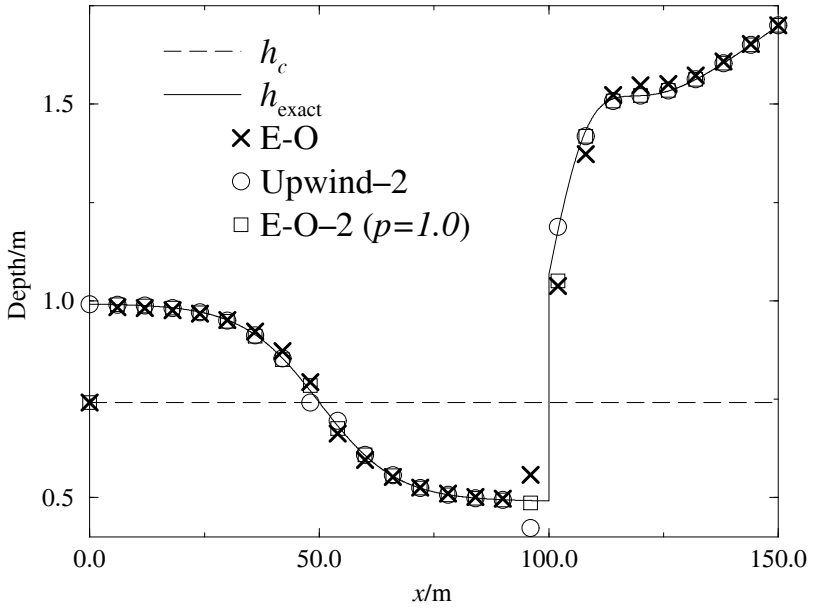


Figure 7.17: Upwind-2 and E-O-2 ($p = 1$) for problem 6 (Δx =6m).

Figure 7.17 compares results for the upwind-2 scheme and the E-O-2 scheme with $p = 1$ for problem 6. Away from the transitions the two schemes give almost identical results and give better accuracy than the Engquist-Osher scheme which is also shown. Near the smooth transition the E-O-2 scheme gives the most accurate solution, whilst the upwind-2 scheme is the least accurate. As in the case of Roe's scheme with upwinded source term (Figure 7.13), the upwind-2 scheme is extremely accurate at the grid point downstream of the jump, but undershoots the solution by a significant amount at the grid-point upstream of the jump. In fact the solutions for these two schemes are very similar due to the fact that they reduce to the same scheme away from transitions. They also appear to behave very similarly at transitions. The E-O-2 scheme is less accurate than the upwind-2 scheme at the grid-point downstream of the jump, but considerably more accurate, with no undershoot, at the grid-point upstream of the jump. To investigate how sensitive the accuracy of the E-O-2 scheme is to the value of the parameter p, we plot the errors as a function of p for problem 6. Each of the Figures 7.18(a-d) shows the error at a particular point along the

channel as a function of p, for three different grid spacings ($N = 20$, $N = 80$ and $N = 320$). The behaviour is very similar at each of the four points. For a particular grid spacing there are two regions of constant error, separated by a transition region. The constant region for low p corresponds to first order accuracy since the error reduces roughly by a factor of 4 in going from $N = 20$ to $N = 80$ and from $N = 80$ to $N = 320$. The constant region for high p corresponds to second order accuracy since the error reduces roughly by a factor of 16 in going from $N = 20$ to $N = 80$ and from $N = 80$ to $N = 320$. In the transition region the error behaves somewhat erratically. In almost all of the cases shown, there is a spike with the error unexpectedly increasing. The position of the transition region moves to lower values of p as N increases and for this problem is about one order of magnitude wide.

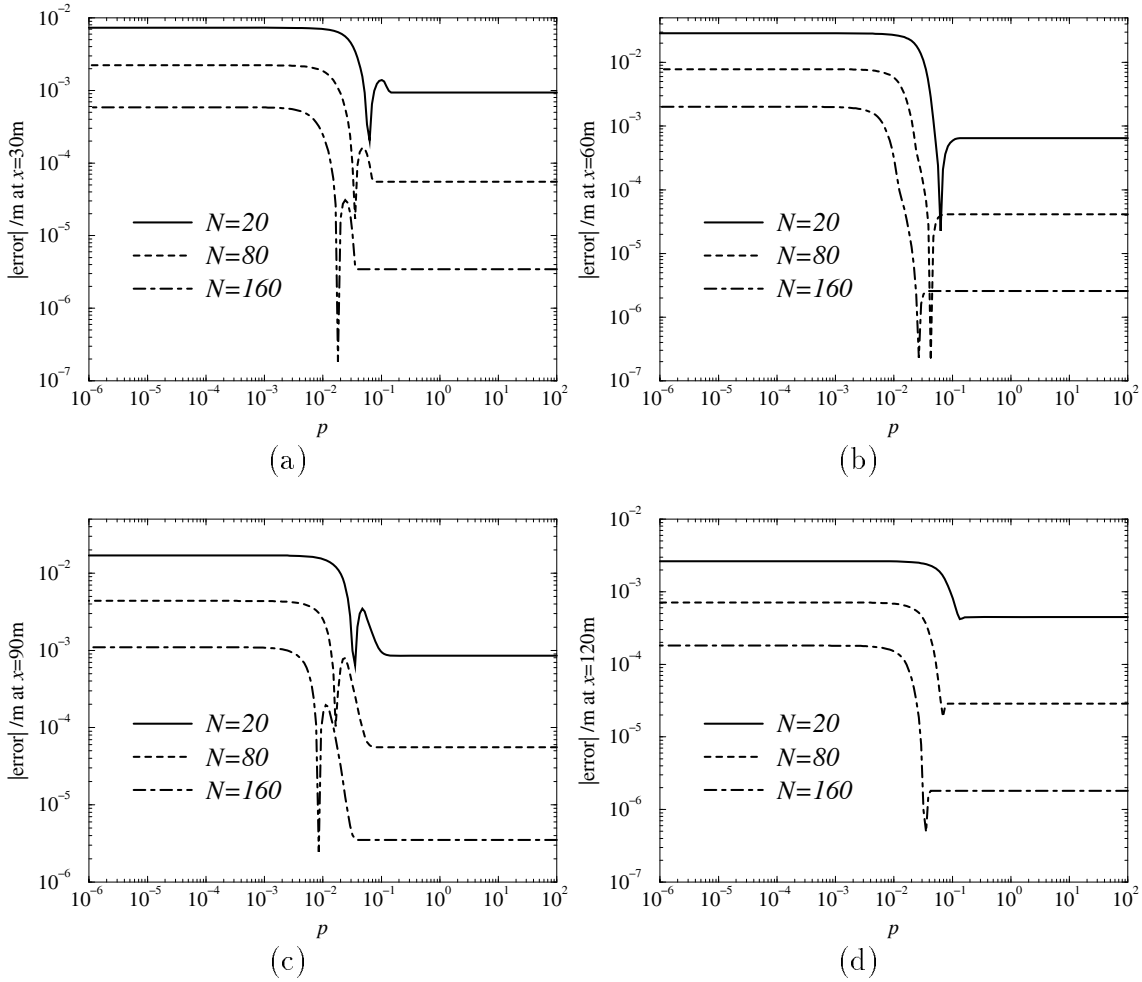


Figure 7.18: Errors for the E-O-2 scheme and problem 6 as a function of p.

7.3.2 High Order TVD Schemes

The technique in the previous section achieves higher accuracy at steady state by modifying the discretisation of the source term to create $O(\Delta x)$ terms in the truncation error, which cancel out all terms of the same order from the approximation to the advective part of the differential equation. A more common technique of obtaining higher order accuracy is to increase the order of accuracy of the approximation to the advective part, allowing high order accuracy not just at steady state but also in the transient solution. Classical second order approximations such as the Lax-Wendroff scheme are available, but these are found to be useless since they become unstable near discontinuities. The instabilities arise due to the lack of sufficient numerical dissipation in the schemes to damp out any oscillations at discontinuities.

A method of overcoming the problem of instabilities, which is discussed in section 3.5, is to add a term to the second order approximation which increases the numerical dissipation of the scheme only near a discontinuity. By adding numerical dissipation only at a discontinuity, the accuracy in smooth parts of the solution should be unaffected. The numerical dissipation is controlled by a nonlinear function of the solution, known as a limiter function. To ensure that the resulting scheme is stable (i.e. oscillation free) the limiter function is chosen such that the resulting scheme is TVD (see section 3.4).

The approach discussed above leads to a class of methods known as high order TVD schemes. An example of such a scheme is given in section 3.5. This example is not immediately suitable for steady state calculations, because the numerical flux function depends on the time step and so at steady state the difference equations also depend on the time step. However the only consequence of removing the terms which depend on Δt is a reduction in the temporal order of accuracy from second order to first order. The temporal order of accuracy is not important for steady state calculations. Other high order TVD schemes often require the removal of second order time accuracy terms before they are suitable for steady state computations. Yee[72] reviews many different high order TVD schemes, and in particular a family

of schemes with numerical flux function written as

$$g_{j+\frac{1}{2}} = \frac{1}{2}\left(f(h_j) + f(h_{j+1}) + \phi_{j+\frac{1}{2}}\right),$$

for different functions $\phi_{j+\frac{1}{2}}$. In general the resulting scheme is a five point scheme and this adds difficulties at the boundaries. In the case of the three point schemes, the values of the boundary nodes h_0^n and h_N^n are fixed regardless of whether these represent physical boundary conditions. We treat the boundaries for five-point schemes in the same unsophisticated manner, by now fixing the values of h_0^n, h_{-1}^n and h_N^n, h_{N+1}^n. This gives the scheme

$$\frac{h_j^{n+1} - h_j^n}{\Delta t} + \mathcal{T}_j^{\mathrm{TVD}} h^n = 0, \qquad j = 1, 2, \ldots, N-1$$
$$u_{-1}^n = u_0^n = \gamma_0, \quad u_{N+1}^n = u_N^n = \gamma_1,$$

where

$$\mathcal{T}_j^{\mathrm{TVD}} h = \frac{g_{j+\frac{1}{2}} - g_{j-\frac{1}{2}}}{\Delta x} + D_j.$$

We apply two different forms of this scheme, and these are the Harten-Yee upwind scheme and the Yee-Roe-Davis symmetric scheme. Details of the particular forms of $\phi_{j+\frac{1}{2}}$ are given below.

Harten-Yee Upwind Scheme This scheme is a variation due to Yee[72] of the modified flux approach of Harten[20], and is given by

$$\phi_{j+\frac{1}{2}} = \frac{1}{2}\left|s_{j+\frac{1}{2}}\right|(\eta_j + \eta_{j+1}) - \left|s_{j+\frac{1}{2}} + \gamma_{j+\frac{1}{2}}\right|(h_{j+1} - h_j),$$

where $s_{j+\frac{1}{2}}$ is given by (3.14),

$$\gamma_{j+\frac{1}{2}} = \frac{1}{2}\left|s_{j+\frac{1}{2}}\right| \begin{cases} \dfrac{\eta_{j+1} - \eta_j}{h_{j+1} - h_j} & h_{j+1} \neq h_j \\ 0 & h_{j+1} = h_j \end{cases}$$

and

$$\eta_j = \eta(h_{j+1} - h_j, h_j - h_{j-1}).$$

The function η is given by

$$\eta(x, y) = \mathrm{minmod}(x, y),$$

where the minmod function is given by

$$\text{minmod}(x, y) = \text{sgn}(x) \max\{0, \min\{|x|, y \cdot \text{sgn}(x)\}\}.$$

Other forms of the function η are given in [72].

Yee-Roe-Davis Symmetric Scheme This scheme is a generalisation by Yee ([71], [70]) of the schemes of Roe[56] and Davis[10], and is given by

$$\phi_{j+\frac{1}{2}} = -\left|s_{j+\frac{1}{2}}\right| \left((h_{j+1} - h_j) - \psi_{j+\frac{1}{2}}\right),$$

where $s_{j+\frac{1}{2}}$ is given by (3.14) and

$$\psi_{j+\frac{1}{2}} = \text{minmod}(h_{j+1} - h_j, h_j - h_{j-1}) + \text{minmod}(h_{j+1} - h_j, h_{j+2} - h_{j+1}) - (h_{j+1} - h_j).$$

Alternative forms for $\psi_{j+\frac{1}{2}}$ are given in [72].

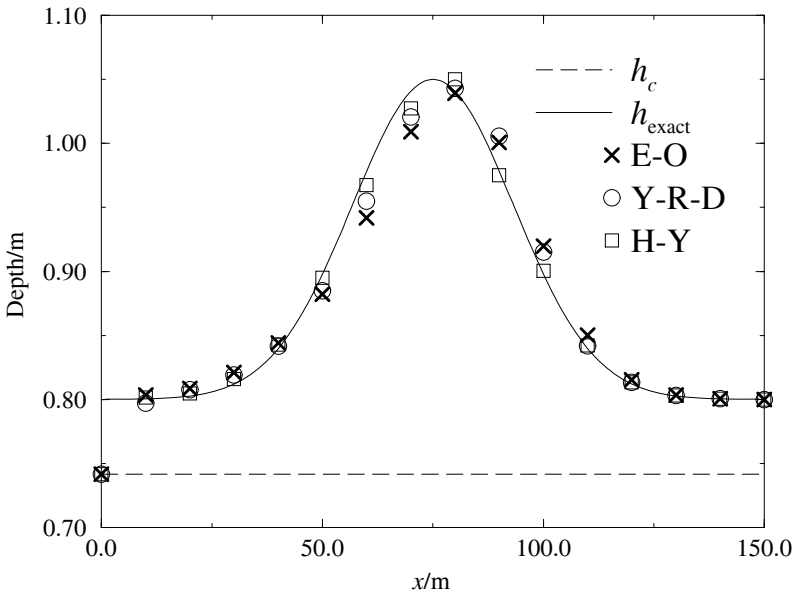


Figure 7.19: Y-R-D and H-R schemes for problem 1 (Δx =10m).

Figure 7.19 compares the results for the Harten-Yee upwind scheme (H-Y) and the Yee-Roe-Davis symmetric scheme (Y-R-D) for problem 1. Both these schemes give superior accuracy to the Engquist-Osher scheme, although the H-Y scheme is significantly more accurate than the Y-R-D scheme. Figure 7.20 compares the results for the Y-R-D and H-Y schemes for problem 6. Both schemes on average give

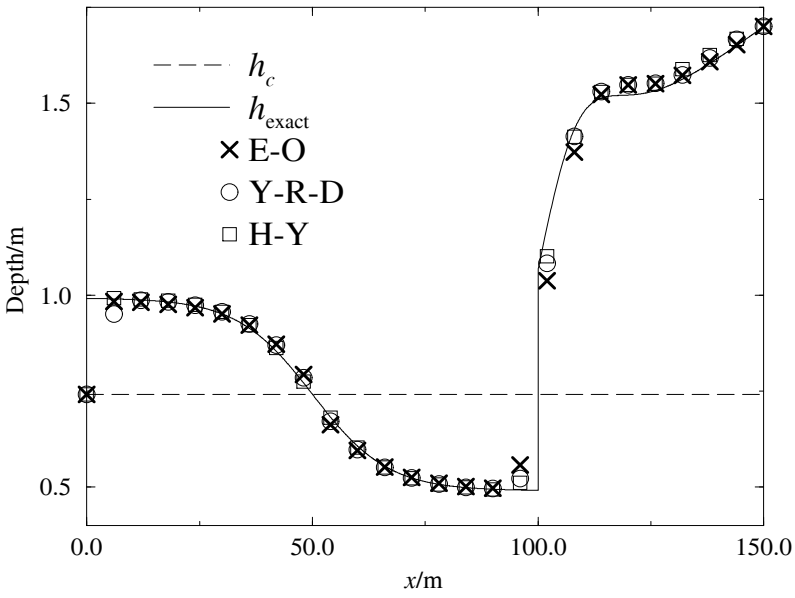


Figure 7.20: Y-R-D and H-R schemes for problem 1 (Δx =6m).

better accuracy than the Engquist-Osher scheme, and again the H-Y scheme is more accurate than the Y-R-D scheme. A feature to note is the large error at the second grid-point for the Y-R-D scheme, as compared to the other schemes. This is due to the symmetric nature of the scheme. The region of non-uniform convergence at the boundaries is greater than for upwind schemes.

Figures 7.21 (a) and (b) show the L_1 and L_∞ errors for problem 1 for the two high order TVD schemes as well as the upwind-2 scheme and Roe's scheme with upwinded source term. The high order TVD schemes give inferior accuracy to both Roe's scheme and the upwind-2 scheme. The H-Y scheme is more accurate than the Y-R-D scheme and exhibits second order accuracy in both measures. The L_∞ error for the Y-R-D scheme remains roughly constant as the number grid-points decreases, and this is due to nonuniform convergence at the inflow boundary. For the upwind type schemes, the solution in the interior of the domain is not influenced by the fact that the nodes at the inflow boundary are fixed to values which do not approximate the exact solution. This is not the case for symmetric schemes, where we have already observed that the error at the first interior node is considerably larger than for the other schemes. In fact the error at x_1 remains roughly constant as N increases, and it is this error that prevents the L_∞ error from decreasing. The nonuniform

convergence at the inflow boundary also explains why, in the L_1 measure, the order of accuracy has degraded from two to 1.6.

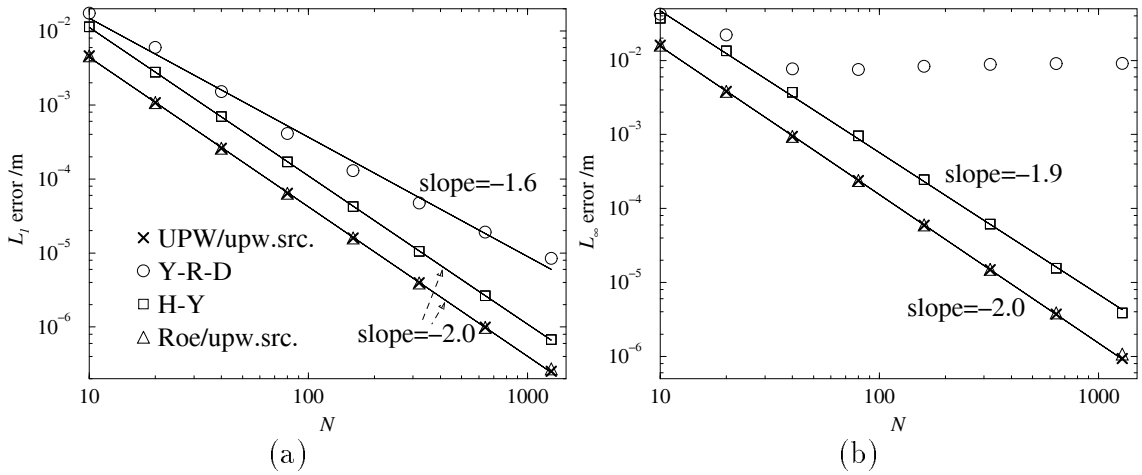


Figure 7.21: L_1 and L_∞ errors for problem 1.

Figure 7.22 compares the L_1 accuracy of the Y-R-D, H-Y, Upwind-2 and E-O-2 schemes and Roe's scheme with upwinded source term for problem 6. All the schemes demonstrate roughly first order accuracy. The accuracy of the upwind-2 scheme and Roe's scheme with upwinded source term is very similar and slightly superior to that of the E-O-2 scheme. These three schemes are more accurate than both of the high order TVD schemes with the H-Y scheme again being more accurate than the Y-R-D scheme. Figures 7.23(a-d) show plots of the accuracy at the points $x = 30$m, $x = 60$m, $x = 90$m and $x = 120$m, respectively. At all four points the upwind-2 scheme, Roe's scheme with upwinded source term and the E-O-2 all give very similar accuracy. These schemes are more accurate than the high order TVD schemes with the H-Y scheme again more accurate than the Y-R-D scheme. The schemes all demonstrate at least second accuracy, except at $x = 90$m where the high order TVD schemes reduce to first order accuracy. It is well known that such schemes reduce to first order accuracy at points of extrema and this appears to be what is happening at this point.

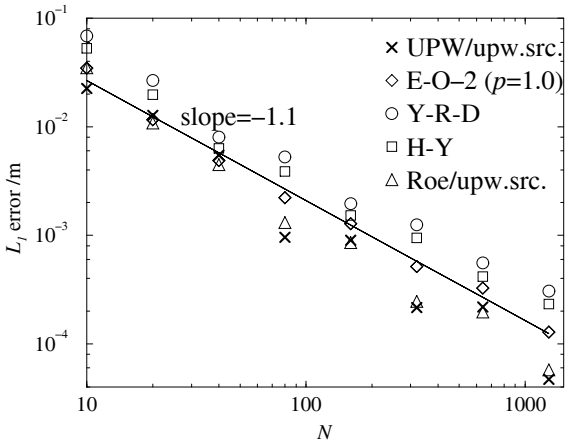


Figure 7.22: L_1 errors for problem 6.

7.4 Conclusions

The monotone Engquist-Osher and Godunov schemes and the first-order upwind scheme are all found to give comparable accuracy. The Godunov and first-order upwind schemes give sharper jumps than the Engquist-Osher scheme, whereas the Engquist-Osher and Godunov schemes give more accurate smooth transitions than the first-order upwind scheme. The Godunov scheme is thus the best of these three schemes. The Lax-Friedrichs scheme is found to be too diffusive to be of use. The a-priori estimate for the time step for the time stepping iteration is in general found to be of poor quality, in that significant saving in computation time can be obtained by taking a larger time-step. This is most-likely due to the poor quality of the a-priori bounds on the solution. Roe's scheme with pointwise source terms is in many cases found to be more accurate than the scalar schemes, even though the corresponding discharge deviates significantly from the correct constant discharge. Upwinding the source term is found not only to yield a constant discharge, but to give second-order accuracy for the depth at steady state. We showed that this was because the scheme reduced to the Trapezium rule in smooth regions of the solution. The idea of upwinding the source term was applied to the scalar methods. The upwind-2 scheme gave almost identical accuracy to Roe's scheme with upwinded source term. The E-O-2 scheme was in general slightly less accurate. Of the two high order TVD schemes we applied, neither were as accurate as the schemes with upwinded source term. The Harten-Yee upwind scheme being more accurate than the Yee-Roe-Davis

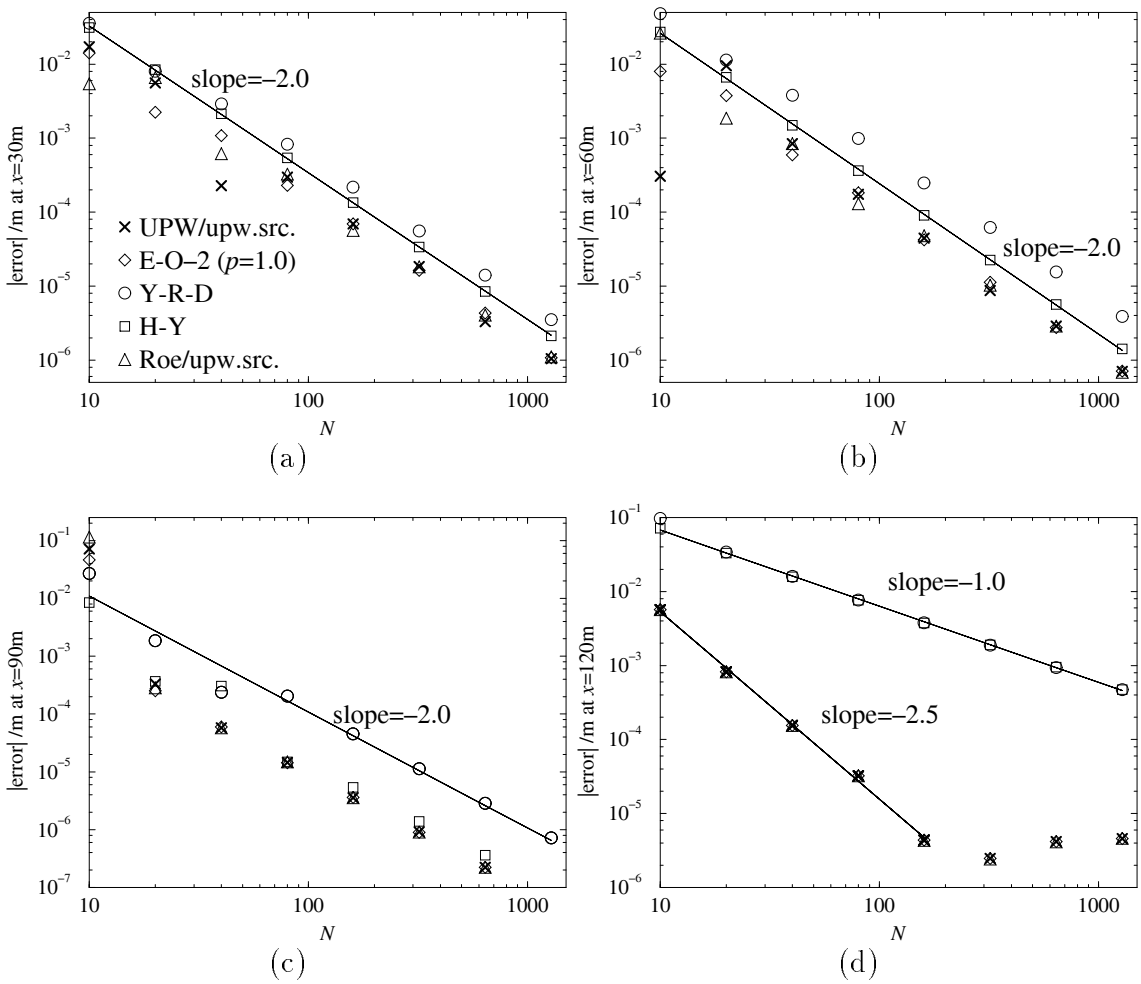


Figure 7.23: Errors for problem 6.

symmetric scheme.

Chapter 8

Computational Efficiency

This chapter explores different methods for solving the difference equations which arise from the various numerical schemes described in the previous chapter. The only method discussed so far is the time stepping iteration. We hope to take advantage of the simplicity of the difference equations to allow easy application of other potentially more efficient methods such as Newton's method. Even with the time stepping iteration it is reasonable to expect the "scalar approach" to be less computationally intensive than Roe's scheme, since we only solve for half the number of unknowns, for the same grid spacing. Therefore a 50% reduction in CPU time over Roe's scheme seems likely.

8.1 The Time Stepping Iteration

The performance of the time stepping iteration used to solve the Engquist-Osher scheme for problem 6 is illustrated in Table 8.1. For each grid spacing the performance is given for two different time steps, the optimum time step (the time step which results in the fastest convergence rate) and the largest time step which satisfies the CFL condition of Theorem 10 (and hence ensures the convergence of the iteration). Measures of performance are given by the number of time steps and the total CPU time taken. It should be noted that the CPU times are not particularly accurate, especially for small time intervals, and so should only be used to indicate the magnitude of the CPU time used.

N	Optimum			Monotonicity		
	Δt	No. Steps	CPU time/s	Δt	No. Steps	CPU time/s
10	0.084	48	0.108×10^{-1}	0.046	95	0.154×10^{-1}
20	0.058	56	0.151×10^{-1}	0.033	107	0.295×10^{-1}
40	0.035	96	0.512×10^{-1}	0.022	158	0.868×10^{-1}
80	0.015	204	0.226	0.011	281	0.311
160	0.0073	441	0.931	0.0059	548	0.115×10^{1}
320	0.0035	864	0.367×10^{1}	0.003	1010	0.408×10^{1}
640	0.0016	1921	0.161×10^{2}	0.0015	2049	0.190×10^{2}
1280	0.0008	3840	0.639×10^{2}	0.00076	4043	0.719×10^{2}

Table 8.1: Time stepping for the Engquist-Osher scheme for Problem 6

The criterion used to test for convergence is

$$\sqrt{\frac{1}{N-1}\sum_{j=1}^{N-1}(\mathcal{T}_j h^n)^2} < tol_h, \tag{8.1}$$

where $tol_h = 10^{-8}$. This is also used for all other iterative methods for solving the difference equations. In the particular case of the time stepping iteration it is equivalent to

$$\sqrt{\frac{1}{N-1}\sum_{j=1}^{N-1}\left(\frac{h_j^{n+1}-h_j^n}{\Delta t}\right)^2} < tol_h. \tag{8.2}$$

The optimum time step is in all cases found to be greater than that arising from the CFL condition. For problem 6 this is by a factor of about 2 for small N. The two values become much closer as N becomes larger. The number of iterations and hence the CPU time varies inversely proportionally to the size of the time step for time steps below the optimum value (this is because the iteration process is modelling the transient behaviour of a PDE and so the time at which steady state is attained $t_{\text{end}} = n_{\text{end}}\Delta t$ is essentially independent of Δt). Hence for problem 6, the CFL time step yields a CPU time of at most twice the best attainable. The situation can be considerably worse than this. For example in problem 1 the optimum time step is consistently greater than the CFL time step by a factor of between three and four, for each N. The difference between the CFL time step and the optimum time step

is dependent on the tightness of the bounds $\underline{h}$ and $\overline{h}$, since the CFL condition must hold for depths between these. Section 7.1 discusses the reason why these bounds are not necessarily tight. Even if tighter bounds on the analytic solution are known, convergence is not guaranteed by ensuring the CFL condition is satisfied over this new smaller range, unless the time stepping iteration can also be shown to satisfy these new bounds. Convergence may still occur for time steps slightly higher than the optimum value; however the number of iterations required for convergence increases. The time stepping iteration applied to the Godunov and first-order upwind schemes is found to behave very similarly to the case of the Engquist-Osher scheme

Table 8.2 demonstrates the performance of Roe's scheme for problem 6, for near

	Pointwise Source Term			**Upwinded Source Term**		
N	Δt	**No. Steps**	**CPU time/s**	Δt	**No. Steps**	**CPU time/s**
10	3.2	130	0.599×10^{-1}	3.4	330	0.162
20	1.6	340	0.301	1.6	424	0.402
40	0.75	631	0.111×10^{1}	0.7	593	0.119×10^{1}
80	0.35	1426	0.542×10^{1}	0.32	1743	0.675×10^{1}
160	0.14	3485	0.265×10^{2}	0.15	3276	0.255×10^{2}
320	0.07	8155	0.123×10^{3}	0.07	8102	0.125×10^{3}
640	0.035	13996	0.427×10^{3}	0.03	17333	0.535×10^{3}
1280	0.017	33640	0.204×10^{4}	0.015	39579	0.245×10^{4}

Table 8.2: Roe's scheme for problem 6 using close to optimum time steps

optimal time steps. The performance is not affected to a great extent by the choice of source term discretisation. For other test cases the upwinded discretisation can result in a more noticeable increase in computation time. To obtain a fair comparison between the performance of Roe's scheme and that of the scalar methods we use the consistent convergence criterion

$$\sqrt{\frac{1}{N+1}\sum_{j=0}^{N}\left(\frac{h_j^{n+1}-h_j^n}{\Delta t}\right)^2} < tol_h,$$

$$\sqrt{\frac{1}{N+1}\sum_{j=0}^{N}\left(\frac{Q_j^{n+1}-Q_j^n}{\Delta t}\right)^2} < tol_Q,$$

where $tol_Q = 10^{-7}$.

Comparing the computational expense of Roe against the Engquist-Osher scheme, it can be seen that if the optimum time step is used the Engquist-Osher scheme is in the order of fifty times cheaper than Roe's scheme. Even if the CFL time step is used the Engquist-Osher scheme is still around 25 times cheaper. The performance differential may be slightly exaggerated by the fact that the Roe computer code is not as streamlined as the Engquist-Osher code. Even when taking this into account, the difference is still significant and is typical of many of the test problems encountered.

We now ask whether the CPU savings observed above still occur when the time stepping iteration is applied to the second order scalar schemes. Appendix A gives the necessary changes to the theory in Chapter 5 to yield a modified CFL condition for the E-O-2 scheme, provided that the parameter p is below a certain critical value. Unlike the case of the first order schemes where only computational efficiency is lost by insisting that the CFL be satisfied, in the second order case solution accuracy may also be lost. For this reason we will in general choose p to be higher than the critical value. There is no corresponding theory for the upwind-2 scheme.

Table 8.3 illustrates the performance of the time stepping iteration applied to upwind-2 and the E-O-2 scheme ($p = 1$) for problem 6. These two schemes give very similar levels of performance, and in a number of cases can converge in fewer iterations than the first order schemes, although they are more complicated to implement and so on average are more expensive. In Section 7.3.1 the dependence of the accuracy of the E-O-2 scheme on the parameter p is described in terms of three regions, a first order region, a second order region and a transitional region. The behaviour of the time stepping iteration can also be described in terms of these ranges of p. In the first order region the method not surprisingly behaves almost identically to the first order Engquist-Osher scheme. In the second order region (which for the results given in Table 8.3 includes $p = 1$) the performance is very similar to the upwind-2 scheme. The performance changes smoothly between the two levels in the transitional region, although there may be values for the which the performance is

N	Upwind-2			E-O-2 ($p = 1$)		
	Δt	No. Steps	CPU time/s	Δt	No. Steps	CPU time/s
10	0.083	47	0.995×10^{-2}	0.078	51	0.129×10^{-1}
20	0.067	59	0.170×10^{-1}	0.074	50	0.156×10^{-1}
40	0.027	142	0.795×10^{-1}	0.036	100	0.710×10^{-1}
80	0.015	249	0.276	0.016	242	0.340
160	0.0074	470	0.116×10^{1}	0.0074	482	0.130×10^{1}
320	0.0032	976	0.450×10^{1}	0.0035	1035	0.532×10^{1}
640	0.0015	2492	0.235×10^{2}	0.0017	1904	0.189×10^{2}
1280	0.0008	3729	0.699×10^{2}	0.0008	3786	0.775×10^{2}

Table 8.3: Time stepping for the upwind-2 and E-O-2 ($p = 1$) schemes for problem 6 using optimal time steps

unexpectedly low.

Finally we consider the performance of the time stepping iteration when using the high order TVD schemes. Table 8.4 gives examples of the performances of these schemes for test problem 6. The Yee-Roe-Davis scheme is about twice as expensive in CPU time as the previous scalar schemes considered. This is partly due to the extra overhead in computing the limiter functions, but the method also seems to inherently require more iterations to converge. The time stepping iteration for the Harten-Yee scheme succeeds when the number of grids points is small, although it is rather expensive. As the number of grid-points becomes larger the iteration appears not to converge for any choice of time step, no matter how small. The residual decreases and then becomes trapped in some limit cycle. Such an effect is called residual plateauing and is common in steady state calculations when using schemes with nonlinear limiters. Techniques to prevent this are discussed in [3]. The method fails on all the tested problems for larger numbers of grid-points.

Having investigated the time stepping iteration we now consider other possibly more efficient methods for solving the difference equations. We start by looking at Newton's method.

	Yee-Roe-Davis			Optimum/Harten-Yee		
N	Δt	No. Steps	CPU time/s	Δt	No. Steps	CPU time/s
10	0.086	119	0.325×10^{-1}	0.123	76	0.179×10^{-1}
20	0.026	175	0.748×10^{-1}	0.045	387	0.175
40	0.013	390	0.333	0.0103	1459	0.132×10^{1}
80	0.0063	664	0.128×10^{1}	Fails to converge for all Δt		
160	0.0031	1357	0.479×10^{1}	Fails to converge for all Δt		
320	0.0015	1967	0.145×10^{2}	Fails to converge for all Δt		
640	0.00075	5415	0.857×10^{2}	Fails to converge for all Δt		
1280	0.00036	8247	0.253×10^{3}	Fails to converge for all Δt		

Table 8.4: Time stepping for the high order TVD schemes for problem 6 using optimal time steps

8.2 Newton's Method

The nonlinear system of difference equations can be written in vector notation as

$$\boldsymbol{\mathcal{T}}(\boldsymbol{h}) = 0, \tag{8.3}$$

where $\boldsymbol{\mathcal{T}}(\boldsymbol{h}) = (\mathcal{T}_1 h, \mathcal{T}_2 h, \ldots, \mathcal{T}_{N-1} h)^T$, $\boldsymbol{h} = (h_1, h_2, \ldots, h_{N-1})^T$ and $h_0 = h_{-1} = \gamma_0$, $h_N = h_{N+1} = \gamma_1$. The time stepping iteration is essentially a Picard iteration applied to this system. Such methods only give a linear convergence rate, i.e. the residual is inversely proportional to number of iterations. Newton's method however is well known to give a quadratic convergence rate, i.e. the residual is inversely proportional to the number of iterations squared. The drawback of Newton's method is that in general, global convergence is not obtained, i.e. convergence will not occur for all initial guesses of the solution. The theory of Newton's method and other related methods can be found in [46].

Applying Newton's method to the system of difference equations yields the following algorithm:

$$\boldsymbol{h}^{n+1} = \boldsymbol{h}^n + s^n \boldsymbol{d}^n, \tag{8.4}$$

where $\boldsymbol{d}^n = (d_1^n, d_2^n, \ldots, d_{N-1}^n)^T$ solves the linear system

$$\boldsymbol{\mathcal{T}}'(\boldsymbol{h}^n)\boldsymbol{d}^n = -\boldsymbol{\mathcal{T}}(\boldsymbol{h}^n), \tag{8.5}$$

and $\boldsymbol{\mathcal{T}}'$ denotes the Jacobian of the system (8.3). Strictly speaking this is only Newton's method if we take $s^n = 1$ at every iteration. However it is found that the method often fails due to components of the iterate becoming negative. The method is found to be much more robust if at each iteration the value of s^n is chosen to prevent this. We use the formula

$$s^n = \min_{0<j<N}\{s_j^n\}, \tag{8.6}$$

where

$$s_j^n = \begin{cases} 1 & \text{if} \quad d_j^n \geq 0 \\ \min\{-s\dfrac{h_j^n}{d_j^n}, 1\} & \text{if} \quad d_j^n < 0. \end{cases}$$

Essentially s^n is taken as unity unless this would cause the solution to become negative, in which case it is taken as just below the largest value which gives a non-negative iterate. The parameter $0 \ll s \leq 1$ is chosen to stop the solution from approaching too close to zero in one iteration and is taken as 0.95. Other approaches are possible. For example Newton's method is a steepest descent method for the functional $\boldsymbol{\mathcal{T}}(\boldsymbol{h})^T\boldsymbol{\mathcal{T}}(\boldsymbol{h})$. Because of this it is always possible at each iteration to find a value $0 < s^n < 1$ such that this functional decreases. This may lead to a method that is globally convergent. The value of s^n can be computed using a line search algorithm, but this increases the CPU overhead considerably since it may require many evaluations of the function $\boldsymbol{\mathcal{T}}$. Such methods are described in [46] and [51].

For $\mathcal{T}$ of the form 5.20 the Jacobian is given by

$$\boldsymbol{\mathcal{T}}'(\boldsymbol{h}) = \begin{pmatrix} q_1 & r_1 & & & & & \\ p_2 & q_2 & r_2 & & & & \\ & \ddots & \ddots & \ddots & & & \\ & & p_j & q_j & r_j & & \\ & & & \ddots & \ddots & \ddots & \\ & & & & p_{N-2} & q_{N-2} & r_{N-2} \\ & & & & & p_{N-1} & q_{N-1} \end{pmatrix}, \tag{8.7}$$

where

$$\begin{aligned} p_j &= -\frac{g_v(h_j, h_{j-1})}{\Delta x}, \\ q_j &= \frac{g_v(h_{j+1}, h_j) - g_u(h_j, h_{j-1})}{\Delta x} + D_h(x_j, h_j), \\ r_j &= \frac{g_u(h_{j+1}, h_j)}{\Delta x}. \end{aligned}$$

The Jacobian does not strictly exist in the case of Godunov and the first-order upwind schemes because of the switching, however on the curves where the function g is not differentiable, either the partial derivatives from the left or right can be used. The Jacobian exists at all points for the Engquist-Osher and Lax-Friedrichs schemes since the numerical flux functions are differentiable in these cases. For monotone schemes (i.e. under the conditions of Theorem 9 with $\boldsymbol{\alpha} \leq \boldsymbol{h} \leq \boldsymbol{\beta}$) we have that $g_u \leq 0$, $g_v \geq 0$ and $D_h > 0$ so that p_j, $r_j \leq 0$ and $q_j > 0$. We observe that

$$|p_2| = \frac{g_v(h_2, h_1)}{\Delta x} < |q_1|,$$

$$|p_{j+1}| + |r_{j-1}| = \frac{g_v(h_{j+1}, h_j) - g_u(h_j, h_{j-1})}{\Delta x} < |q_j| \qquad j = 2, \ldots, N-2,$$

$$|r_{N-2}| = -\frac{g_u(h_{N-1}, h_{N-2})}{\Delta x} < |q_{N-1}|,$$

which demonstrates that the transpose of the Jacobian is strictly diagonally dominant and the Jacobian itself is non-singular. In fact the Jacobian is a special type of matrix known as an *M-matrix*. Properties of this class of matrices are discussed in [46]. One property is that all entries of the inverse matrix are non-negative.

Table 8.5 illustrates the performance of the Newton algorithm applied to the Engquist-Osher scheme for problem 6. The method gives a minimum of 50% reduction over the time stepping iteration and in most cases more. For example with 1280 grid points Newton's method is more than eight times more efficient. As well as being significantly faster, Newton's method has the advantage that the performance is not dependent on the choice of some parameter such as the time step in the time stepping iteration. Newton's method combined with the Engquist-Osher scheme is found to give a very good performance for all the problems 1-8 as long as the number of grid-points is not too large. For problems without hydraulic jumps the iteration count is found to be almost independent of the number of grid-points. For problems

N	E-O		E-O-2	
	No. iter	**CPU time/s**	**No. iter**	**CPU time/s**
10	9	0.372×10^{-2}	17	0.752×10^{-2}
20	10	0.524×10^{-2}	31	0.221×10^{-1}
40	30	0.308×10^{-1}	30	0.382×10^{-1}
80	44	0.691×10^{-1}	44	0.112
160	61	0.252	66	0.369
320	83	0.525	158	0.167×10^{1}
640	141	0.243×10^{1}	144	0.304×10^{1}
1280	267	0.728×10^{1}	553	0.236×10^{2}

Table 8.5: Performance of Newton's method for the E-O and E-O-2 ($p = 1$) schemes for problem 6

with jumps the method can fail to converge when the number of grid-points is high. For example in problem 5 it fails for $N = 1280$. This can be avoided by solving the scheme on a sequence of grids of increasing resolution. The solution is transferred from grid to grid using linear interpolation. This approach ensures that on each grid the initial guess is closer to the final solution and hence is more likely to be within the radius of convergence. This technique also can also improve efficiency, since fewer Newton iterations are required on the more expensive finer grids. Table 8.6 illustrates the performance of such an approach. For each value of N we started by solving on the grid $N = 10$ and then repeatedly doubled the grid until reaching the final grid. On each intermediate grid the solution of the difference equations is only solved to a tolerance of half the magnitude of that used on the final grid. The number of iterations given in the table correspond to the number of iterations taken on the final grid. The results show a significant improvement in efficiency when N becomes large.

Combining Newton with the Godunov or first order upwind schemes gives a similar level of performance to that for Engquist-Osher. For these schemes the function $\boldsymbol{\mathcal{T}}(\boldsymbol{h})$ is not differentiable on manifolds of the solution space, and this might be expected to be detrimental to Newton's method. The method often fails to converge

N	E-O		E-O-2	
	No. iter	**CPU time/s**	**No. iter**	**CPU time/s**
20	6	0.920×10^{-2}	8	0.151×10^{-1}
40	7	0.156×10^{-1}	7	0.276×10^{-1}
80	8	0.332×10^{-1}	10	0.579×10^{-1}
160	9	0.740×10^{-1}	18	0.109
320	9	0.142	10	0.314
640	100	0.158×10^{1}	11	0.514
1280	11	0.181×10^{1}	12	0.102×10^{1}

Table 8.6: Performance of Newton's method for the E-O and E-O-2 ($p = 1$) schemes for problem 6 using grid refinement

at a lower number of grid-points than for the Engquist-Osher scheme. For example for problem 5 both Godunov and the first-order upwind schemes fail for $N = 640$, unlike the Engquist-Osher scheme which does converge. Again this situation can in most cases be improved by the use of a grid refining technique as described above.

In the case of the upwind-2 scheme the function $\boldsymbol{\mathcal{T}}(\boldsymbol{h})$ does not even depend continuously on its variables, let alone be differentiable. Consequently for almost all problems that contain transitions and almost all values of N, Newton's method does not converge. The second order modification to the Engquist-Osher scheme, however, maintains the differentiability of the function $\boldsymbol{\mathcal{T}}(\boldsymbol{h})$, and hence Newton's method is expected to be well behaved. The Jacobian in this case is given by (8.7), where in the notation of Appendix A

$$\begin{aligned} p_j &= -\frac{1}{\Delta x}\frac{\partial \hat{g}_{j-\frac{1}{2}}}{\partial h_{j-1}}, \\ q_j &= \frac{1}{\Delta x}\left(\frac{\partial \hat{g}_{j+\frac{1}{2}}}{\partial h_j} - \frac{\partial \hat{g}_{j-\frac{1}{2}}}{\partial h_j}\right) + D_h(x_j, h_j) \\ &= p_{j+1} + r_{j-1} + D_h(x_j, h_j), \\ r_j &= \frac{1}{\Delta x}\frac{\partial \hat{g}_{j+\frac{1}{2}}}{\partial h_{j+1}}. \end{aligned}$$

Appendix A shows that if $\alpha \leq \boldsymbol{h} \leq \beta$ and p satisfies (7.6) then the off-diagonal elements of the Jacobian are non-negative and the diagonal elements are positive.

Hence we have

$$|q_j| = |p_{j+1}| + |r_{j-1}| + D_h(x_j, h_j) > |p_{j+1}| + |r_{j-1}|.$$

Hence the transpose of the Jacobian is strictly diagonally dominant and the Jacobian itself is non-singular, and as for the first order monotone schemes the Jacobian is an M-matrix.

Newton's method applied to the second order Engquist-Osher scheme is in practice found to be well-behaved and Table 8.6 shows the performance for problem 1 with $p = 1$. At best for this example the method converges in roughly the same number of iterations as for the first order scheme and at worst takes three times as many iterations. In the worst case, the method still yields a considerable improvement over the time stepping iteration. The author of [33] suggests that the method may fail if the parameter p is too high (violating (7.6)). This is not found to be the case. For some test cases there are found to be isolated values of p for which the method does not converge, however these always appear to be in the range corresponding to the transition between first order and second order. In general the number of iterations required reaches its maximum in this transition range and there have been no difficulties encountered by taking p well into the second order range. As for the first order scheme, the method can fail as the number of grid-points becomes large. A grid refining mechanism can again in many cases remedy this and also help for awkward values of p. The possible gain in efficiency obtainable by using a grid refining technique can be seen by comparing Tables 8.6 and 8.5

Finally we consider the application of Newton's method to the high order TVD schemes. The non-standard algebraic form of the limiter functions means that it is impractical to compute the full Jacobian of the system. In any case these limiter functions are not even continuous. We choose therefore only to compute an approximation to the Jacobian. Before calculating the Jacobian the limiter functions are set to zero, which is a standard technique for such schemes. This approach is used in [72] to linearise implicit schemes, a topic discussed in the next section. For the two high order TVD schemes considered, the approximate Jacobian corresponds to the Jacobian of the system for the first order upwind scheme (i.e. $\mathcal{T}' \approx (\mathcal{T}^{\mathrm{FOU}})'$). This approach also has the advantage that the Jacobian is tri-diagonal rather than

having a band-width of five, as would be the case if one attempted to use the full Jacobian. We expect the fact that only an approximate Jacobian is used to reduce the performance of the Newton's method. Applying the method to the Yee-Roe-Davis scheme and the Harten-Yee scheme we find that convergence is obtained if the number of grid-points is very small, although the method is more expensive than for the previous schemes such as the Engquist-Osher scheme. For even a moderate number of grid points and even for solutions without transitions, the method fails to converge for the Yee-Roe-Davis scheme. The Harten-Yee scheme converges for a higher number of grid-points, however for problems with hydraulic jumps fails for significantly smaller N than for example the Engquist-Osher scheme. The robustness of the method can in some cases be improved by use of a grid refining approach, but the method is not really suited to solving these schemes.

8.3 The Implicit Time Stepping Iteration

We now consider a generalisation of the time stepping algorithm. The particular implementation of the algorithm relates it closely to the Newton algorithm described in the previous section. The efficiency of the time stepping method is restricted by the fact that the time step must satisfy some CFL condition in order for the iterative process to be stable (i.e. converge). For steady state computations a standard method for relaxing or even removing altogether the CFL restriction is to use an implicit time stepping iteration. Such methods are discussed in section 3.6 and can be written as follows:

$$\boldsymbol{R}(\boldsymbol{h}^{n+1}) = \frac{\boldsymbol{h}^{n+1} - \boldsymbol{h}^n}{\Delta t} + \theta \boldsymbol{\mathcal{T}}(\boldsymbol{h}^{n+1}) + (1-\theta)\boldsymbol{\mathcal{T}}(\boldsymbol{h}^n) = 0, \qquad (0 < \theta \leq 1). \quad (8.8)$$

Note that in the case $\theta = 0$ this degenerates to the original explicit time stepping. In the case $\theta \neq 0$ the solution at the next time level can only be determined by solving the nonlinear system of equations (8.8). This can be carried out using exactly the same algorithm as described in the previous section. The complete method can now be written as follows:

$$\boldsymbol{h}^{n+1,k+1} = \boldsymbol{h}^{n+1,k} + s^{n+1,k}\boldsymbol{d}^{n+1,k}, \qquad k = 1, 2, \ldots \quad (8.9)$$

where $\boldsymbol{d}^{n+1,k} = (d_1^{n+1,k}, d_2^{n+1,k}, \ldots, d_{N-1}^{n+1,k})^T$ solves the linear system

$$\boldsymbol{\mathcal{R}}'(\boldsymbol{h}^{n+1,k})\boldsymbol{d}^{n+1,k} = -\boldsymbol{\mathcal{R}}(\boldsymbol{h}^{n+1,k}), \tag{8.10}$$

$$\boldsymbol{\mathcal{R}}' = \frac{I}{\Delta t} + \theta \boldsymbol{\mathcal{T}}',$$

$$s^{n+1,k} = \min_{0<j<N}\{s_j^{n+1,k}\},$$

and we take

$$s_j^{n+1,k} = \begin{cases} 1 & \text{if} \quad d_j^{n+1,k} \geq 0 \\ \min\{-s\dfrac{h_j^{n+1,k}}{d_j^{n+1,k}}, 1\} & \text{if} \quad d_j^{n+1,k} < 0. \end{cases}$$

We take the most obvious choice of initial guess, $\boldsymbol{h}^{n+1,0} = \boldsymbol{h}^n$, and set $\boldsymbol{h}^{n+1} = \boldsymbol{h}^{n,k+1}$ if the iteration converges to the required tolerance, i.e.

$$\sqrt{\frac{\boldsymbol{\mathcal{R}}(\boldsymbol{h}^{n+1,k})^T \boldsymbol{\mathcal{R}}(\boldsymbol{h}^{n+1,k})}{N-1}} < tol'.$$

For the monotone schemes and the E-O-2 scheme it was shown in the previous section that, if certain assumptions hold, then $\boldsymbol{\mathcal{T}}'$ must be nonsingular because the transpose of this matrix is strictly diagonally dominant. It follows that if $\boldsymbol{\mathcal{T}}'$ is strictly diagonally dominant, then

$$(\boldsymbol{\mathcal{R}}')^T = \frac{I}{\Delta t} + \theta(\boldsymbol{\mathcal{T}}')^T,$$

is also strictly diagonally dominant, since we are only adding positive terms to the positive diagonal entries of $\boldsymbol{\mathcal{T}}'$.

It may appear that at each time step we may be required to perform as much work as completely solving the system of difference equations using the Newton algorithm. However if the time step is not too large, the initial guess for the solution for each internal iteration (which is the solution at the current time level) is likely to be close to the solution of system (8.8) and so convergence may occur in a small number of internal iterations. It is well known and also observed in practice that the best performance occurs for the case $\theta = 1$, where for certain numerical flux functions the method is unconditionally TVD. In the case $\theta = 1$ the algorithm can be seen to approach the purely Newton algorithm of the previous section as the time step Δt tends towards infinity. Thus for cases where the Newton algorithm converges we

expect the implicit algorithm to have no restriction on the time step. This is found to be the case in practice. A measure of the performance for this method is given by the number of Jacobian inversions required for convergence to occur. The number of inversions required decreases as the time step increases and approaches the number of inversions required for the Newton algorithm as Δt becomes very large. Even for large time steps where the number of Jacobian inversions required for the two methods are almost identical, the implicit algorithm is significantly more expensive due to the expense of setting up the internal iteration.

The implicit time stepping algorithm has an advantage over Newton for some of the cases where the Newton algorithm fails. In many of these cases it is possible to choose a time step small enough such that the iteration converges. The implicit time stepping was applied to the upwind-2 scheme and is found to converge as long as the time step is small enough. However the time step is required to be so small that the method does not perform even as well as the explicit method. Furthermore spurious solutions are encountered for some values of the time step. These spurious solutions, which are unrelated to the physical solution of the problem, vary depending on the time step and so the system of difference equations appears to have many solutions. Such spurious solutions have not been encountered using the explicit time stepping algorithm.

Again, for the five point TVD schemes only the approximate Jacobian is used for implicit time stepping. The method is found to converge in all the cases tested, provided that the time step is small enough. This is better than both the explicit time stepping and Newton's methods which have difficulties with either the Yee-Roe-Davis or Harten-Yee schemes for large N.

It is found that the performance of the implicit method is dependent on the tolerance placed on the convergence of the internal iteration. The higher this tolerance, the more efficient the method is found to be. In fact the best performance is obtained when the tolerance is so high that at most one Newton iteration is performed at each time step. If we intend to perform only one Newton iteration at each time step then the method can be simplified to.

$$\boldsymbol{h}^{n+1} = \boldsymbol{h}^n + s^n \boldsymbol{d}^n, \tag{8.11}$$

where $\boldsymbol{d}^n = (d_1^n, d_2^n, \ldots, d_{N-1}^n)^T$ solves the linear system

$$\boldsymbol{\mathcal{R}}'(\boldsymbol{h}^n)\boldsymbol{d}^n = -\boldsymbol{\mathcal{R}}(\boldsymbol{h}^n), \tag{8.12}$$

and s^n is taken as in Newton's method. Again for $\theta = 1$ this method approaches the Newton algorithm as Δt grows large. The method is known as a linearised implicit scheme (see section 3.6) since it can also be derived using Taylor's expansions to linearise the implicit part of the operator. The linearised method is more efficient to implement than the method allowing a variable number of inner iterations, and each iteration requires a very similar expenditure to one iteration of the Newton algorithm. Table 8.7 shows the performance of the linearised method for the Engquist-Osher

	E-O			**E-O-2** ($p = 1$)		
N	Δt	**No. iter.**	**CPU time/s**	Δt	**No. iter.**	**CPU time/s**
10	35.34	9	0.185×10^{-2}	3.0	13	0.472×10^{-2}
120	13.4	10	0.470×10^{-2}	1.0	16	0.112×10^{-1}
40	2.4	31	0.299×10^{-1}	1.0	36	0.308×10^{-1}
80	5.0	44	0.729×10^{-1}	1.0	56	0.106
160	13.5	60	0.198	1.0	97	0.369
320	16.0	82	0.646	1.0	180	0.121×10^{1}
640	39.0	141	0.191×10^{1}	21.0	138	0.332×10^{1}
1280	84.0	266	0.749×10^{1}	36.0	260	0.116×10^{2}

Table 8.7: Linearised implicit algorithm for the E-O and E-O-2 ($p = 1$) schemes for problem 6

scheme and its second order modification ($p = 1$) for problem 6. In each case performance is given for a time step which gives close to the optimum convergence rate. By comparing Tables 8.7 and 8.5 it can be seen that the linearised method and Newton's method are indeed very similar in performance. Of course for the linearised method the performance is dependent on the choice of time step and there is no way to predict the optimal value in advance. However, it is found that good performance is obtained over a much wider range of time steps than for the explicit time stepping method. There also appears to be no limit on the size of the time step

in many cases. In the case of the high order TVD schemes, the linearised method

	Yee-Roe-Davis			Harten-Yee		
N	Δt	No. Steps	CPU time/s	Δt	No. iter.	CPU time/s
10	4.0	36	0.137×10^{-1}	2.35	25	0.741×10^{-2}
20	1.0	174	0.108	0.51	37	0.238×10^{-1}
40	0.1	56	0.655×10^{-1}	0.42	45	0.547×10^{-1}
80	0.06	68	0.159	0.19	84	0.218
160	0.049	115	0.543	0.13	157	0.778
320	0.027	124	0.119×10^{1}	0.053	300	0.308×10^{1}
640	0.071	63	0.112	0.021	605	0.141×10^{2}
1280	0.013	271	0.124×10^{2}	0.013	1072	0.497×10^{2}

Table 8.8: Linearised implicit algorithm for the Y-R-D and H-Y schemes for problem 6

gives the best performance of all the methods previously encountered. This is not surprising since this method is exactly the method (except for the modification to prevent negative values) recommended by the authors of [72] for computing steady solutions of these schemes. The method does not perform to the same standard as for the Engquist-Osher scheme. To start with it is more expensive, secondly convergence only occurs below a certain value of time step and lastly the efficiency is more sensitive to the choice of time step. Table 8.8 illustrates the performance of the method for the two schemes for problem 6, the performance being comparable to that of Newton's method when Newton converges and greatly superior to the explicit time stepping.

The idea of a linearised implicit scheme has been successfully applied in [17] to Roe's approximate Riemann solver with the aim of efficiently computing steady solutions. Carrying this out is considerably more complex than for the scalar schemes, and it is difficult to see how this can be more efficient than for the scalar equivalents.

8.4 Conclusions

In this chapter the effectiveness of four different methods of solving the difference equations was considered. These were the explicit time-stepping iteration, Newton's method, the implicit time-stepping iteration and the linearised implicit time-stepping iteration. The Engquist-Osher and first-order upwind schemes were found to be particularly amenable to solution by Newton's method. Not only was this method the most efficient and robust, but also had the advantage that the efficiency and robustness did not depend on the appropriate choice of some parameter. However the method can fail to converge when the number of grid-points becomes large. This situation can often be remedied by solving the difference equations on finer and finer grids. This technique can also further increase the efficiency, even in cases where Newton's method would converge without it. Alternatively, since Newton's method corresponds to the linearised implicit method with an infinite time step, convergence can be obtained by using the latter method with a finite time step.

The only method suitable for the upwind-2 scheme, due to the discontinuous way in which the source term switches, was the explicit time-stepping iteration. On the other hand, Newton's method was almost as robust and efficient for the E-O-2 scheme as it was for the first order scheme and the same comments hold for improving the robustness at large N. For the high order TVD schemes, even the time stepping iteration failed to converge in many cases. However good performance was observed using the linearised implicit method. We conclude that the best strategy in order to obtain accuracy, efficiency and robustness is the combination of the E-O-2 scheme and Newton's iteration (and grid refinement for large values of N).

Chapter 9

Non-Prismatic Channels

9.1 Scalar Schemes

Thus far we have only considered numerical methods for the case of prismatic channels. The steady flow equation for a non-prismatic channel is given by

$$\frac{d}{dx}F(x,h) = D(x,h). \tag{9.1}$$

The difference between the non-prismatic case and the prismatic case is the additional explicit dependence of the quantity F on the distance along the channel x. In order to model this dependence numerically, we allow the numerical flux function to also depend on x. A possible first order accurate approximation of (9.1) is given by

$$\mathcal{T}_j h \equiv \frac{g(x_{j+q}, h_{j+1}, h_j) - g(x_{j+q-1}, h_j, h_{j-1})}{\Delta x} + D(x_j, h_j) = 0,$$

for any q, where $x_{j+q} = (j+q)\Delta x$ and

$$g(x,h,h) \equiv f(x,h) \equiv -F(x,h),$$

for all x and h. The parameter q controls how the "x" evaluation points of the numerical fluxes are staggered. If we consider the time accurate method

$$\frac{h_j^{n+1} - h_j^n}{\Delta t} + \mathcal{T}_j h^n = 0,$$

for the scalar PDE (4.6), then the term $g(x_{j+q}, h_{j+1}, h_j)$ approximates the time average flux across the cell interface at $x = x_{j+\frac{1}{2}}$. The choice $q = 1/2$ is therefore the

most natural. Other choices may be more appropriate for reasons of computational efficiency. If q is not an integer then certain quantities (such as the wetted area) will need be computed at both x_j and x_{j+q} for each j. Alternatively if we take $q = 0$ or $q = 1$ then these quantities are only required at the grid-points. The numerical flux functions described thus far in this thesis can, in general, be modified simply by adding the argument x to any evaluations of the function f or its derivatives. In the case of the Engquist-Osher scheme this leads to

$$g^{\text{E-O}}(x, u, v) = f_-(x, u) + f_+(x, v) + f(x, c), \tag{9.2}$$

where

$$\begin{aligned} f_-(x, u) &= \int_c^u \min\{f_h(x, s), 0\} ds, \\ f_+(x, u) &= \int_c^u \max\{f_h(x, s), 0\} ds \end{aligned}$$

and $c > 0$ is arbitrary. The Godunov numerical flux function is given by

$$g^{\text{God}}(x, u, v) = \begin{cases} \max\{f(x, w) : u \le w \le v\} & \text{for } u \le v \\ \min\{f(x, w) : v \le u \le v\} & \text{for } v \le u, \end{cases} \tag{9.3}$$

and first-order upwind numerical flux is given by

$$g^{\text{FOU}}(x, u, v) = \frac{1}{2} \left(f(x, u) + f(x, v) - |s|(u - v) \right), \tag{9.4}$$

where

$$s = \begin{cases} \dfrac{f(x, v) - f(x, u)}{v - u} & u \ne v \\ f_h(x, u) & u = v. \end{cases}$$

If at each cross-section there is a unique critical depth, $h_c(x)$, and the width does not approach zero as the depth becomes large, then these three upwind schemes have the property that:

$$\begin{aligned} h_i, h_{i+1} > h_c(x_{j+q}) &\implies g(x_{j+q}, h_{j+1}, h_j) = f(x_{j+q}, h_{j+1}), \\ h_i, h_{i+1} < h_c(x_{j+q}) &\implies g(x_{j+q}, h_{j+1}, h_j) = f(x_{j+q}, h_j). \end{aligned}$$

We look at the truncation error of the scheme to investigate which of the values of q gives the best approximation to the differential equations. In a region of the solution

where the flow is subcritical, i.e. $h_{j-1},\, h_j > h_c(x_{j+q-1})$ and $h_j,\, h_{j+1} > h_c(x_{j+q})$, the three schemes all reduce to

$$\frac{f(x_{j+q}, h_{j+1}) - f(x_{j+q-1}, h_j)}{\Delta x} + D(x_j, h_j) = 0,$$

with truncation error

$$\begin{aligned} \text{T.E.} &= \frac{\Delta x}{2}\left(h'' f_h + (h')^2 f_{hh} + 2q h' f_{hx} + (2q-1) f_{xx}\right) + O(\Delta x^2) \\ &= \frac{\Delta x}{2}\frac{d}{dx}\left(-D + 2(q-1) f_x\right) + O(\Delta x^2). \end{aligned} \tag{9.5}$$

In a region of the solution where the flow is supercritical, i.e. $h_{j-1},\, h_j < h_c(x_{j+q-1})$ and $h_j,\, h_{j+1} < h_c(x_{j+q})$, the three schemes reduce to

$$\frac{f(x_{j+q}, h_j) - f(x_{j+q-1}, h_{j-1})}{\Delta x} + D(x_j, h_j) = 0,$$

with truncation error

$$\begin{aligned} \text{T.E.} &= \frac{\Delta x}{2}\left(-h'' f_h - (h')^2 f_{hh} + 2(q-1) h' f_{hx} + (2q-1) f_{xx}\right) + O(\Delta x^2) \\ &= \frac{\Delta x}{2}\frac{d}{dx}\left(D + 2q f_x\right) + O(\Delta x^2). \end{aligned} \tag{9.6}$$

We see that the truncation error is $O(\Delta x)$ for all values of q and various terms may be eliminated by certain choices of q, but without necessarily leading to a reduction in the magnitude of the truncation error. We will see below that there is no single value of q that performs best in all situations.

Figures 9.1 to 9.6 show results for problems 9-14 (given in Appendix B) for the Engquist-Osher scheme with $N = 20$. All results are computed using the Newton algorithm described in section 8.2.

The solution to problem 9 (Figure 9.1) represents entirely subcritical flow, the depth profile is a hump which is symmetric about the center of the reach. The solution for $q = 1/2$ is clearly the most accurate, with the solution for $q = 0$ overestimating the maximum depth by a significant amount and the solution for $q = 1$ underestimating the maximum depth by a significant amount. The solutions were computed for a range of values of N, namely $N = 10, 20, \ldots, 640, 1280$. In all cases the solution for $q = 1/2$ is found to the most accurate in both the L_1 and L_∞ measures.

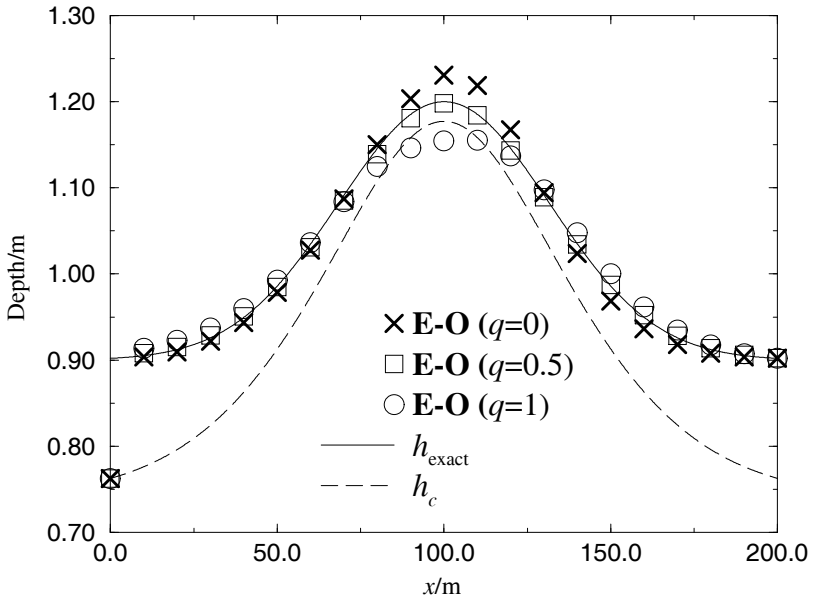


Figure 9.1: Results for Engquist-Osher problem 9 (Δx =10m).

The depth profile for problem 10 (Figure 9.2) is similar to the previous example, but in this case it corresponds to supercritical flow. It can be seen that the solution for $q = 0$ is now the most accurate. The solutions for $q = 1/2$ and $q = 1$ both underestimate the maximum by considerable amounts, with the solution for $q = 1$ being the least accurate. The value $q = 0$ is found to be the most accurate over all those values of N tried.

In the case of problem 11 (Figure 9.3) the flow is subcritical until approximately one third distance along the channel where it accelerates smoothly to supercritical flow. The choice $q = 0$ gives the best solution in the subcritical region of flow, whilst the solution for $q = 1/2$ is the most accurate in the supercritical region of flow. In terms of the L_1 error, the solution for $q = 1/2$ is found to be the most accurate for all the values of N considered.

The solutions for problem 12 are shown in Figure 9.4. The solution has a hydraulic jump at $x = 120$m and the solution for $q = 1/2$ is the most accurate in both the subcritical and supercritical regions of flow. Again this value gives the smallest L_1 error for all the values of N. One anomaly observed is that the error at $x = 80$m for $q = 1$, whilst initially greater than that for $q = 1/2$, decreases at a rate consistent with second order accuracy, and eventually becomes less than that for $q = 1/2$.

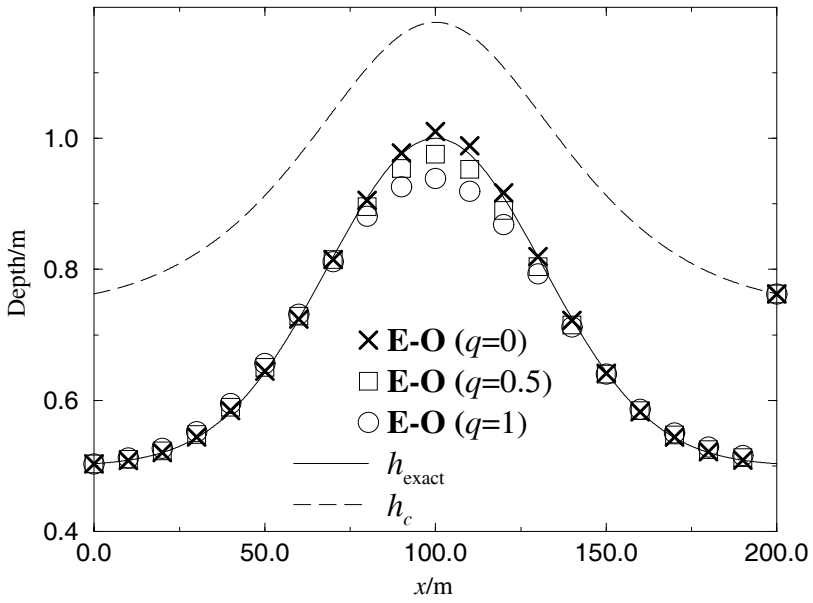


Figure 9.2: Results for Engquist-Osher problem 10 (Δx =10m).

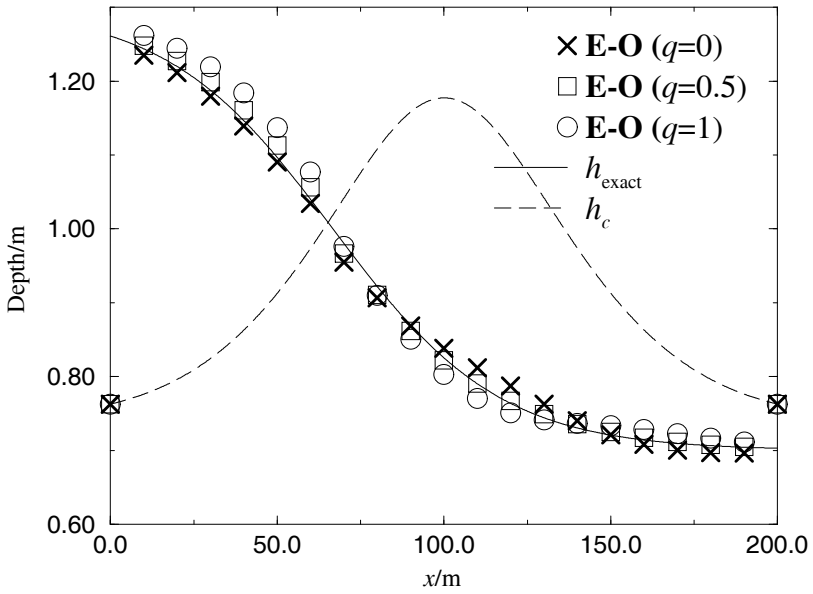


Figure 9.3: Results for the Engquist-Osher scheme for problem 11 (Δx =10m).

There is no obvious reason for this superconvergence.

The solution for problem 13 (Figure 9.5) as for problem 9 represents an entirely subcritical flow. However in this case the choice $q = 1$ yields the best solution (although it is not the most accurate at the second peak). This choice gives the lowest L_1 and L_∞ errors for all the values of N considered.

The final case is problem 14 (Figure 9.6) whose solution has two transitions, a

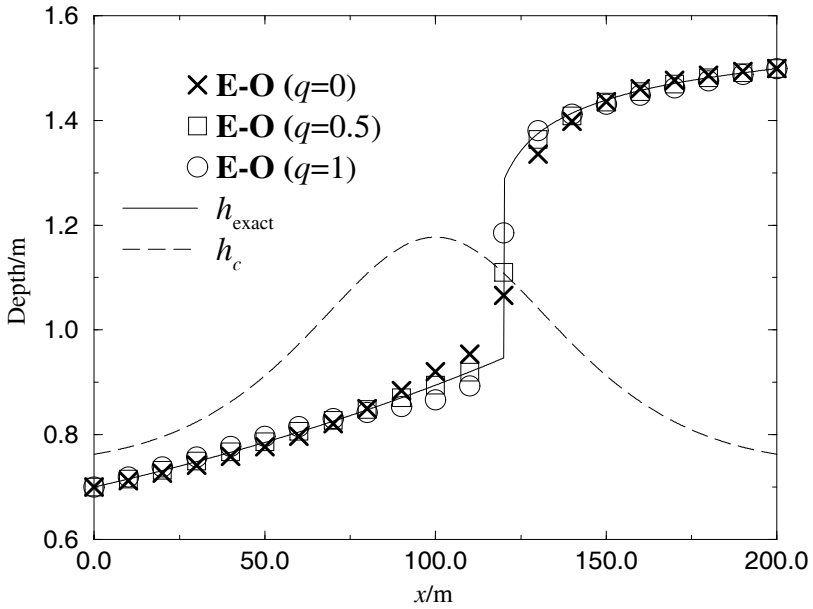


Figure 9.4: Results for the Engquist-Osher scheme for problem 12 (Δx =10m).

smooth transition at 1/8 distance and a hydraulic jump at 1/4 distance. The solution for $q = 1/2$ is the most accurate in the region of subcritical flow downstream of the jump. In fact the solutions for the other values of q are particularly poor. The solution for $q = 1/2$ remains the most accurate (in the L_1 measure) for all the values of N considered.

The choice $q = 1/2$ is found to give the most accurate solution in more cases than not, but there is no clear trend to allow us to predict when another choice will be more accurate. For example the solutions to problems 9 and 13 both represent subcritical flows; however in the first case $q = 1/2$ is the most accurate whilst in the second case $q = 1$ is the most accurate. In cases where the choice $q = 1/2$ does not give the most accurate solution, it is never found to give the least accurate solution. Almost identical conclusions to those given above can be made for the generalisation of the first-order upwind scheme, which is found to give very similar accuracy to the Engquist-Osher scheme.

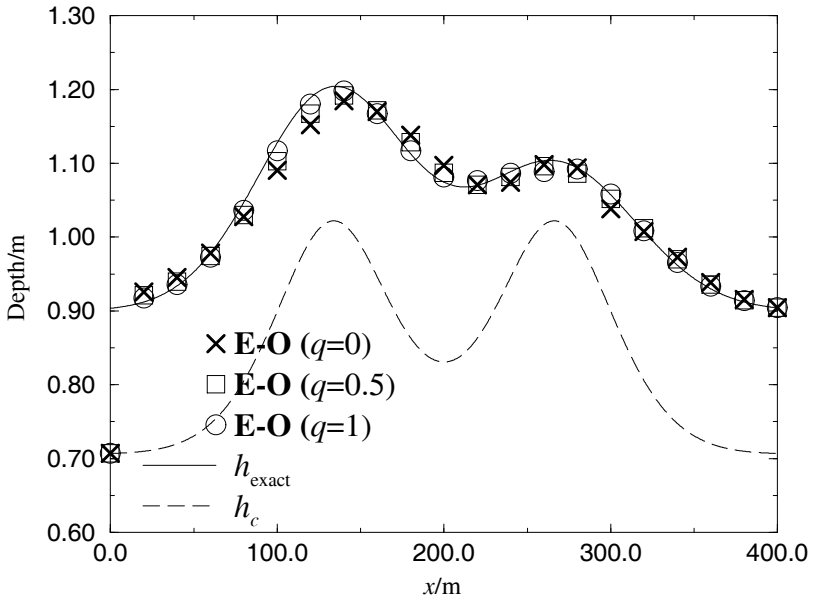


Figure 9.5: Results for the Engquist-Osher scheme for problem 13 (Δx =20m).

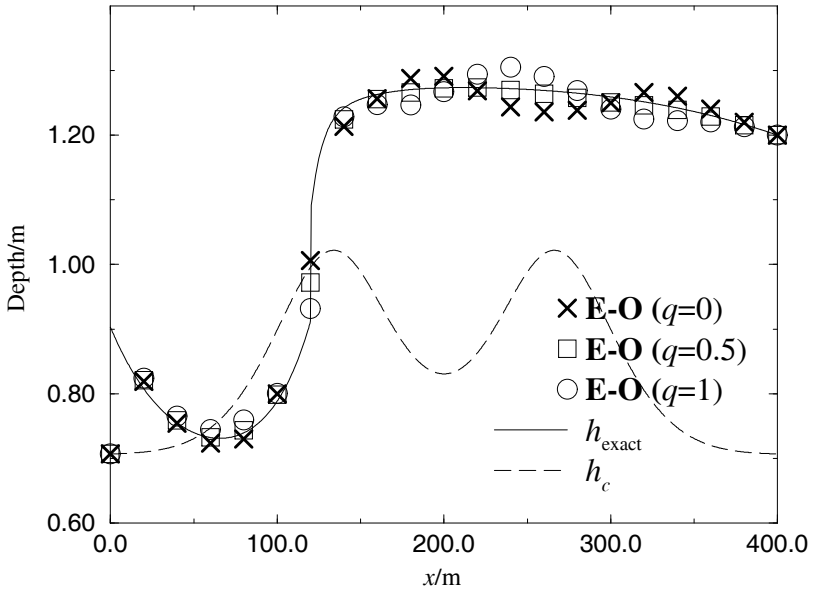


Figure 9.6: Results for the Engquist-Osher scheme problem 14 (Δx =20m).

The straightforward extension of the E-O-2 scheme to the non-prismatic case is given by

$$\begin{aligned}\mathcal{T}_j^{\text{E-O-2}} h &= \frac{g^{\text{E-O}}(x_{j+q}, h_{j+1}, h_j) - g^{\text{E-O}}(x_{j+q-1}, h_j, h_{j-1})}{\Delta x} \\ &\quad + \chi_j^- D_{j-1} + \chi_j^0 D_j + \chi_j^+ D_{j+1},\end{aligned}$$

where

$$\chi_j^- = \chi\left(\frac{p f_h(x_{j-1}, h_{j-1})}{\sqrt{\Delta x}}\right),$$

$$\chi_j^0 = 1 - \chi\left(\frac{pf_h(x_j,h_j)}{\sqrt{\Delta x}}\right) - \chi\left(\frac{-pf_h(x_j,h_j)}{\sqrt{\Delta x}}\right) = 1 - \chi_{j+1}^- - \chi_{j-1}^+,$$
$$\chi_j^+ = \chi\left(\frac{-pf_h(x_{j+1},h_{j+1})}{\sqrt{\Delta x}}\right),$$

and the function χ is given as before. For a prismatic channel the scheme gives second order accuracy. To see whether this remains the case for a non-prismatic channel we again consider the truncation error in regions of subcritical and regions of supercritical flow. In a region of subcritical flow (sufficiently far from being critical) the scheme reduces to

$$\frac{f(x_{j+q},h_{j+1}) - f(x_{j+q-1},h_j)}{\Delta x} + \frac{D(x_j,h_j) + D(x_{j+1},h_{j+1})}{2} = 0.$$

To obtain the truncation error for this scheme we simply add the terms

$$\begin{aligned}\frac{D(x_j,h_j) + D(x_{j+1},h_{j+1})}{2} - D(x_j,h_j) &= \frac{\Delta x}{2}\left(D_x + h'D_h\right) + O(\Delta x^2)\\ &= \frac{\Delta x}{2}\frac{d}{dx}D + O(\Delta x^2)\end{aligned}$$

to (9.5) to obtain the truncation error

$$T.E. = 2(q-1)\Delta x\frac{d}{dx}f_x + O(\Delta x^2).$$

Thus the scheme is only second order accurate in regions of subcritical flow if $q = 1$. In a region of supercritical flow (again sufficiently far from being critical) the scheme reduces to

$$\frac{f(x_{j+q},h_j) - f(x_{j+q-1},h_{j-1})}{\Delta x} + \frac{D(x_{j-1},h_{j-1}) + D(x_j,h_j)}{2} = 0.$$

We again obtain the truncation error by adding the terms

$$\begin{aligned}\frac{D(x_j,h_j) + D(x_{j-1},h_{j-1})}{2} - D(x_j,h_j) &= -\frac{\Delta x}{2}\left(D_x + h'D_h\right) + O(\Delta x^2)\\ &= -\frac{\Delta x}{2}\frac{d}{dx}D + O(\Delta x^2)\end{aligned}$$

to (9.6) to obtain the truncation error

$$T.E. = 2q\Delta x\frac{d}{dx}f_x + O(\Delta x^2).$$

Therefore the scheme is second order accurate in regions of supercritical flow in the case $q = 0$. We conclude that the approach of upwinding the source term does not

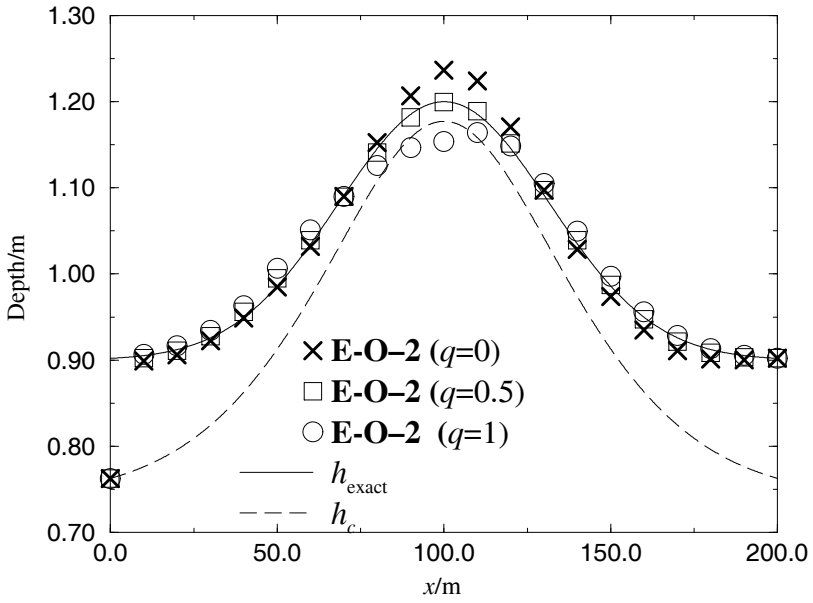


Figure 9.7: Results for the E-O-2 scheme ($p = 0.1$) for problem 9 (Δx =10m).

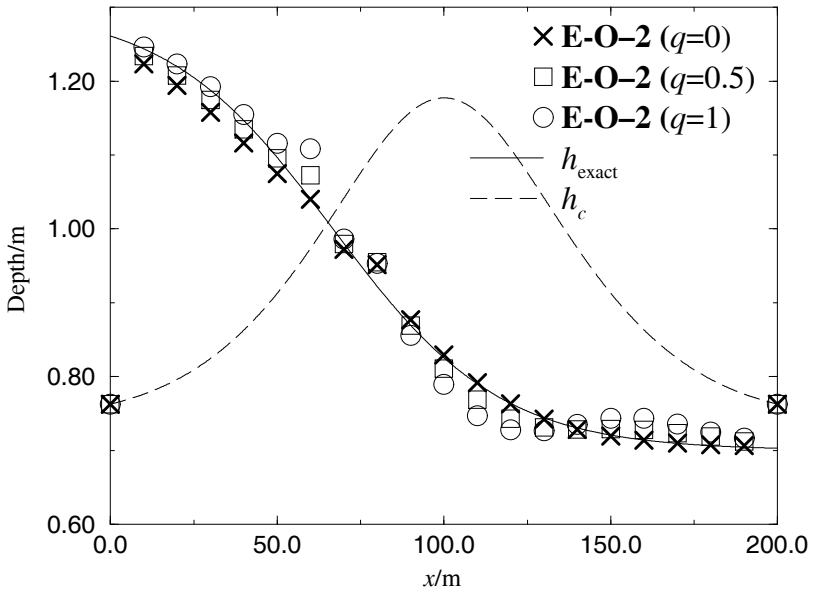


Figure 9.8: Results for the E-O-2 scheme ($p = 0.1$) for problem 11 (Δx =10m).

give second order accuracy in all regions of a transcritical flow. This suggests that some kind of switching of the value of q is required depending on the type of the flow. However, such a switching would destroy the conservative nature of the scheme.

Figures 9.7-9.10 show results for the E-O-2 scheme ($N = 20, p = 0.1$) for problems 9, 11, 12 and 14 respectively. The solution to problem 9 (Figure 9.7) is a subcritical flow, thus from the above analysis we expect that the choice $q = 1$, $p = 0.1$ will give

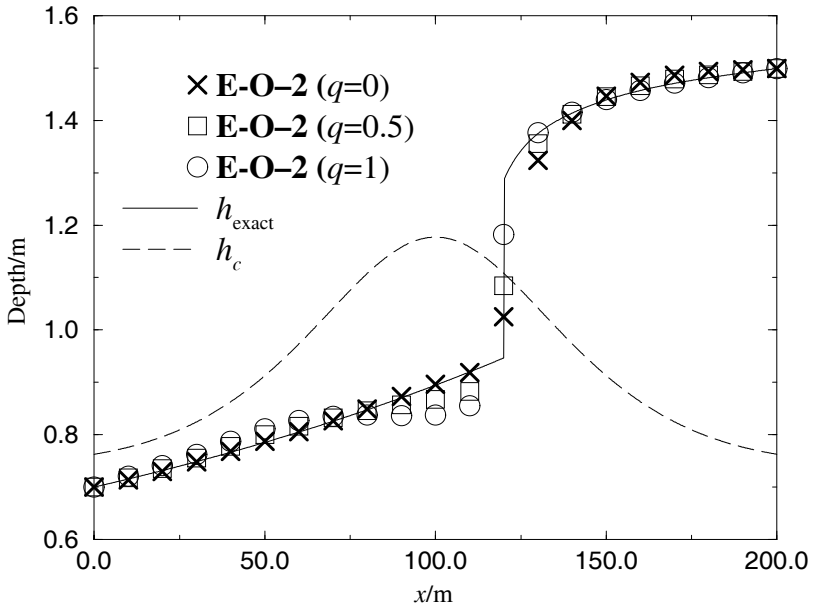


Figure 9.9: Results for the E-O-2 scheme ($p = 0.1$) for problem 12 (Δx =10m).

second order accuracy. However comparing this solution against the corresponding solution for the first-order scheme (Figure 9.1) we see that there is no visible improvement in the accuracy. Figure 9.15 shows the L_1 errors as a function of N for the E-O and E-O-2 schemes. We see that the errors for the E-O-2 scheme with $q = 1$ do not start to decrease at a rate consistent with second order accuracy until N is greater than 80, and the scheme remains less accurate than the E-O scheme ($q = 1/2$) until $N = 640$. The initially slow rate at which the errors decrease is undoubtedly due to the phantom transitions near the maxima of the depth profile.

For problem 11 (Figure 9.8) the errors in the subcritical flow region for the E-O-2 scheme with $q = 1$ are found to decrease with roughly second order accuracy, similarly for $q = 0$ the errors in the supercritical flow region are observed to decrease with roughly second order accuracy. However the quality of the overall solutions are in general inferior to those from the E-O scheme (Figure 9.3). The solutions from the E-O-2 scheme behave wildly at the smooth transition. This behaviour, which degrades the accuracy of the solution well away for the transition, is found to be sensitive to the value of the parameter p which controls the rate at which the source term discretisation switches across a transition. Reducing p improves the quality of the solution near the transition, but may also destroy the partial second order

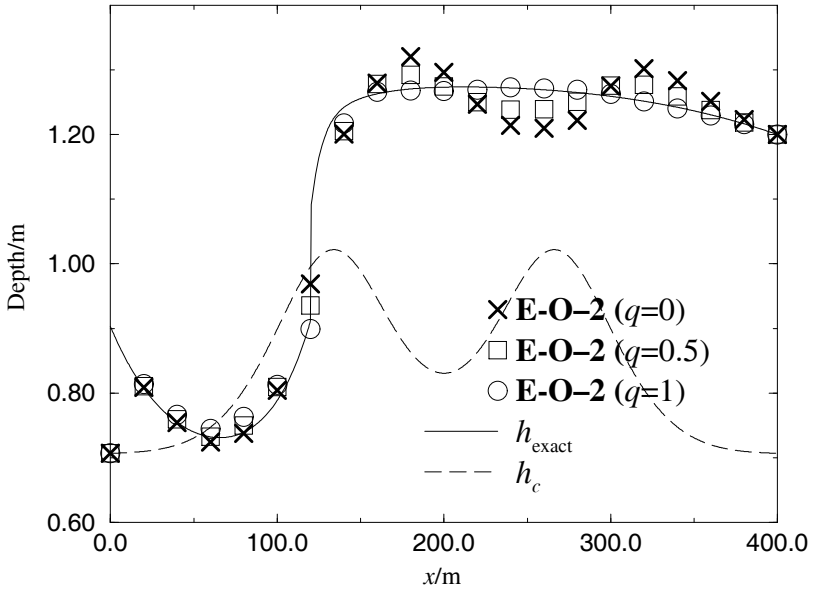


Figure 9.10: Results for the E-O-2 scheme ($p = 0.1$) for problem 14 (Δx =20m).

accuracy of the scheme. The L_1 errors of the E-O and E-O-2 schemes are shown in Figure 9.16. Clearly the E-O scheme with $q = 1/2$ is more accurate than the E-O-2 for all three values of q.

The E-O-2 scheme with $q = 0$ or $q = 1$ again demonstrates second order accuracy for problem 12 in the region of flow of the appropriate type. In this case, Figure 9.16 shows that the E-O-2 scheme with $q = 0$ gives the best L_1 accuracy out of all the schemes, for all the values of N considered.

The L_1 errors for problem 14 are shown in Figure 9.18 and in this case the E-O-2 scheme with $q = 1$ is the most accurate method. In Figure 9.10 we see that the solution for $q = 1$ is clearly the most accurate downstream of the jump, with the solutions for the other values of q significantly less accurate than the solutions for the E-O scheme.

We have encountered two main difficulties with the E-O-2 scheme. Firstly the scheme can perform badly across smooth transitions. Secondly any gain in accuracy over the E-O scheme in a region of one flow type is often at the expense of accuracy in a region of the opposite flow type. We conclude that it is only beneficial to use the E-O-2 scheme for problems where the flow is predominantly of one type, such as problem 14 where the E-O-2 scheme with $q = 1$ gives the best overall solution over the range of N. The generalisation of the upwind-2 scheme (7.5) is found to be

extremely problematic, since even the explicit time stepping iteration fails to converge more often than not.

9.2 Roe's Approximate Riemann Solver

We can use the same principle as in the previous section to extend Roe's approximate Riemann solver to the non-prismatic case. The generalised numerical flux is given by

$$\boldsymbol{g}^{\text{Roe}}_{j+\frac{1}{2}} = \frac{1}{2}\left(\boldsymbol{F}(x_{j+q}, \boldsymbol{w}_j) + \boldsymbol{F}(x_{j+q}, \boldsymbol{w}_{j+1}) - |\tilde{J}_{j+\frac{1}{2}}|(\boldsymbol{w}_{j+1} - \boldsymbol{w}_j)\right).$$

where now

$$\boldsymbol{F}(x_{j+q}, \boldsymbol{w}_{j+1}) - \boldsymbol{F}(x_{j+q}, \boldsymbol{w}_j) = \tilde{J}_{j+\frac{1}{2}}(\boldsymbol{w}_{j+1} - \boldsymbol{w}_j).$$

The scheme can be written as

$$\frac{\boldsymbol{w}_j^{n+1} - \boldsymbol{w}_j^n}{\Delta t} + \left(\tilde{J}^-_{j+\frac{1}{2}}\right)^n \frac{(\boldsymbol{w}_{j+1}^n - \boldsymbol{w}_j^n)}{\Delta x} + \left(\tilde{J}^+_{j-\frac{1}{2}}\right)^n \frac{(\boldsymbol{w}_j^n - \boldsymbol{w}_{j-1}^n)}{\Delta x} + \frac{\boldsymbol{F}(x_{j+q}, \boldsymbol{w}_j^n) - \boldsymbol{F}(x_{j+q-1}, \boldsymbol{w}_j^n)}{\Delta x} = \boldsymbol{D}_j^n.$$

At each time step, as well as distributing the increments $\left(\boldsymbol{\Phi}^{\pm}_{j+\frac{1}{2}}\right)^n$ (see section 3.8), we now need to add a quantity to each cell due to the new term

$$\frac{\boldsymbol{F}(x_{j+q}, \boldsymbol{w}_j^n) - \boldsymbol{F}(x_{j+q-1}, \boldsymbol{w}_j^n)}{\Delta x} \approx F_x(x_j, \boldsymbol{w}_j^n). \tag{9.7}$$

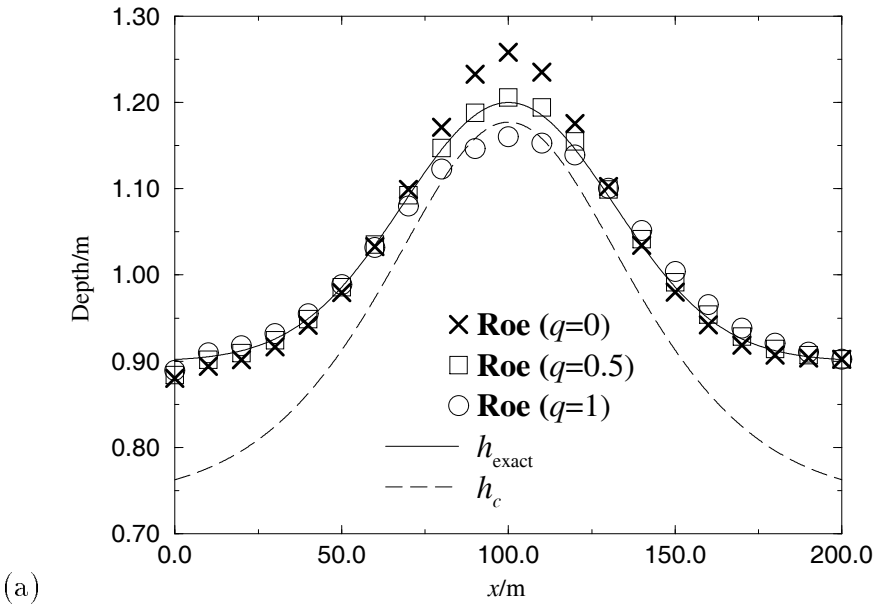
The approach of [53] for a rectangular channel of variable width is to use the approximation (9.7) and to absorb the term $F_x(x_j, \boldsymbol{w}_j^n)$ into the source term. This approach, however, renders the scheme non-conservative. No increment is added due to the term (9.7) at the boundary cells and this makes the treatment of the boundary conditions inadequate. This can be seen in the results by anomalies at boundaries where one of the flow variables is unspecified. Further work is required to improve this situation. The only other change to the algorithm from the prismatic case is that when calculating the wave speeds

$$\tilde{\lambda}_{j+\frac{1}{2},1} = \tilde{u}_{j+\frac{1}{2}} - \tilde{c}_{j+\frac{1}{2}}, \qquad \tilde{\lambda}_{j+\frac{1}{2},2} = \tilde{u}_{j+\frac{1}{2}} + \tilde{c}_{j+\frac{1}{2}},$$

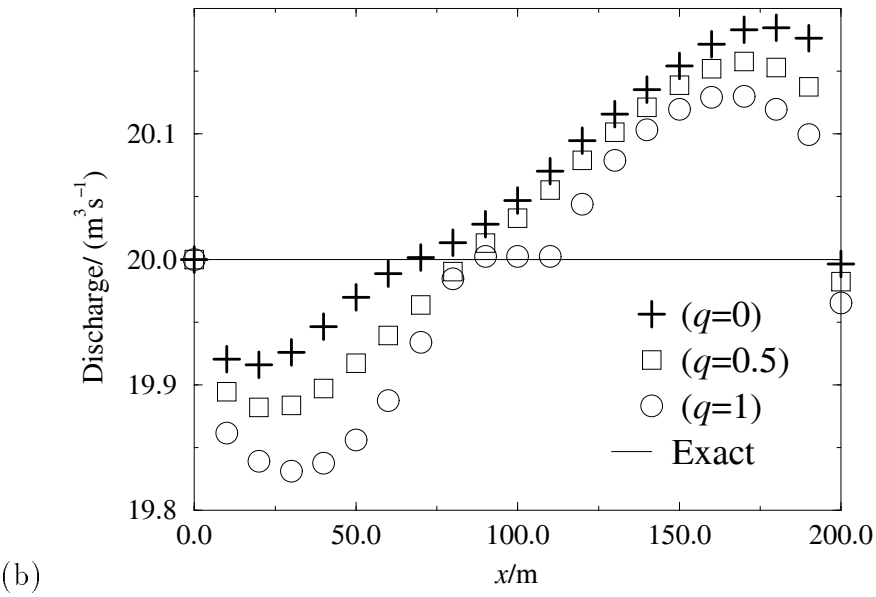
we use the modified formula

$$\left(\tilde{c}_{j+\frac{1}{2}}\right)^2 = \begin{cases} g\left(\dfrac{I_1(x_{j+q}, A_{j+1}) - I_1(x_{j+q}, A_j)}{A_{j+1} - A_j}\right) & A_j \neq A_{j+1} \\ \\ \dfrac{gA_j}{T(x_{j+q}, A_j)} & A_j = A_{j+1}, \end{cases}$$

where now the functions T and I_1 give there respective quantities as a function of cross-section and wetted area.



(a)



(b)

Figure 9.11: Roe's scheme for problem 9 (Δx =10m).

Figures 9.11 to 9.14 show results for Roe's scheme for problems 9, 11, 12 and 14 with $N = 20$, respectively. In each case both the depth and discharge fields are

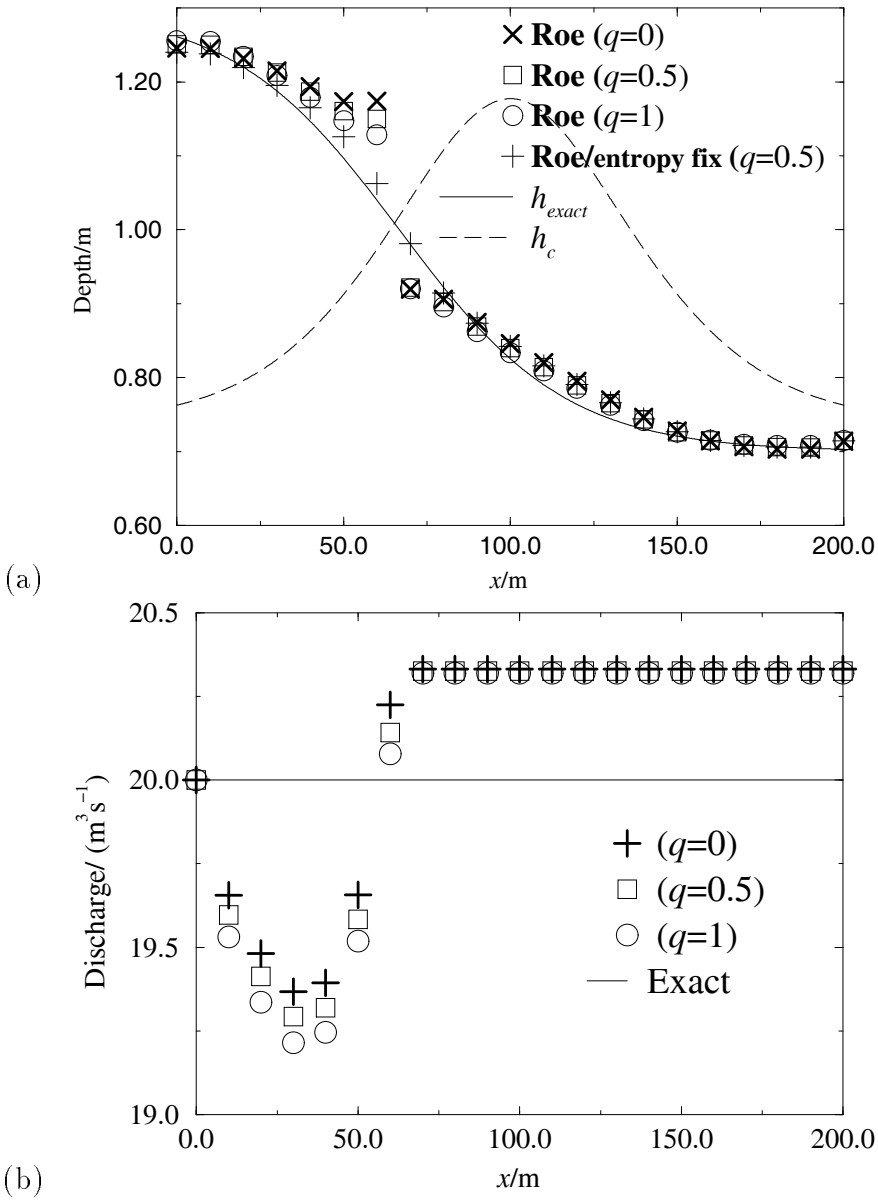


Figure 9.12: Roe's scheme for problem 11 (Δx =10m).

shown. For problem 9 the most accurate depth profile is for $q = 1/2$, although from the plot of the L_1 errors (Figure 9.15) it can be seen that this is less accurate than the Engquist-Osher scheme with the same value of q. The solutions at the smooth transition for problem 11 are completely wrong. The behaviour is similar to that of the E-O-2 scheme, but here it is much more pronounced. It appears as though the scheme has captured a jump rather than a smooth transition, although the solution converges to a smooth transition as the grid is refined. This situation is not as serious as it may appear, since it can be remedied simply by the addition of an entropy fix

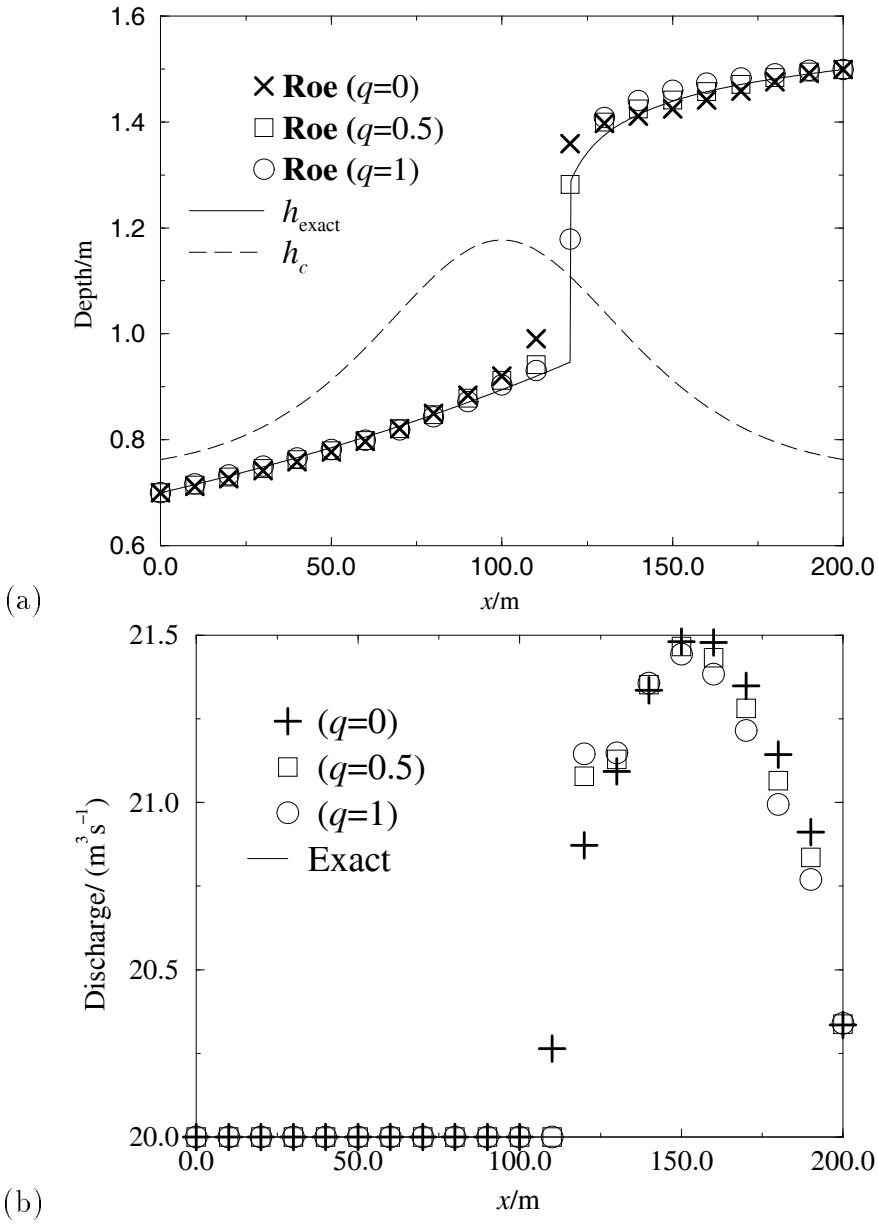


Figure 9.13: Roe's scheme for problem 12 (Δx =10m).

(see [30], Chapter 18) which prevents the formation of unphysical discontinuities. Results with the addition of an entropy fix are shown in Figure 9.12 for this case.

The results for problem 12 (Figure 9.13) can be seen to be very comparable to those from the Engquist-Osher scheme. Looking at the plot of the L_1 errors (Figure 9.17) we see in the case $q = 1$ that the error for Roe's scheme increases suddenly for no obvious reason after $N = 160$.

Roe's scheme is the least accurate of all the schemes in terms of L_1 errors for problem 14 and the scheme has even greater difficulty than the other schemes in the

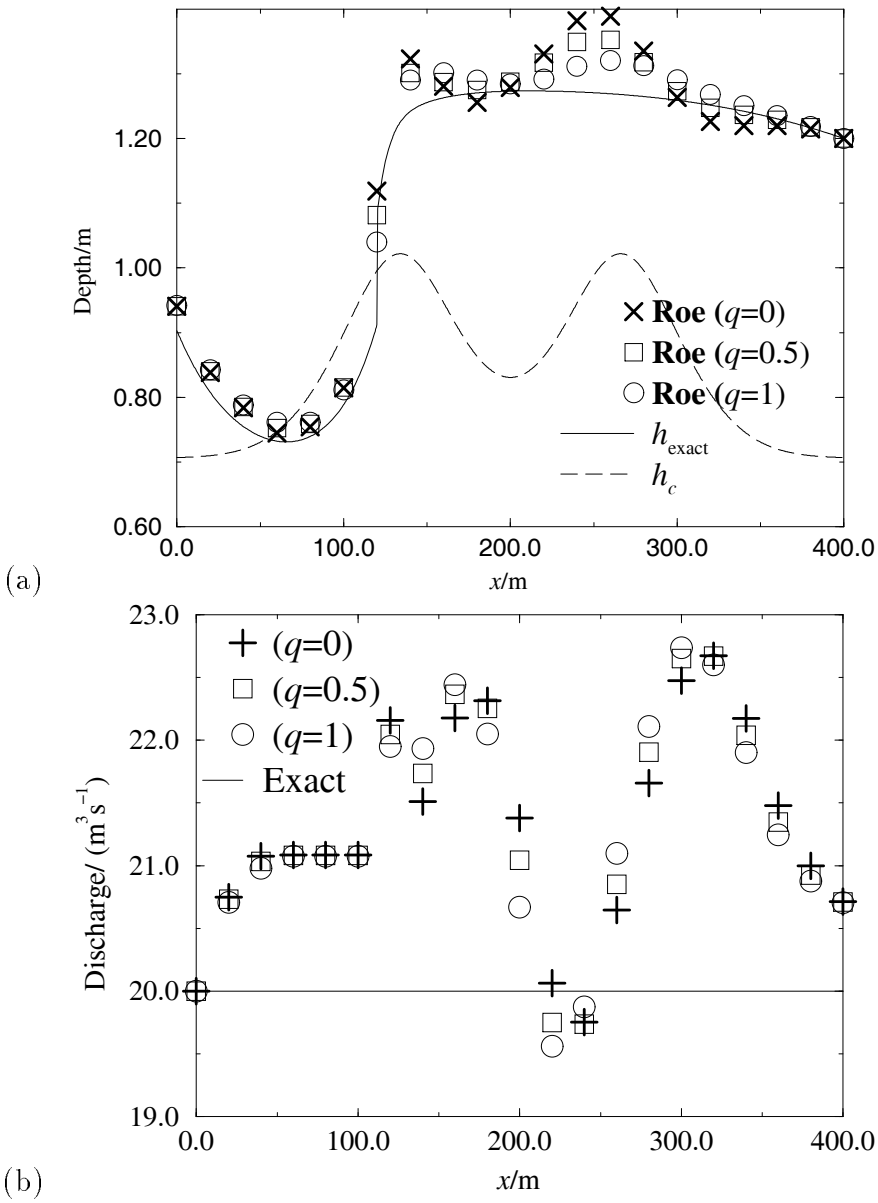


Figure 9.14: Roe's scheme for problem 14 (Δx =20m).

subcritical region of flow downstream of the jump.

We conclude that Roe's scheme is in general less accurate than the Engquist-Osher scheme. The reason for this is most likely the large deviations in the discharge. For regions of supercritical flow we find (for the reason discussed in section 7.2) that the discharge is constant (although not necessarily at the correct level). In regions of subcritical flow, the scheme is not consistent at steady state with a constant discharge, resulting in large deviations in the discharge field. This was remedied in section 7.2 by upwinding the source term. This does not work for a non-prismatic

channel, because of the additional term (9.7) which has not been decomposed onto the eigenvectors of the Roe matrix. The approach of Priestley[53], which absorbs the additional term into the source term and upwinds the modified source term, does however result in a constant discharge (away from jumps), but also results in a non-conservative scheme.

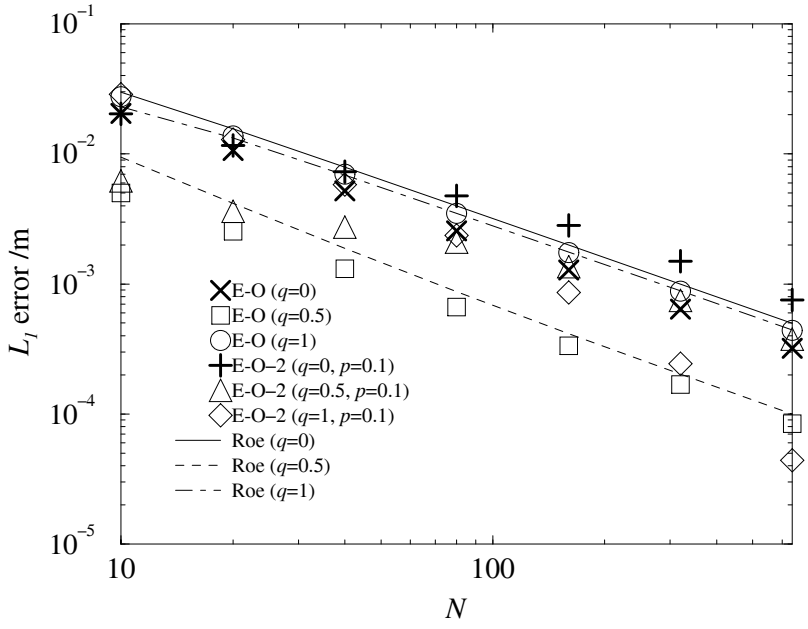


Figure 9.15: L_1 errors for problem 9.

9.3 Conclusions

We have seen in this chapter that the "scalar approach" can be extended to the case of non-prismatic channels. A parameter q was introduced to control the cross-sections at which the numerical flux was evaluated, with the choice $q = 1/2$ (so that the fluxes were evaluated at the cell interfaces) giving the best accuracy in more cases than not. The technique of improving the accuracy of the solution by upwinding the source term was found be very much less effective than for the prismatic case, since depending on the choice of the parameter q, the scheme is only second order accurate in regions of subcritical flow or regions of supercritical flow. For the E-O-2 scheme it was often found that a gain in accuracy in a region of one type of flow results in a reduction in accuracy in regions of the opposite type of flow. This approach

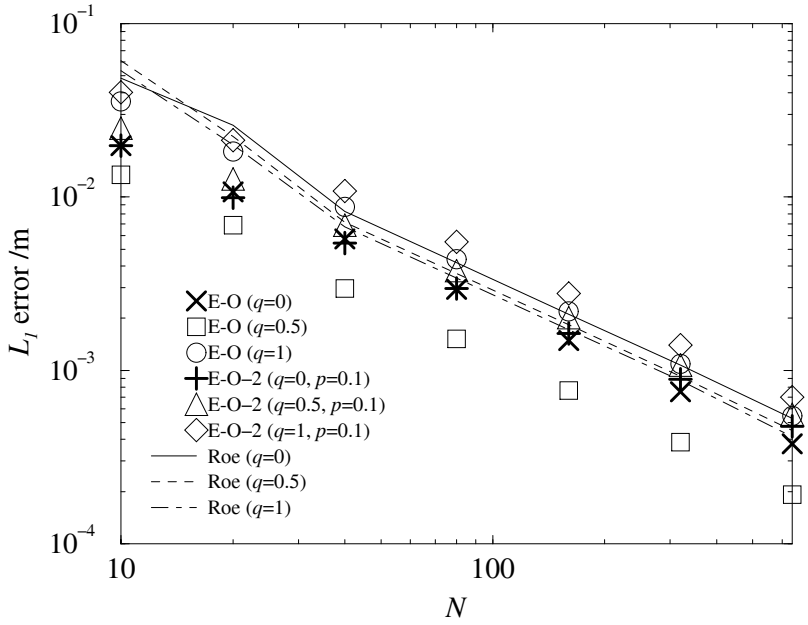


Figure 9.16: L_1 errors for problem 11.

may therefore only be of any benefit for solutions of predominantly one type of flow. The generalisation of the upwind-2 scheme was found to be extremely problematic due non-convergence of the time-stepping iteration, whereas the E-O and E-O-2 schemes were, as for the prismatic case, very amenable to Newton's method. Roe's scheme was also generalised to the non-prismatic case. The scheme has difficulties capturing smooth transitions but this could be cured by addition of an entropy fix. Roe's method was in general found to less accurate than the scalar schemes, due to the large amount of deviation from the expected constant discharge. This cannot be remedied by upwinding the source term without a special treatment of the new term arising from the variation of the cross-section of the channel. This last subject requires further work.

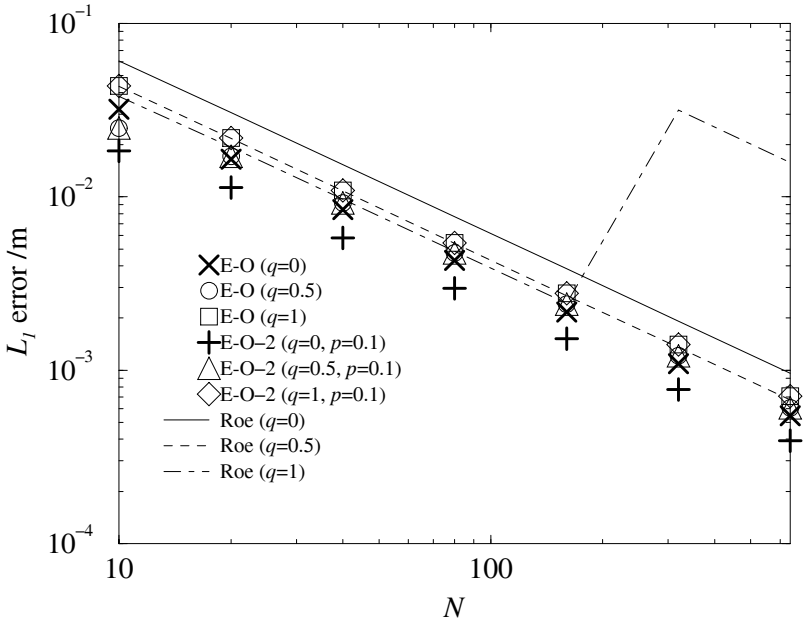


Figure 9.17: L_1 errors for problem 12.

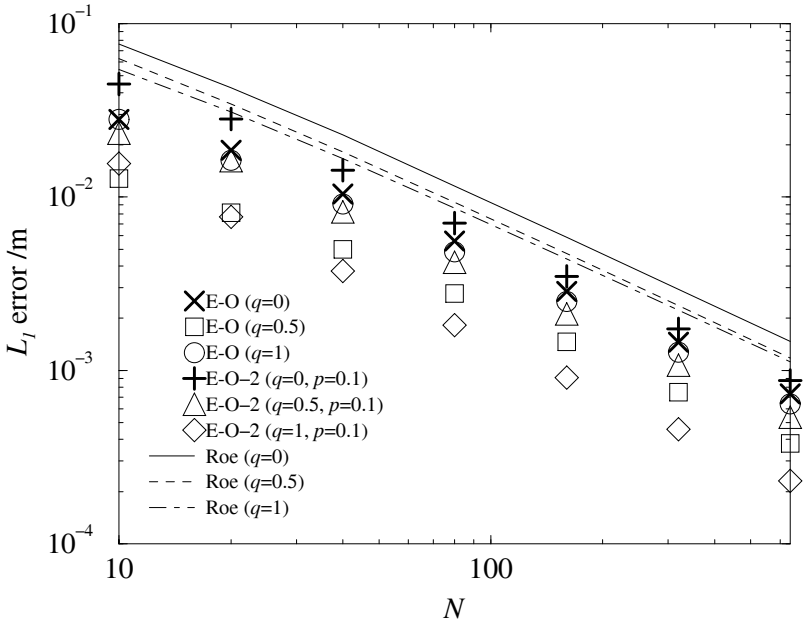


Figure 9.18: L_1 errors for problem 14.

Chapter 10

Conclusions and Further Work

In this thesis we have considered both analytical and numerical aspects of the steady state Saint-Venant problem, with particular emphasis on the subject of discontinuous solutions.

The main analytical results were presented in Chapter 4 where it was shown under certain conditions that the steady flow problem has at most one physically allowable solution for any particular set of boundary conditions. Information was also obtained regarding which sets of boundary conditions actually provide a solution. The major requirements for the theory to apply are that the channel be prismatic, have positive bed slope and have only a single critical depth. In addition to other minor technical requirements, the conveyance, which gives the form of the friction law, must be a strictly increasing function of depth. Provided the bed slope is positive and sufficiently smooth, then all the necessary conditions were shown to hold for trapezoidal channels (including the special cases of rectangular and triangular channels) with either the Manning or Chezy forms of the conveyance. For problems where the bed slope is not everywhere positive, we demonstrated that there can be multiple physical solutions for a given set of boundary conditions. The existence of multiple solutions is also anticipated for non-prismatic channels, thus it may not be possible to extend the theory to a significantly wider class of problems.

The "scalar approach" for numerically computing steady solutions was introduced, which arose from the observation that, under certain conditions, the physical solutions of the steady flow problem are exactly the steady entropy satisfying solu-

tions of a particular scalar conservation law. Scalar shock capturing schemes could thus be used to compute discontinuous solutions of the steady flow problem. A family of such schemes, which are particularly rich in theoretical results, are the monotone schemes. In Chapter 5 we demonstrated that this richness carries over to the computation of solutions of the steady flow problem. Under the same conditions as the theory in Chapter 4 we demonstrated that the numerical schemes define a solution which converges to the unique physical solution of the steady flow problem (as the grid-spacing vanishes).

In order to assess the performance of particular methods, it is useful to have model problems for which the exact solution is known. Other than for certain idealised cases, there appear to be no such model problems for the steady flow problem. In Chapter 6 we presented a relatively simple procedure for constructing test problems with known solutions. This allowed us to construct a series of test problems, including problems with discontinuous solutions. The test problems were constructed using an inverse approach, in that we calculated the bed slope necessary to give a specified solution.

We commenced Chapter 7 by applying the Engquist-Osher, Godunov and Lax-Friedrichs monotone schemes and the first-order upwind scheme to a selection of the test cases and compared the numerical solutions against the exact solutions. The Engquist-Osher, Godunov and first-order upwind schemes were all found to give good representations of the actual solutions, differing in accuracy only near transitions. The Lax-Friedrichs scheme was found to be considerably more diffusive for values of the free parameter necessary to make the scheme monotone over the depth range of interest. We investigated the effectiveness of certain a-priori estimates arising from the theory, and in general these were found to be too pessimistic to be of great use. In particular the time step predicted was often found to be many times smaller than the optimum value.

The results for the Engquist-Osher scheme were compared against those obtained from time accurately integrating the Saint-Venant system using Roe's approximate Riemann solver. When for the latter approach the source term was discretised in a pointwise manner, comparable accuracy to the Engquist-Osher scheme was observed.

However in general the discharge for Roe's schemes was found to be far from constant at steady state. This was remedied by using an upwind discretisation of the source term. Moreover for a particular form of source term averaging it was found that the scheme gave second order accuracy at steady state. This was explained by showing that, at steady state, the scheme effectively reduces to the trapezium rule. Upwinding of the source term was also used to obtain second order accurate three-point scalar schemes. The resulting generalisations of the Engquist-Osher and first-order upwind schemes were found to give comparable accuracy to Roe's scheme with an upwinded source term. Another more traditional approach to obtaining second order accuracy is through the use of nonlinear limiter functions, leading to five-point TVD schemes. Examples of such schemes were found to be significantly less accurate than the schemes with upwinded source term.

The accuracy of a numerical method is only one important factor in the overall performance of the method. Another important factor is the efficiency with which solutions can be computed. The "scalar approach" reduces to solving a system of nonlinear difference equations and the most natural method for carrying this out is through a time stepping iteration which effectively models the transient behaviour of a scalar conservation law. This algorithm is found to be significantly more efficient than the time integration of the system of equations. However in Chapter 8 we investigated potentially more efficient techniques for solving the difference equations. For the first order schemes (Engquist-Osher, first-order upwind and Godunov schemes) we observed that Newton's method can give a considerable improvement in efficiency and is also very robust. This is also the case for the second order modification to the Engquist-Osher scheme. Newton's method is found to fail completely for the first-order upwind scheme with upwinded source term, because of the discontinuous manner with which the source term discretisation switches across a transition. Newton's method is also found to have severe difficulties with the five-point TVD schemes, although the linearised implicit algorithm still gave much better performance than the explicit time stepping iteration. We demonstrated that the efficiency of Newton's method can often be improved by solving the difference equations on a series of grids of increasing fineness, using linear interpolation to transfer solutions

between grids.

The schemes discussed thus far are only applicable in the case of a prismatic channel. In Chapter 9 we extended the scalar schemes to the non-prismatic case by allowing the numerical flux functions to depend on the distance along the channel. We investigated different ways of staggering the evaluation points, and for the Engquist-Osher scheme we found that, in the majority of test cases, the most accurate solution was obtained by evaluating the numerical flux at the cell interfaces. The version of the Engquist-Osher scheme with upwinded source term was no longer found to give second order accuracy in all regions of the solution. A particular staggering of the evaluation points may give second order accuracy in one flow regime, but the scheme remains only first order accurate in regions of the opposite type of flow. We concluded that upwinding the source term was only beneficial for solutions of predominantly one type of flow. Further work is required to develop a conservative scheme that is second-order accurate in both flow regimes for a non-prismatic channel.

The same idea as above was used to extend Roe's scheme to the case of non-prismatic channels, and this method was found to less accurate than the Engquist-Osher scheme. We concluded that this is due to the non-constant discharge. Unlike the prismatic case, however, this cannot be rectified by upwinding the source term, because the additional term, due to the variation of the channel cross-section with distance, is not decomposed onto eigenvectors of the Roe matrix. It may be sufficient to simply decompose the additional term onto the eigenvectors in such a way that the scheme remains conservative.

The performance of the numerical schemes discussed in this thesis can be improved by allowing variable grid spacing, and this is relatively simple. To make full use of the variable grid spacing requires some kind of grid adaptivity, so that extra grid points are only placed where really necessary, for example near discontinuities. This is a topic for future work.

For the "scalar approach" we have treated the solution at the ends of the channel in an unsophisticated manner, fixing the values of the depth at the boundaries, regardless of whether these represent physical boundary conditions. For schemes which are upwind in nature this does not affect the accuracy of the solution away from the

boundaries, however some applications may require an accurate approximation of the solution at the boundaries. Further work is therefore necessary to investigate ways of achieving this.

One approach is to solve the system of difference equations as usual and then, if necessary, extrapolate the solution onto the boundaries. For an upwind scheme, the appropriate one-sided form of the scheme may used to perform the extrapolation, maintaining the accuracy of the scheme at the boundaries. A difficulty with this technique is to decide when extrapolation is required, for example we must differentiate between a hydraulic jump close to the boundary (for which extrapolation is not appropriate) and a boundary layer.

A more elegant approach is to build the treatment of the boundary conditions into the numerical scheme. In the case of the time stepping iteration, this effectively models the transient solution of a partial differential equation, and we could therefore treat the boundary conditions in the same manner as we would for a scalar partial differential equation. This is relatively straightforward for an upwind scheme, although a problem with this approach is that the solution may consequently depend on the initial guess of the solution. For example, suppose the initial data has a discontinuity and this leaves the domain as time progresses, overriding a boundary condition. The eventual steady solution will not now satisfy the overridden boundary condition, even if a steady solution satisfying the boundary condition does exist.

An application where a special treatment is required at the ends of the channel is that of solving for the steady flow in a network of channels. The need to compute steady solutions for large looped networks of channels was a major motivation for developing more efficient steady state solvers. The "scalar approach" is not particularly suitable for this application, since the constant discharges in each channel in the network are not initially known. One approach is to construct an iterative procedure to compute the discharges and match up the depths at the junctions. However preliminary investigations into such an approach proved disappointing.

The "scalar approach" has inherent limitations due to the fact that, in general, a problem may have more than a single physical steady solution satisfying a given set of boundary conditions. In this case we cannot expect to determine the actual steady

solution without referring to the transient flow. Work is required to investigate how common the existence of multiple solutions is and the behaviour of the scalar schemes in such cases.

Bibliography

[1] L Abrahamsson and S Osher. Monotone difference schemes for singular perturbation problems. *SIAM J. Numer. Anal.*, 19(5):979–992, October 1982.

[2] M J Baines, A Maffio, and A Di Filippo. Unsteady 1-D flows with steep waves in plant channels: The use of Roes's upwind TVD difference scheme. *Advances in Water Resources*, 15:89–94, 1992.

[3] P A Burton and P K Sweby. A dynamical approach study of some explicit and implicit TVD schemes and their convergence to steady-state solutions. Numerical Analysis Report 5/95, University of Reading, Department of Mathematics, PO Box 220, Reading RG6 6AF, UK, 1995.

[4] V T Chow. *Open Channel Hydraulics.* McGraw-Hill Publishing Company, New York, 1959.

[5] R Courant and K O Friedrichs. *Supersonic Flow and Shock Waves.* Springer-Verlag, New York, 1948.

[6] R Courant, K O Friedrichs, and H Lewy. Uber die partiellen differenzengleichungen der mathematisches physik. *Math. Ann.*, 100:32–74, 1928.

[7] R Courant, K O Friedrichs, and H Lewy. On the partial differential equations of mathematical physics. *IBM Journal*, 11:215–234, 1967.

[8] M G Crandall and A Majda. Monotone difference approximations for scalar conservation laws. *Mathematics of Computation*, 34(149):1–21, January 1980.

[9] J A Cunge, F M Holly, and A Verwey. *Practical Aspects of Computational River Hydraulics.* Pitman Publishing Ltd., New York, 1980.

[10] S F Davis. TVD finite difference schemes and artificial viscosity. ICASE Report 84-20, 1984.

[11] B de Saint-Venant. Théorie du mouvement non-permanent des eaux avec application aux crues des rivières et à l'introduction des marées dans leur lit. *Acad. Sci. Comptes rendus*, 73:148–154,237–240, 1871.

[12] D Dee. Standard validation document:- definition and guidelines. Technical report, Delft Hydraulics, Netherlands, 1994.

[13] P G Drazin. *Nonlinear Systems*. Cambridge University Press, Cambridge, 1992.

[14] B Engquist and S Osher. Stable and entropy satisfying approximations for transonic flow calculations. *Mathematics of Computation*, 34(149):45–75, 1980.

[15] B Engquist and S Osher. One sided difference approximations for nonlinear conservation laws. *Mathematics of Computation*, 36:321–351, 1981.

[16] R H French. *Open Channel Hydraulics*. McGraw-Hill Publishing Company, New York, 1986.

[17] P Garcia-Navarro, A Priestley, and F Alcrudo. An implicit method for water flow modelling in channels and pipes. *Journal of Hydraulic Research*, 32(5):721–742, 1994.

[18] P Glaister. Second order difference schemes for hyperbolic conservation laws with source terms. Numerical Analysis Report 6/87, University of Reading, Department of Mathematics, PO Box 220, Reading RG6 6AF, UK, 1987.

[19] S K Godunov. A finite difference method for the numerical computation of discontinuous solutions of the equation of fluid dynamics. *Mat. Sb.*, 47:271, 1959.

[20] A Harten. On a class of high resolution total-variation-stable finite difference schemes. *SIAM J. Num. Anal.*, 21:1–23, 1984.

[21] A Harten, B Engquist, S Osher, and S Chakravarthy. Uniformly high order accurate essentially non-oscillatory schemes. *J. Comput. Phys.*, 71:231, 1987.

[22] A Harten and J M Hyman. Self-adjusting grid methods for one-dimensional hyperbolic conservation laws. *J. Comput. Phys.*, 50:235–269, 1983.

[23] A Harten, J M Hyman, and P D Lax. On finite-difference approximations and entropy conditions for shocks. *Communications on Pure and Applied Mathematics*, 29:297–322, 1976.

[24] F M Henderson. *Open Channel Flow.* The Macmillan Company, New York, 1966.

[25] F A Howes. Boundary-interior layer interactions in nonlinear singular perturbation theory. *Mem. Amer. Math. Soc*, (203), 1978.

[26] H B Humpidge and W D Moss. The development of a comprehensive computer program for the calculation of flow profiles in open channels. *Proc. of the Inst. of Civ. Engrs.*, 50:49–64, 1971.

[27] L K Jackson. Subfunctions and second-order ordinary differential inequalities. *Adv. in Math.*, 2:307–363, 1968.

[28] D W Jordan and P Smith. *Nonlinear Ordinary Differential Equations.* Oxford University Press, Oxford, second edition, 1987.

[29] P Lax and B Wendroff. Systems of conservation laws. *Communications on Pure and Applied Mathematics*, 13:217–237, 1960.

[30] R J LeVeque. *Numerical Methods for Conservation Laws.* Lectures in Mathematics. Birkhaüser, Basel, 1992.

[31] J A Liggett. Stability. In K Mahmood and V Yevjevich, editors, *Unsteady Flow in Open Channels*, volume 1, chapter 6, pages 259–282. Water Resources Publications, Fort Collins, Colorado, 1975.

[32] J Lorenz. Nonlinear boundary value problems with turning points and properties of difference schemes. In W Eckhaus and E M de Jager, editors, *Lecture*

Notes in Applied Mathematics, volume 942 of *Springer Lecture Notes in Math.*, pages 150–169. Springer Verlag, New York, 1982.

[33] J Lorenz. Numerical solution of singular perturbation problem with turning points. In H W Knoblock and K Schmitt, editors, *Equadiff 82*, volume 1017 of *Lecture Notes in Applied Mathematics*, pages 432–439. Springer Verlag, New York, 1983.

[34] J Lorenz. Analysis of difference schemes for a stationary shock problem. *SIAM J. Numer. Anal.*, 21(6):1038–1053, December 1984.

[35] J Lorenz. Convergence of upwind schemes for a stationary shock. *Mathematics of Computation*, 46(173):45–57, January 1986.

[36] J Lorenz and R Sanders. On the rate of convergence of viscosity solutions and boundary value problems. *SIAM J. Math. Anal.*, 18(2):306–320, March 1987.

[37] I MacDonald. Test problems with analytic solutions for steady open channel flow. Numerical Analysis Report 6/94, University of Reading, Department of Mathematics, PO Box 220, Reading RG6 6AF, UK, 1994.

[38] I MacDonald, M J Baines, N K Nichols, and Samuels P G. Analytic benchmark solutions for open channel flows. To appear in the Journal of Hydraulic Engineering, ASCE.

[39] I MacDonald, M J Baines, N K Nichols, and P G Samuels. Comparison of some steady state Saint-Venant solvers for some test problems with analytic solutions. Numerical Analysis Report 2/95, University of Reading, Department of Mathematics, PO Box 220, Reading RG6 6AF, UK, 1995.

[40] I MacDonald, M J Baines, N K Nichols, and P G Samuels. Steady open channel test problems with analytic solutions. Numerical Analysis Report 3/95, University of Reading, Department of Mathematics, PO Box 220, Reading RG6 6AF, UK, 1995.

[41] E M Murman. Analysis of embedded shockwaves calculated by relaxation methods. *AIAA Journal*, 12:636, 1974.

[42] E M Murman and Cole J D. Calculation of plane steady transonic flows. *AIAA Journal*, 9:114, 1971.

[43] M Nagumo. Über die differentialgleichung $y'' = f(x, y, y')$. *Proc. Phys. Math. Soc. of Japan*, 19:861–866, 1937.

[44] O Oleinik. Discontinuous solutions of non-linear differential equations. *Usp. Mat. Nauk.*, 12:3–73, 1957. Translation in Amer. Math. Soc. Transl. Ser. 2, 96, 95-172, 1963.

[45] R E O'Malley. *Introduction to Singular Perturbations.* Academic Press, New York, 1974.

[46] J M Ortega and W C Rheinboldt. *Iterative Solution of Nonlinear Equations in Several Variables.* Academic Press, New York, 1970.

[47] S Osher. Nonlinear singular perturbation problems and one sided difference schemes. *SIAM J. Numer. Anal.*, 18(1):129–144, February 1981.

[48] S Osher. Riemann solvers, the entropy condition and difference approximations. *SIAM J. Numer. Anal.*, 21:217–235, 1984.

[49] S Osher and F Solomon. Upwind difference schemes for hyperbolic systems of conservation laws. *Mathematics of Computation*, 38(158):339–374, April 1982.

[50] C E Pearson. On the differential equation of boundary layer type. *J. Maths. Phys.*, 47:134–154, 1968.

[51] W H Press, S A Teukolsky, W T Vetterling, and B P Fanner. *Numerical Recipes in Fortran.* Cambridge University Press, Cambridge, 2nd edition, 1992.

[52] R K Price and P G Samuels. A computational hydraulic model for rivers. *Proc. of the Inst. of Civ. Engrs.*, 69(2):87–96, 1980.

[53] A Priestley. Roe type schemes for super-critical flow in rivers. Numerical Analysis Report 13/89, University of Reading, Department of Mathematics, PO Box 220, Reading RG6 6AF, UK, 1989.

[54] A Priestley. A quasi-Riemann method for the solution of one-dimensional shallow water flow. *J. Comput. Phys.*, 106(1):139–146, 1993.

[55] P L Roe. Approximate Riemann solvers, parameter vectors, and difference schemes. *Journal of Computational Physics*, 43:357–372, 1981.

[56] P L Roe. Generalised formulation of TVD Lax-Wendroff schemes. ICASE Report 84-53, 1984.

[57] P L Roe. Upwind differencing schemes for hyperbolic conservation laws with source terms. In C Carasso, P A Raviart, and D Serre, editors, *Nonlinear Hyperbolic problems*, Springer Lecture Notes in Mathematics. Springer Verlag, 1986.

[58] P G Samuels. *Modelling River and Flood Plain Flow using the Finite Element Method.* PhD thesis, University of Reading, Department of Mathematics, 1985.

[59] P G Samuels. Backwater lengths in rivers. *Proc. InstnCiv. Engrs.*, 87:571–582, 1989.

[60] P G Samuels and C P Skeels. Stability limits for Preissmann's scheme. *Journal of Hydraulic Engineering*, 116(8):997–1012, August 1990.

[61] C P Skeels. *One Dimensional River Modelling.* PhD thesis, Department of Numerical Analysis, Faculty of Mathematics, University of Oxford, 1992.

[62] J Smoller. *Shock Waves and Reaction-Diffusion Equations.* Springer-Verlag, New York, 1982.

[63] D S G Stirling. *Mathematical Analysis.* Ellis Horwood Ltd., Chichester, U.K., 1987.

[64] J J Stoker. *Water Waves.* Interscience Publishers, Inc., New York, 1986.

[65] K R Stromberg. *An Introduction to Classical Real Analysis*. Wadsworth, Inc., Belmont, California, 1981.

[66] P K Sweby. *Shock Capturing Schemes*. PhD thesis, University of Reading, Department of Mathematics, 1982.

[67] P K Sweby. High resolution schemes using flux limiters for hyperbolic conservation laws. *SIAM J. Numer. Anal.*, 21(5):995–1011, October 1984.

[68] J P Vila. Simplified Godunov schemes for 2x2 systems of conservation laws. *SIAM J. Numer. Anal.*, 23(6):1173–1192, December 1986.

[69] S Wolfram. *Mathematica, A System for Doing Mathematics by Computer*. Addison-Wesley, New York, 1988.

[70] H C Yee. Construction of a class of symmetric TVD schemes. *The IMA Volumes in Mathematics and its Applications*, 2:381–396, 1986.

[71] H C Yee. Construction of explicit and implicit symmetric TVD schemes and their applications. *Journal of Computational Physics*, 68:151–179, 1987.

[72] H C Yee. A class of high-resolution explicit and implicit shock-capturing methods. NASA Technical Memorandum 101088, Ames Research Center, Moffet Field, California, 1989.

[73] B C Yen. Open-channel flow equations revisited. *Journal of Engineering Mechanics Division, ASCE*, 99(EM5):979–1009, 1973.

Appendix A

Theory for the Second Order Modification to Engquist-Osher

In this appendix we present theory for the second order modification of the Engquist-Osher scheme due to [33]. This is essentially the analogue of Theorem 10 and is likewise proven by Lemma 5.1. Under the conditions of Theorem 3 and if the parameter p satisfies a condition over some range of depths, then the system of difference equations have a unique solution in this range. The uniqueness is not global since in general there is not one value of p which satisfies the condition over all depth ranges. This is exactly the same as for the parameter λ for the Lax-Friedrichs scheme. The result can be obtained via the theory of M-functions (see [46]), however the method used here has the advantage of yielding a modified CFL condition which guarantees the convergence of the time-stepping iteration. Unlike for the first order scheme, we have no results concerning the convergence of the discrete solution in the limit as the grid spacing vanishes. The system of difference equations are as follows

$$\begin{aligned} \mathcal{T}_j^{\text{E-O-2}} h = 0, \qquad & j = 1, 2, \ldots, N-1 \\ h_0 = \gamma_0, \quad & h_N = \gamma_1. \end{aligned} \tag{A.1}$$

Here the operator $\mathcal{T}_j^{\text{E-O-2}}$ is given by

$$\begin{aligned}\mathcal{T}_j^{\text{E-O-2}}h &= \frac{g^{\text{E-O}}(h_{j+1},h_j)-g^{\text{E-O}}(h_j,h_{j-1})}{\Delta x}\\ &\quad+\chi_j^- D(x_{j-1},h_{j-1})+\chi_j^0 D(x_j,h_j)+\chi_j^+ D(x_{j+1},h_{j+1}),\end{aligned}$$

where

$$\begin{aligned}\chi_j^- &= \chi\left(\frac{pf'(h_{j-1})}{\sqrt{\Delta x}}\right),\\ \chi_j^0 &= 1-\chi\left(\frac{pf'(h_j)}{\sqrt{\Delta x}}\right)-\chi\left(\frac{-pf'(h_j)}{\sqrt{\Delta x}}\right)=1-\chi_{j+1}^- -\chi_{j-1}^+,\\ \chi_j^+ &= \chi\left(\frac{-pf'(h_{j+1})}{\sqrt{\Delta x}}\right),\end{aligned}$$

$p\geq 0$ is a parameter and χ is the smooth increasing function

$$\chi(r)=\begin{cases}0 & r<0\\ r^2 & 0\leq r\leq\frac{1}{2}\\ \frac{1}{2}-(1-r)^2 & \frac{1}{2}\leq r\leq 1\\ \frac{1}{2} & r>1\end{cases}$$

connecting the values 0 and $\frac{1}{2}$. It is convenient to write

$$\mathcal{T}_j^{\text{E-O-2}}h = \frac{\hat{g}_{j+\frac{1}{2}}-\hat{g}_{j-\frac{1}{2}}}{\Delta x}+D(x_j,h_j),$$

where

$$\hat{g}_{j+\frac{1}{2}}=g^{\text{E-O}}(h_{j+1},h_j)+\Delta x\left(\chi_j^+ D(x_{j+1},h_{j+1})-\chi_{j+1}^- D(x_j,h_j)\right).$$

This is now in a similar form to the first order scheme. In the next lemma we show that under certain conditions that $\hat{g}_{j+\frac{1}{2}}$ is non-increasing in h_{j+1} and non-decreasing in h_j.

Lemma A.1 *Under the conditions of Theorem 3, if $f\equiv -F$, $0<\alpha\leq\underline{h}$, $\beta\geq\overline{h}$ and $p>0$ satisfies*

$$p\sqrt{\Delta x}M_1\leq 1,\qquad 4p^2M_2\leq 1, \tag{A.2}$$

where $|D_h(x_j,h)|\leq M_1$ and $|f''(h)D(x_j,h)|\leq M_2$ for all $h\in[\alpha,\beta]$ and $0\leq j\leq N$, then

$$\frac{\partial\hat{g}_{j+\frac{1}{2}}}{\partial h_{j+1}}\leq 0,$$

for $0 \le j+1 \le N$, $h_{j+1} \in [\alpha, \beta]$ *and*

$$\frac{\partial \hat{g}_{j+\frac{1}{2}}}{\partial h_j} \ge 0,$$

for $0 \le j \le N$, $h_j \in [\alpha, \beta]$.

Proof

$$\frac{\partial \hat{g}_{j+\frac{1}{2}}}{\partial h_j} = f'_+(h_j) - \Delta x \chi\left(\frac{pf'(h_j)}{\sqrt{\Delta x}}\right) D_h(x_j, h_j) - \sqrt{\Delta x} p \chi'\left(\frac{pf'(h_j)}{\sqrt{\Delta x}}\right) f''(h_j) D(x_j, h_j).$$

This is zero if $f'(h_j) \le 0$. In the case $f'(h_j) > 0$

$$\begin{aligned}
\frac{\partial \hat{g}_{j+\frac{1}{2}}}{\partial h_j} &\ge f'(h_j) - \Delta x \chi\left(\frac{pf'(h_j)}{\sqrt{\Delta x}}\right) M_1 - \sqrt{\Delta x} p \chi'\left(\frac{pf'(h_j)}{\sqrt{\Delta x}}\right) M_2 \\
&= \frac{\sqrt{\Delta x}}{p}\left(r - p\sqrt{\Delta x} M_1 \chi(r) - p^2 M_2 \chi'(r)\right) \\
&\ge \frac{\sqrt{\Delta x}}{p}\left(r - \chi(r) - \frac{1}{4}\chi'(r)\right),
\end{aligned}$$

where

$$r = \frac{pf'(h_j)}{\sqrt{\Delta x}} > 0.$$

Elementary analysis of the function χ shows that

$$r - \chi(r) - \frac{1}{4}\chi'(r) \ge 0,$$

for $r \ge 0$. Next we have

$$\begin{aligned}
\frac{\partial \hat{g}_{j+\frac{1}{2}}}{\partial h_{j+1}} &= f'_-(h_{j+1}) + \Delta x \chi\left(\frac{-pf'(h_{j+1})}{\sqrt{\Delta x}}\right) D_h(x_{j+1}, h_{j+1}) \\
&\quad - \sqrt{\Delta x} p \chi'\left(\frac{-pf'(h_{j+1})}{\sqrt{\Delta x}}\right) f''(h_{j+1}) D(x_{j+1}, h_{j+1}).
\end{aligned}$$

This is zero if $f'(h_j) \ge 0$. In the case $f'(h_j) < 0$

$$\begin{aligned}
\frac{\partial \hat{g}_{j+\frac{1}{2}}}{\partial h_{j+1}} &\le f'(h_{j+1}) + \Delta x \chi\left(\frac{-pf'(h_{j+1})}{\sqrt{\Delta x}}\right) M_1 + \sqrt{\Delta x} p \chi'\left(\frac{-pf'(h_{j+1})}{\sqrt{\Delta x}}\right) M_2 \\
&= \frac{\sqrt{\Delta x}}{p}\left(-r + p\sqrt{\Delta x} M_1 \chi(r) + p^2 M_2 \chi'(r)\right) \\
&\le \frac{\sqrt{\Delta x}}{p}\left(-r + \chi(r) + \frac{1}{4}\chi'(r)\right) \\
&\le 0,
\end{aligned}$$

where

$$r = \frac{-pf'(h_j)}{\sqrt{\Delta x}} > 0.$$

The above result leads to the following theorem.

Theorem 11 *Suppose the conditions of Theorem 3 hold and for* $0 < \alpha \le \underline{h}$, $\beta \ge \overline{h}$ *that* $p > 0$ *satisfies* (A.2). *If* $\Delta t > 0$ *satisfies*

$$\Delta t \left(\frac{|f'(h)|}{\Delta x} + D_h(x_j, h) + \frac{p}{\sqrt{\Delta x}} |f''(h) D(x_j, h)| \right) \le 1, \qquad \text{for all } h \in [\alpha, \beta] \text{ and } 0 \le j \le N, \tag{A.3}$$

(such a value exists since the coefficient of Δt *is positive and bounded on* $[\alpha, \beta]$*) then the mapping*

$$\boldsymbol{G} : [\boldsymbol{\alpha}, \boldsymbol{\beta}] \longrightarrow I\!R^{N+1},$$

given by

$$\boldsymbol{G}(\boldsymbol{h}) = \begin{bmatrix} \gamma_0 \\ h_1 - \Delta t \mathcal{T}_1^{E\text{-}O\text{-}2} h \\ \vdots \\ h_j - \Delta t \mathcal{T}_j^{E\text{-}O\text{-}2} h \\ \vdots \\ h_{N-1} - \Delta t \mathcal{T}_{N-1}^{E\text{-}O\text{-}2} h \\ \gamma_1 \end{bmatrix}, \tag{A.4}$$

has exactly one fixed point $\boldsymbol{h}$ *which is the only solution in* $[\boldsymbol{\alpha}, \boldsymbol{\beta}]$ *of the difference equations* (A.1). *Furthermore, for any initial guess* $\boldsymbol{h}^0 \in [\boldsymbol{\alpha}, \boldsymbol{\beta}]$ *the iteration*

$$\boldsymbol{h}^{n+1} = \boldsymbol{G}(\boldsymbol{h}^n), \qquad n = 0, 1, 2, \ldots$$

converges to the fixed point as $n \to \infty$ *and we have the convergence rate estimate*

$$\|\boldsymbol{h}^n - \boldsymbol{h}\|_1 \le (1 - \Delta t \delta)^n \|\boldsymbol{h}^0 - \boldsymbol{h}\|_1 \le \|\boldsymbol{h}^0 - \boldsymbol{h}\|_1 e^{-n \Delta t \delta},$$

where δ *is given by* (5.17) *and* $0 \le 1 - \Delta t \delta < 1$.

Proof To prove this result we simply apply Lemma 5.1. For $0 < j < N$ we have

$$\boldsymbol{G}(\boldsymbol{\alpha})\big|_j = \alpha - \Delta t \left(\chi_j^- D(x_{j-1}, \alpha) + \chi_j^0 D(x_j, \alpha) + \chi_j^+ D(x_{j+1}, \alpha) \right),$$

with the non-negative coefficients χ_j^-, χ_j^0, χ_j^+ being evaluated at $h_{j-1} = h_j = h_{j+1} = \alpha$. Since $0 < \alpha \le \underline{h} = \min\{\gamma_0, \gamma_1 . m\}$, then $D(x_j, \alpha) \le 0$ for $j = 0, 1, \ldots, N$, and we have

$$\boldsymbol{G}(\boldsymbol{\alpha}) \ge \boldsymbol{\alpha}.$$

Similarly for $0 < j < N$ we have

$$\boldsymbol{G}(\boldsymbol{\beta})|_j = \beta - \Delta t \left(\chi_j^- D(x_{j-1}, \beta) + \chi_j^0 D(x_j, \beta) + \chi_j^+ D(x_{j+1}, \beta)\right),$$

with the non-negative coefficients χ_j^-, χ_j^0, χ_j^+ being evaluated at $h_{j-1} = h_j = h_{j+1} = \beta$. Since $\beta \geq \overline{h} = \max\{\gamma_0, \gamma_1, m\}$, then $D(x_j, \beta) \geq 0$ for $j = 0, 1, \ldots, N$, and we have

$$\boldsymbol{G}(\boldsymbol{\beta}) \leq \boldsymbol{\beta}.$$

As in section 5.2 for $\boldsymbol{u}, \boldsymbol{v} \in [\boldsymbol{\alpha}, \boldsymbol{\beta}]$ we can write

$$\boldsymbol{G}(\boldsymbol{h}) - \boldsymbol{G}(\boldsymbol{v}) = M(\boldsymbol{h} - \boldsymbol{v}),$$

where

$$M = \int_0^1 \boldsymbol{G}'(\boldsymbol{h} + s(\boldsymbol{v} - \boldsymbol{h}))ds$$

and $\boldsymbol{G}'$ is of the form (5.15) with

$$p_j = \frac{\Delta t}{\Delta x}\frac{\partial \hat{g}_{j-\frac{1}{2}}}{\partial h_{j-1}},$$

$$\begin{aligned} q_j &= 1 - \frac{\Delta t}{\Delta x}\left(\frac{\partial \hat{g}_{j+\frac{1}{2}}}{\partial h_j} - \frac{\partial \hat{g}_{j-\frac{1}{2}}}{\partial h_j}\right) - \Delta t D_h(x_j, h_j) \\ &= 1 - \Delta t D_h(x_j, h_j) - p_{j+1} - r_{j-1}, \end{aligned}$$

$$r_j = -\frac{\Delta t}{\Delta x}\frac{\partial \hat{g}_{j+\frac{1}{2}}}{\partial h_{j+1}}.$$

The previous lemma shows that the p_j and r_j are non-negative. For $j = 0, 1, 2 \ldots, N$ we can write

$$\begin{aligned} q_j &= 1 - \Delta t\Bigg(\frac{|f(h_j)|}{\Delta x} + (1 - \chi(r) - \chi(-r))D_h(x_j, h_j) \\ &\quad + (\chi'(-r) - \chi'(r))\frac{pf''(h_j)}{\sqrt{\Delta x}}D(x_j, h_j)\Bigg) \qquad \left(r = \frac{pf'(h_j)}{\sqrt{\Delta x}}\right) \\ &\geq 1 - \Delta t\left(\frac{|f(h_j)|}{\Delta x} + D_h(x_j, h_j) + \frac{p}{\sqrt{\Delta x}}|f''(h_j)D(x_j, h_j)|\right) \\ &\geq 0, \end{aligned}$$

using the fact that $|\chi'(-r) - \chi'(r)| \leq 1$ and condition (A.3).

We estimate the L_1 norm of the matrix $\boldsymbol{G}'(h)$ by computing the sum of each column. The sum of the first column is p_1. Consider the expression q_0 where for the sake of argument take $h_{-1} = h_0$. We have shown that $q_0 \geq 0$, and we can write

$$p_1 = 1 - \Delta t D_h(x_0, h_0) - r_{-1} - q_0 \leq 1 - \Delta t\delta,$$

since $r_{-1} \geq 0$ from Lemma A.1. The sum of the second column is given by

$$q_1 + p_2 = 1 - \Delta t D_h(x_1, h_1) - r_0 \leq 1 - \Delta t\delta,$$

since $r_0 \geq 0$. For the j^{th} column $(3 \leq j \leq N-2)$ the sum is given by

$$r_{j-2} + q_{j-1} + p_j = 1 - \Delta t D_h(x_{j-1}, h_{j-1}) \leq 1 - \Delta t\delta.$$

The same argument shows that the remaining two column sums satisfy the same bound, hence we conclude that

$$\|\boldsymbol{G}'(h)\|_1 \leq 1 - \Delta t\delta.$$

It follows that

$$\|M\|_1 = \left\| \int_0^1 \boldsymbol{G}'(\boldsymbol{h} + s(\boldsymbol{v} - \boldsymbol{h}))ds \right\|_1 \leq \int_0^1 \|\boldsymbol{G}'(\boldsymbol{h} + s(\boldsymbol{v} - \boldsymbol{h}))\|_1 ds \leq 1 - \Delta t\delta < 1.$$

Thus Lemma 5.1 holds with $k = 1 - \Delta t\delta$. This completes the proof.

Appendix B

Test Problems for Non-Prismatic Channels

In this appendix we give the details of six test problems with non-prismatic channels, constructed using the method described in Chapter 6. Table B.1 gives the parameters for these problems.

The channel for problems 9, 10, 11 and 12 is rectangular with width given by

$$B_1(x) = 10 - 5\exp\left(-10\left(\frac{x}{200} - \frac{1}{2}\right)^2\right).$$

The width profile is illustrated in Figure B.1.

Problem	B/m	Z	L/m	n	$Q/(\mathrm{m^3 s^{-1}})$	h_{in}/m	h_{out}/m
9	$B_1(x)$	0	200	0.03	20		0.902921
10	$B_1(x)$	0	200	0.03	20	0.503369	
11	$B_1(x)$	0	200	0.03	20		
12	$B_1(x)$	0	200	0.03	20	0.700000	1.215485
13	$B_2(x)$	2	400	0.03	20		0.904094
14	$B_2(x)$	2	400	0.03	20		1.200000

Table B.1: Information for test problems 9-14

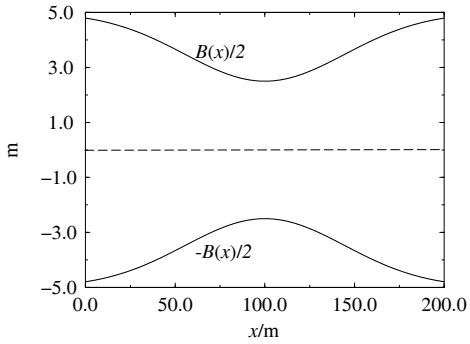


Figure B.1: Channel width for problems 9, 10, 11 and 12.

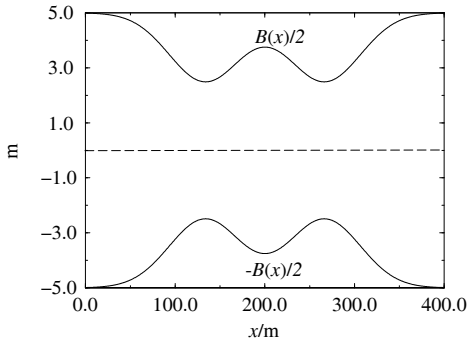


Figure B.2: Bottom width for problems 13 and 14.

The channel for problems 13 and 14 is trapezoidal with bottom width given by

$$B_2(x) = 10 - 5\exp\left(-50\left(\frac{x}{400} - \frac{1}{3}\right)^2\right) - 5\exp\left(-50\left(\frac{x}{400} - \frac{2}{3}\right)^2\right).$$

The bottom profile is illustrated in Figure B.2.

In each case the bed slope is given by

$$\begin{aligned} S_0(x) &= \left(1 - \frac{Q^2\left(B(x) + 2Z\hat{h}(x)\right)}{9.08665\left(\hat{h}(x)\right)^3\left(B(x) + Z\hat{h}(x)\right)^3}\right)\hat{h}'(x) \\ &\quad + Q^2n^2\frac{\left(B(x) + 2\hat{h}(x)\sqrt{1+Z^2}\right)^{4/3}}{\left(\hat{h}(x)\right)^{10/3}\left(B(x) + \hat{h}(x)\right)^{10/3}} \\ &\quad - \frac{Q^2B'(x)}{9.08665\left(\hat{h}(x)\right)^2\left(B(x) + Z\hat{h}(x)\right)^3}. \end{aligned}$$

It now only remains to specify the depth profile for each of the test cases.

Problem 9 (subcritical flow) In this case the depth is given by

$$\hat{h}(x) = 0.9 + 0.3\exp\left(-20\left(\frac{x}{200} - \frac{1}{2}\right)^2\right),$$

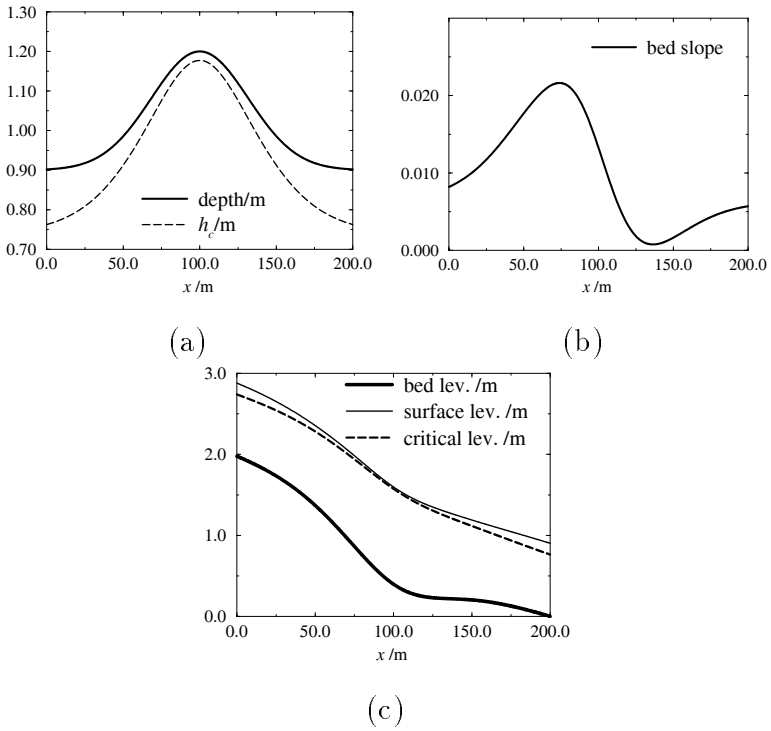


Figure B.3: Depth, bed slope, bed level and surface level for test problem 9

and the problem is illustrated by Figure B.3.

Problem 10 (supercritical flow) In this case the depth is given by

$$\hat{h}(x) = 0.5 + 0.5 \exp\left(-20\left(\frac{x}{200} - \frac{1}{2}\right)^2\right),$$

and the problem is illustrated by Figure B.4.

Problem 11 (smooth transition) In this case the depth is given by

$$\hat{h}(x) = 1.0 - 0.3\tanh\left(4\left(\frac{x}{200} - \frac{1}{3}\right)\right),$$

and the problem is illustrated by Figure B.5.

Problem 12 (hydraulic jump) In this case the depth is given by

$$\hat{h}(x) = 0.7 + 0.3\left(\exp\left(\frac{x}{200}\right) - 1\right),$$

for $x \leq 120$m, and for the remainder of the reach is of the form (6.6) with $x^* = 120$m, $x^{**} = 200$m, $M = 2$, $k_0 = -0.154375$, $k_1 = -0.108189$, $k_2 = -2.014310$, $p = 0.1$ and

$$\phi(x) = 1.5 \exp\left(-0.1\left(\frac{x}{200} - 1\right)\right).$$

This problem is illustrated by Figure B.6.

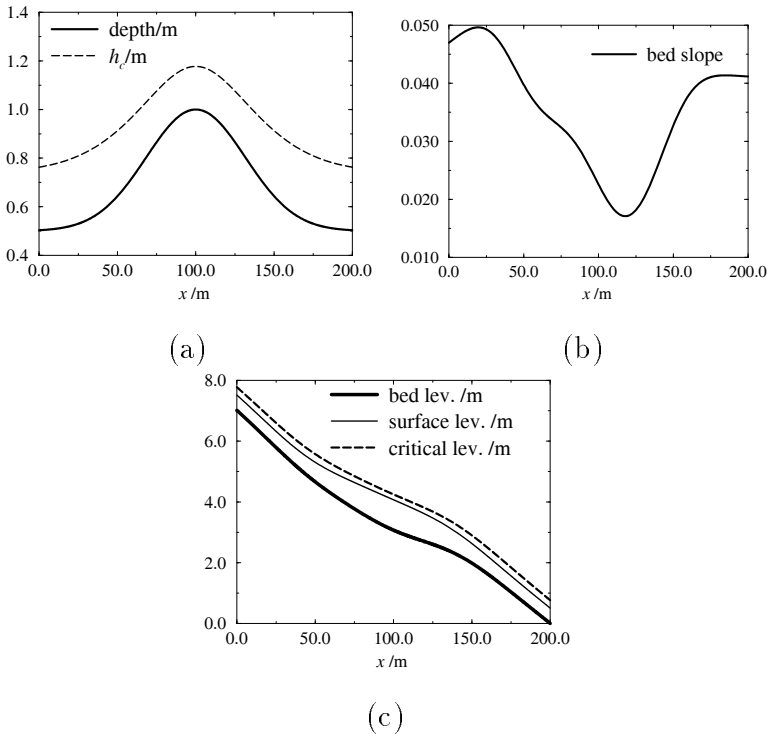


Figure B.4: Depth, bed slope, bed level and surface level for test problem 10

Problem 13 (subcritical flow) In this case the depth is given by

$$\hat{h}(x) = 0.9 + 0.3\exp\left(-40\left(\frac{x}{400} - \frac{1}{3}\right)^2\right) + 0.2\exp\left(-35\left(\frac{x}{400} - \frac{2}{3}\right)^2\right),$$

and the problem is illustrated by Figure B.7.

Problem 14 (smooth transition followed by hydraulic jump) In this case the depth is given by

$$\hat{h}(x) = 0.9 + 0.25\left(\exp\left(-\frac{x}{40}\right) - 1\right) + 0.25\exp\left(15\left(\frac{x}{40} - \frac{3}{10}\right)\right),$$

for $x \leq 120$m, and for the remainder of the reach is of the form (6.6) with $x^* = 120$m, $x^{\star\star} = 400$m, $M = 2$, $k_0 = -0.183691$, $k_1 = 1.519577$, $k_2 = -18.234429$, $p = 0.09$ and

$$\phi(x) = 1.5\exp\left(0.16\left(\frac{x}{400} - 1\right)\right) - 0.3\exp\left(2\left(\frac{x}{400} - 1\right)\right).$$

This problem is illustrated by Figure B.8.

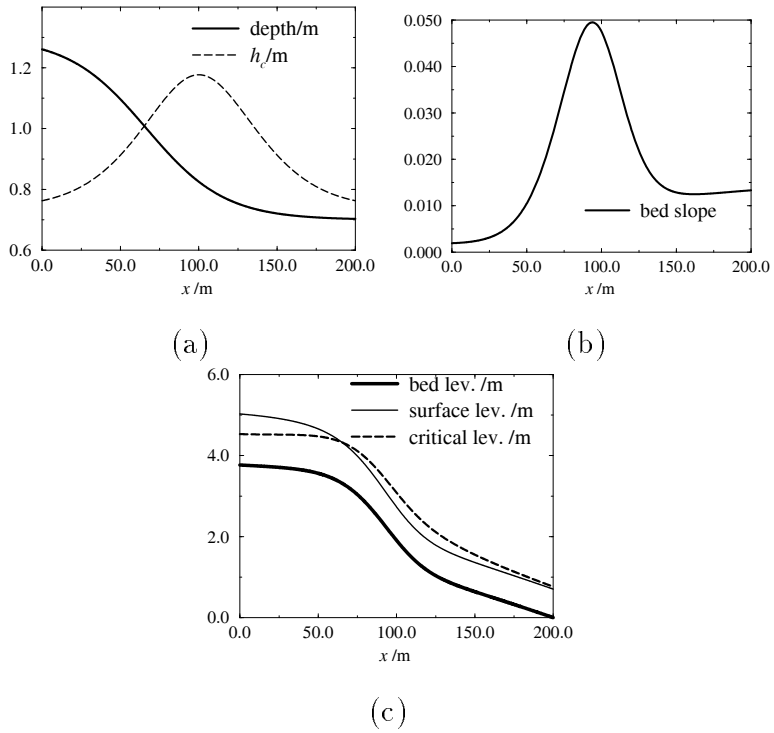


Figure B.5: Depth, bed slope, bed level and surface level for test problem 11

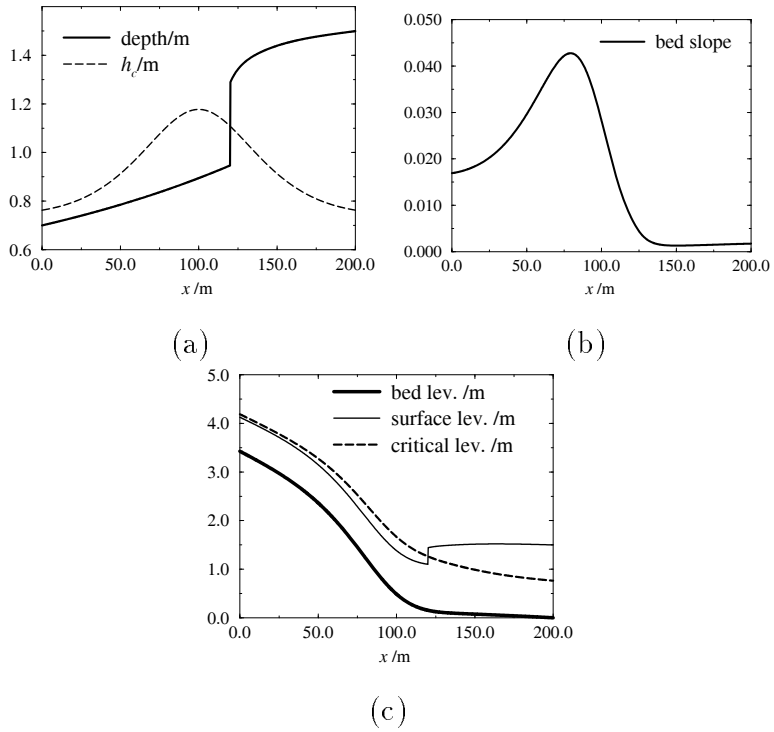


Figure B.6: Depth, bed slope, bed level and surface level for test problem 12

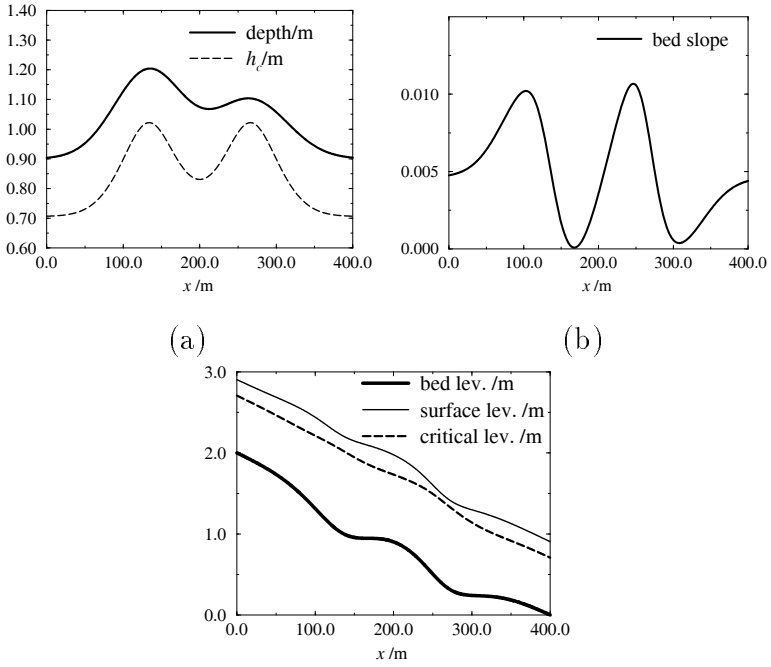


Figure B.7: Depth, bed slope, bed level and surface level for test problem 13

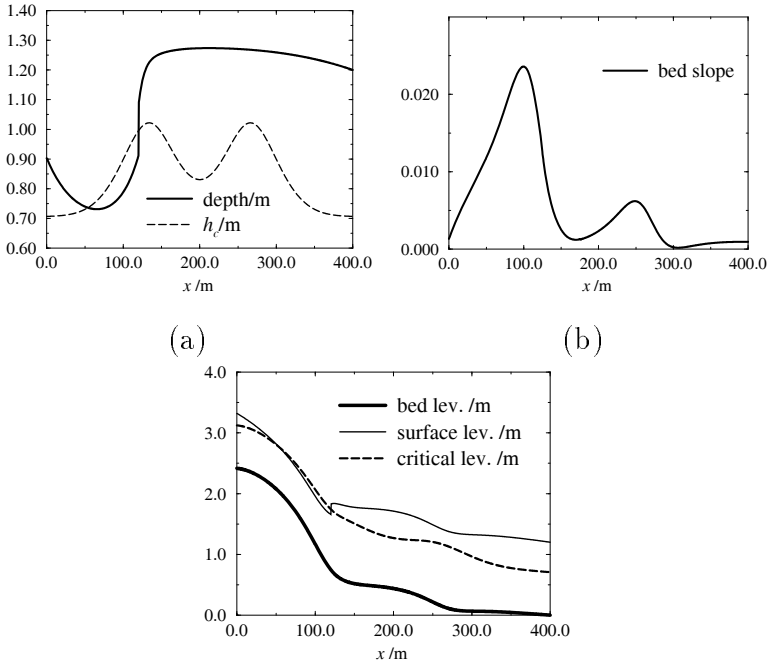


Figure B.8: Depth, bed slope, bed level and surface level for test problem 14